Weili He
Yixin Fang
Hongwei Wang *Editors*

Real-World Evidence in Medical Product Development

Springer

Real-World Evidence in Medical Product Development

Weili He • Yixin Fang • Hongwei Wang
Editors

Real-World Evidence in Medical Product Development

Springer

Editors
Weili He
AbbVie (United States)
Westfield, NJ, USA

Yixin Fang
AbbVie (United States)
Lincolnshire, IL, USA

Hongwei Wang
AbbVie (United States)
Vernon Hills, IL, USA

Disclaimer from Authors

The views expressed in this book are those of the authors and not necessarily reflective of the positions, policies, or practices of the authors' respective organizations.

ISBN 978-3-031-26327-9 ISBN 978-3-031-26328-6 (eBook)
https://doi.org/10.1007/978-3-031-26328-6

Mathematics Subject Classification: 62-02, 62D20, 62P99, 92C50

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The organic and evolving nature of real-world data (RWD) and real-world evidence (RWE) responding to the fit-for-purpose requirements for expanding applications of RWE to address payers, patients, physicians' need along with supporting regulatory decisions is a defining characteristic of this arena. Randomized controlled clinical trials (RCTs) have been the gold standard for the evaluation of efficacy and safety of medical interventions. However, the costs, duration, practicality, and limited generalizability have incentivized many to look for alternative ways to optimize it and address unique real-world research questions. In recent years, we have seen an increasing usage of RWD and RWE in clinical development and life-cycle management. The major impetus behind the interest in the use of RWE is the increased efficiency in drug development, resulting in savings of cost and time, ultimately getting drugs to patients sooner.

However, even with the encouragement from regulators and available guidance and literature on the use of RWD and RWE in recent years, many challenges remain. This book attempts to address these challenges by providing an end-to-end guidance including strategic considerations, state-of-the-art statistical methodology reviews, organization and infrastructure considerations, logistic challenges, and practical use cases. The target audience is anyone involved, or with an interest, in the use of RWE in their research for drug development and healthcare decision-making. In particular, it includes statisticians, clinicians, pharmacometricians, clinical operation specialists, regulators, and decision makers working in academic or contract research organizations, government, and industry. Our goal for this book is to provide, to the extent possible, a balanced and comprehensive coverage of key considerations and methodologies for the uptake of RWE in drug development. This book includes the following four parts:

- Part I: Real-World Data and Evidence to Accelerate Medical Product Development
- Part II: Fit-for-use RWD Assessment and Data Standards
- Part III: Causal Inference Framework and Methodologies in RWE Research
- Part IV: Application and Case Studies

Part I consists of three chapters. Chapter "The Need for Real-World Evidence in Medical Product Development and Future Directions" provides introduction and background on the need for RWE and RWD in clinical development and life-cycle management along with future directions. Chapter "Overview of the Current Real-World Evidence Regulatory Landscape" reviews existing guidance documents and precedents related to RWE by major regulatory agencies across the world. It also outlines the key concepts underpinning evaluation of RWE and discusses similarities and differences in those concepts in guidance documents from different countries. When we talk about fit-for-purpose use of RWE, it is very important to conceptualize right research questions that are clear and feasible to address. Chapter "Key Considerations in Forming Research Questions and Conducting Research in Real-World Setting" discusses key considerations in forming research questions.

Part II consists of four chapters. Chapter "Assessment of Fit-for-Use Real-World Data Sources and Applications" provides valuable information to guide practitioners on how to assess fit-for-use RWD sources via a framework and an example. As RWD sources may contain key data elements in different places, chapter "Key Variables Ascertainment and Validation in RW Setting" presents advanced analytics on how to ascertain key variables such as disease status, exposure, or outcomes. Chapter "Data Standards and Platform Interoperability" examines the role of health data and interoperability standards, their harmonization, and role within data platforms internationally as we see more utilization of platforms to interact with RWD for RWE generation, from a regulatory science and Health Technology Assessment (HTA) perspective. Often, one RWD source may not be sufficient to answer a research question, and multiple RWD sources may need to be linked to enrich the data and address the right research question. Chapter "Privacy-Preserving Record Linkage for Real-World Data" discusses several aspects behind Privacy-Preserving Record Linkage, including data pre-processing, privacy protection, linkage, and performance evaluation.

Part III contains ten chapters. This part covers state-of-art statistical methodologies in causal inference with targeted learning in chapter "Causal Inference with Targeted Learning for Producing and Evaluating Real-World Evidence", use of Estimand framework based on ICH E9 (R1) in RW setting along with examples in chapter "Estimand in Real-World Evidence Study: From Frameworks to Application", clinical studies leveraging RWD using propensity score-based methods in chapter "Clinical Studies Leveraging Real-World Data Using Propensity Score-based Methods", recent statistical development for comparative effectiveness research beyond propensity-score methods in chapter "Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods", innovative hybrid designs and analytical approaches leveraging RWD and Clinical Trial data in chapter "Innovative Hybrid Designs and Analytical Approaches Leveraging Real-World Data and Clinical Trial Data", statistical challenges for causal inference using time-to-event RWD in chapter "Statistical Challenges for Causal Inference Using Time-to-Event Real-World Data", sensitivity analyses for unmeasured confounding in chapter "Sensitivity Analyses for Unmeasured Confounding: This Is the Way", sensitivity analysis in the analysis of RWD when underlaying

assumptions addressing a research question are not met in chapter "Sensitivity Analysis in the Analysis of Real-World Data", personalized medicine with advance analytics in chapter "Personalized Medicine with Advanced Analytics", and use of RWE in HTA submissions in chapter "Use of Real-World Evidence in Health Technology Assessment Submissions".

Part IV contains three chapters. To promote uptake of RWE usage, practical examples will show the way. Chapter "Examples of Applying Causal-Inference Roadmap to Real-World Studies" demonstrates the application of causal-inference roadmap to RW studies via examples. Chapter "Applications Using Real-World Evidence to Accelerate Medical Product Development" presents six application examples where the regulatory contexts are summarized, whether the use of RWE/RWD is pivotal or supplemental for the regulatory decisions, assessment of regulatory quality data sources, statistical methods employed, settings where the approvals were obtained or denied, and any regulatory opinions for the submission and regulatory decision. Finally, chapter "The Use of Real-World Data to Support the Assessment of the Benefit and Risk of a Medicine to Treat Spinal Muscular Atrophy" details a case study where RWD is used to support the assessment of the benefit and risk of a medicine to treat spinal muscular atrophy.

We would like to express our sincerest gratitude to all the contributors who made this book possible. They are the leading experts in the use of RWE and RWD from industry, regulatory, and academia. Their in-depth discussions, thought-provoking considerations, deep knowledge in the field, and innovative approaches based on a wealth of experience make this book unique and valuable for a wide range of audiences. We are indebted to Donna Chernyk of Springer Nature for providing us with the opportunity for publication. Our immense thanks also go out to our families for their unfailing support and understanding of the many nights and weekends that we spent working on this book. Finally, the views expressed in this book are those of the authors and not necessarily reflective of the positions, policies, or practices of the authors' respective organizations.

Westfield, NJ, USA — Weili He, PhD
Lincolnshire, IL, USA — Yixin Fang, PhD
Vernon Hills, IL, USA — Hongwei Wang, PhD

Contents

Editors and Contributors

About the Editors

Weili He has over 25 years of experience working in the biopharmaceutical industry. She is currently a Distinguished Research Fellow and head of Medical Affairs and Health Technology Assessment Statistics at AbbVie. She has a PhD in Biostatistics. Weili's areas of expertise span across clinical trials, real-world studies and evidence generations, statistical methodologies in clinical trials, observational research, innovative adaptive designs, and benefit-risk assessment. She is the lead author or co-author of more than 60 peer-reviewed publications in statistics or medical journals and lead editor of two books on adaptive design and benefit-risk assessment, respectively. She is the co-founder and co-chair of the American Statistical Association (ASA) Biopharmaceutical Section's (BIOP) Real-world Evidence Scientific Working Group from 2018 to 2022. Weili is the BIOP Chair-Elect, Chair, and Past Chair from 2020 to 2022. She is also an Associate Editor of *Statistics in Biopharmaceutical Research* since 2014, and an elected Fellow of ASA since 2018.

Yixin Fang After he received his PhD in Statistics from Columbia University in 2006, Yixin Fang had been working in academia before he joined AbbVie in 2019. Currently, he is a Research Fellow and Director of Statistics in Medical Affairs and Health Technology Assessment Statistics (MA&HTA Statistics) at AbbVie. Within MA&HTA Statistics, he is Head of the therapeutics areas (TAs) of Eye Care and Specialty and Head of Causal Inference Center (CIC). In this role, he is involved with the design and analysis of Phase IV studies and real-world studies in medical affairs and leading HTA submissions in the TA of Eye Care. In addition, he is active in the statistical community with over 100 peer-reviewed manuscripts, and his research interests are in real-world data analysis, machine learning, and causal inference.

Hongwei Wang has close to 20 years' experience working in the biopharmaceutical industry. He is currently a Research Fellow and Director at Medical Affairs and Health Technology Assessment Statistics of AbbVie. Prior to that, Hongwei worked at Sanofi and Merck with increasing responsibilities. He has been leading evidence planning and evidence generation activities across various therapeutic areas in the fields of real-world studies, network meta-analysis, and post-hoc analysis with a mission to support medical affair strategy and optimal reimbursement. Hongwei received his PhD in Statistics from Rutgers University, conducts active methodology research and leads their application to different stage of drug development. He is co-author of about 40 manuscripts in peer reviewed journals and over 100 presentations at scientific congresses.

About the Contributors

Jie Chen is Chief Scientific Officer at ECR Global and Senior Vice President and head of Biometrics at Overland Pharmaceuticals. Jie has over 27 years of experience in biopharmaceutical R&D and his research interest spans from innovative clinical trial design and analysis, dose finding, medical product safety evaluation, multiple comparison and multiple testing, Bayesian methods, causal inference, use of real-world evidence in medical product development, and regulatory decision-making. He is an elected Fellow of the American Statistical Association.

Ruthanna Davi is a Statistician and Vice President of Data Science at Medidata and has a background in pharmaceutical clinical trials with more than 20 years working at the FDA, most recently as a Deputy Division Director in the Office of Biostatistics in the Center for Drug Evaluation and Research. Of late, her work is focused on developing analytical tools to improve the efficiency and rigor of clinical trials with special emphasis on creation and analysis of synthetic or external controls. She is a frequent speaker and author on this topic. Ruthie holds a PhD in Biostatistics from George Washington University.

Sai Dharmarajan is a Senior Mathematical Statistician in the Office of Biostatistics at Center for Drug Evaluation and Research, Food and Drug Administration (FDA). He joined the FDA in June 2018 after obtaining his PhD in Biostatistics from the University of Michigan. At FDA, he supports the review of clinical studies across different therapeutic areas and data sources (clinical trials and observational data). He also oversees various FDA-funded pharmacoepidemiology and safety surveillance research projects in claims and electronic health record databases. He has developed new methods and statistical software in the areas of causal inference, machine learning applications, and quantitative benefit-risk assessment. He is an active member of multiple Drug Information Association (DIA) and American Statistical Association (ASA) Biopharmaceutical Section working groups on benefit-risk and real-world evidence.

Muna El-Khairi is a Senior Statistical Scientist at Roche. She obtained an MSc in Statistics (Medical Statistics) from University College London in 2015, a PhD in Mathematics from Imperial College London in 2013, and an MSci in Mathematics from Imperial College London in 2007. Her research interests include RWE for drug development, Bayesian statistics, and the design and analysis of clinical trials, and she has published manuscripts in peer-reviewed journals.

Douglas Faries has a PhD in Statistics from Oklahoma State University and is currently a Senior Research Fellow in Real-World Analytics and Access at Eli Lilly and Company. In this role, Doug is involved with the design and analysis of observational research including comparative effectiveness, and he leads the development of real-world analytical capabilities for the business. He is active in the statistical community with over 150 peer-reviewed manuscripts, and his research interests are in causal inference and unmeasured confounding.

Dai Feng is currently a Director of the Medical Affairs and Health Technology Assessment Statistics at AbbVie. Before joining AbbVie, he worked at Merck with increasing responsibilities. Dai has over 14 years of pharmaceutical industry experience, ranging from early drug discovery to late clinical development and post-approval activities. His experiences span multiple therapeutic areas, including oncology, neuroscience, and immunology. Dai received his PhD in statistics from the University of Iowa. His research interests include Bayesian experiment design and data analysis, and statistical/machine learning. He has co-authored over 50 publications in statistical and medical journals and as book chapters.

Margaret Gamalo Statistics Head for Inflammation and Immunology in Pfizer Global Product Development. She combines expertise in biostatistics, regulatory and adult and pediatric drug development. Prior to joining Pfizer, she was Research Advisor at Global Statistical Sciences, Eli Lilly and Company and Mathematical Statistician at the Food and Drug Administration. Meg leads the Complex Innovative Design Task Force at the Biotechnology Innovation Organization. She also actively contributes to research topics within the European Forum for Good Clinical Practice – Children's Medicine Working Party. Meg is currently Editor-in-Chief of the *Journal of Biopharmaceutical Statistics* and is actively involved in many statistical activities in the American Statistical Association. She received her PhD in Statistics from The University of Pittsburgh.

Susan Gruber PhD, MPH, MS, is Founder and Principal of Putnam Data Sciences, a statistical consulting and data analytics consulting firm. Dr. Gruber's work focuses on the development and application of data-adaptive methodologies for improving the quality of evidence generated by observational and randomized health care studies. She was the PI on a multi-year FDA-funded project demonstrating a Targeted Learning framework for causal effect estimation using real-world data.

Lisa V. Hampson is based in the Advanced Methodology and Data Science group at Novartis in Switzerland, where her role is to support the development and implementation of innovative statistical methods. Prior to joining the pharmaceutical industry in 2016, Lisa was a Lecturer in Statistics at Lancaster University in the UK and held a UK Medical Research Council (MRC) Career Development Award in Biostatistics. Her research interests are in group sequential and adaptive clinical trials, Bayesian approaches for quantitative decision making, and the use of real-world evidence in clinical drug development.

Nigel Hughes has a 36-year career spanning the NHS in the UK (16 years), NGOs and patient organizations (10 years), and within the pharmaceutical industry (18 years). He has worked clinically in HIV and viral hepatitis, liver disease, and in sales and marketing, medical affairs, market access and health economics, R&D, precision medicine, advanced diagnostics, health IT, and Real-World Data/Real-World Medicine. His experience covers clinical, education, as an advisor, consulting, communications, and lobbying over the years. He is currently the Project Lead for the IMI2 European Health Data & Evidence Network (EHDEN) initiative and was Platform Co-Lead for the IMI1 European Medical Information Framework (EMIF), as well as provides consulting on numerous projects and programs in the domain of RWD/RWE.

Rima Izem is a Director in the Statistical Methodology group in Novartis where she supports use of best practice methodologies as well as development and implementation of novel statistical methods using real-world data or hybrid designs in all phases of clinical development and across therapeutic areas. Her research experience includes pioneering work in regulatory statistics using causal inference for comparative safety, signal detection, and survey research at the US FDA. Her methodological experience also includes comparative effectiveness in rare diseases at Children's National Research Institute and dimension reduction methods at Harvard University. Her experience with real-world data includes work with US insurance claims databases and electronic healthcare data, international registries, and electronic clinical outcome assessments.

Tae Hyun Jung PhD, is a Senior Statistical Reviewer in the Office of Biostatistics at Center for Drug Evaluation and Research, Food and Drug Administration (FDA). He has comprehensive New Drug Applications, Biologics License Applications, and Postmarketing Requirements review experience from phase III/phase IV safety focused reviews, and phase III efficacy reviews including Real-World Data/Evidence (RWD/E) efficacy. In addition, he works on reviews that leverage RWD from Centers for Medicare and Medicaid Services and leads RWE research using machine learning in causal inference and synthetic patient data. Tae Hyun joined the FDA in October 2017 after completing his PhD in Biostatistics from Yale University.

Dipak Kalra is President of The European Institute for Innovation through Health Data (www.i-hd.eu), a Professor of Health Informatics, and a former London gen-

eral practitioner. He plays a leading international role in Electronic Health Record R&D, including the reuse of EHRs for research. He has led the development of ISO standards on EHR interoperability, personal health records, and data protection. He participates in multiple EU Horizon 2020 and IMI projects including the generation of real-world evidence in pregnancy, the governance of patient-centric clinical trials, frameworks for the design and governance of mobile health programmes, scaling up the quality, interoperability and the reuse of health data for research including inputs to the European Health Data Space, scaling up of the collection and use of health outcomes towards more value-based care, and initiatives to explain the value of clinical research to the public.

Charles Lee is currently Executive Regulatory Science Director, Late CVRM at Astrazeneca. He oversees Regulatory science and strategy for the cardiovascular, renal, and metabolism disease areas. Charles represents Regulatory Affairs on CVRM governance committees for research, early development, late development, and medical affairs. He also leads several key Regulatory and Scientific initiatives, including RWE, digital endpoints, and novel endpoint qualification. Charles maintains line management responsibilities for a group of Regulatory Affairs Directors and Senior Directors based in the USA and Europe. Charles holds a BA in Biology from The Johns Hopkins University, MS in Microbiology from The University of Virginia, and MBA from Columbia Business School.

Hana Lee is a Senior Statistical Reviewer of the Office of Biostatistics in the CDER, FDA. She leads and oversees various FDA-funded projects intended to support development of the agency's RWE program including multiple Sentinel projects to develop causal inference framework for conducting non-randomized studies and to enhance analytic capacity using machine learning-based methods. She an FDA lead for RWE demonstration project on potential use of a Targeted Learning framework and Targeted Maximum Likelihood Estimation to support regulatory decision making. She has been taking a leadership role for RWE scientific working group of the American Statistical Association (ASA) Biopharmaceutical Section since 2020.

Heng Li is a Lead Mathematical Statistician in the Division of Biostatistics, CDRH/FDA. He is instrumental in the establishment and evaluation of policies and procedures to maintain a unified and consistent program for biostatistical aspects of the regulation of medical device. He leads a team of statisticians in technical reviews and research. He has made substantial contributions in advancing regulatory statistics, especially propensity score-based methods, including developing and implementing the novel propensity score-integrated approaches for leveraging real-world evidence to support regulatory decision-making. He currently serves as an Associate Editor for *Pharmaceutical Statistics*. He is a recipient of the FDA Scientific Achievement Awards – Excellence in Analytical Science.

Yingyi Liu is a manager from Medical Affairs and Health Technology Assessment Statistics at AbbVie. In this role, Yingyi provides statistical leadership in Health Technology Assessment (HTA) submissions supporting market access and reimbursement and supports all key Medical Affairs business activities. Her current research interests include causal inference, indirect treatment comparison, and real-world evidence research. Yingyi obtained her PhD degree in Quantitative Economics and Econometrics and Master's degree in Statistics from the University of Illinois at Urbana-Champaign. Prior to joining AbbVie, she has worked at Morgan Stanley and U.S. Bank as quantitative model developer employing machine learning techniques for predictive modeling.

Jingyu Julia Luan is currently a Senior Director of Global Regulatory Affairs in AstraZeneca, leading global drug development, regulatory strategies, and regulatory execution. She also leads the regulatory initiatives in Real-World Data/Real-World Evidence and drives external engagement with senior Health Authority officials and Industry Groups globally. Dr. Luan is a frequent speaker, organizer, and chair for international conferences, workshops, and forums. Prior to AstraZeneca, Dr. Luan worked at the US FDA for 13+ years and held various positions of increasing responsibilities, including Statistical Reviewer, Team Leader, and Acting Deputy Division Director. She received more than ten FDA honor awards. Before the FDA, she was a member of the research faculty at Johns Hopkins University and a Statistical Consultant at the University of Kentucky Medical Center. Dr. Luan is the President-elect (2022-2023) and a board member of Chinese Biopharmaceutical Association, and a board member and committee co-chair of FDA Alumni Association.

Julia Ma is currently the therapeutic area lead of neuroscience within the Medical Affairs and Health Technology Assessment (MA&HTA) Statistics group at AbbVie. She had a brief stint in finance before settling in the pharmaceutical industry in 2007. During her more than 15 years of industry experience, Julia held various positions at Bristol Myers Squib, Eli Lilly, Allergan, and then AbbVie through the merger in support of many different therapeutic areas including immunology, oncology, neuroscience, urology, ophthalmology, and women's health. Her experience and knowledge span all different stages of drug development as well as market access and reimbursement of pharmaceuticals. Julia received a PhD from West Virginia University majoring in Statistics and minoring in Mathematics and Computer Science. Her research interests include machine learning, psychometric evaluation of patient-reported outcomes, analytical strategies, and methodologies in HTA submissions.

Tammy McIver is a Senior Principal Statistical Scientist at Roche. Tammy obtained an MSc in Applied Statistics from the University of Greenwich in 1998 and became a Chartered Statistician in 2004. She has 20 years' experience in clinical drug development across multiple therapeutic areas. During this time she has had full responsibility for statistical activities on several new medicines resulting in

approvals and market access. Most recently she was the Project Lead Statistician for risidplam (EVRYSDI®), currently approved in 90 countries worldwide for the treatment of Spinal Muscular Atrophy. Her research interests include RWE for drug development and patient reported outcomes. She has published over 20 manuscripts in peer-reviewed journals.

Herbert Pang is an Expert Statistical Scientist at Roche/Genentech. Herb obtained his MBA from HKUST in 2016, PhD in Biostatistics from Yale in 2008, and BA in Mathematics and Computer Science from Oxford in 2002. He was formerly a tenured Associate Professor at the University of Hong Kong (HKU). His research interests include RWE for drug development, machine learning, biomarker discovery, and the design and analysis of clinical trials. Herb is currently a PI on an FDA grant under the NIH U01 mechanism on RWE. He is also an Editor of the ASA Biopharmaceutical Report in 2022 and a member of the ASA Biopharmaceutical section RWE working group. He remains as an Honorary Associate Professor at HKU and is also an adjunct faculty in the Department of Biostatistics and Bioinformatics at Duke. He has published over 130 methodological and translational peer-reviewed research articles.

Rachael Phillips MS, is a PhD candidate in biostatistics at UC Berkeley. Her research on Targeted Learning and causal inference focuses on semi-parametric statistical estimation and inference. She is a contributor to the TLverse, a repository of software implementing targeted minimum loss-based estimation, the highly adaptive lasso, and super learning. Ms. Phillips was a key member of the team on an FDA-funded project on Targeted Learning for developing real-world evidence. She is also co-developing a personalized online super learner for intensive care units.

Ram Tiwari is the Head of Statistical Methodology at the BMS, New Jersey, since 2021. In his position, Ram is responsible for promoting the use of novel statistical methods and innovative clinical study designs in the drug development. Prior to joining BMS, Ram served 20+ years in the Federal Government as Director of Biostatistics, Center for Devices and Radiological Health, FDA; Associate Director for Science and Policy in the Office of Biostatistics, Center for Drug Evaluation and Research, FDA; and Mathematical Statistician and Program Director at National Cancer Center, NIH. He also spent over 20 years in academia serving as Professor and Chair of the Department of Mathematics at University of North Carolina at Charlotte. Ram received his MS and PhD in Mathematical Statistics from the Florida State University. He is a Fellow of the American Statistical Association, and a past President of the International Indian Statistical Association. Ram has published over 200 papers, and a book on *Signal Detection for Medical Scientists: Likelihood Ratio Test-Based Methodology* by Chapman and Hall/CRC, 2021.

Mark van der Laan PhD, is the Jiann-Ping Hsu/Karl E. Peace Professor in Biostatistics and Statistics at UC Berkeley, where he developed Targeted Learning, targeted minimum loss-based estimation (TMLE), and Super Learning. He is a

founding editor of the *Journal of Causal Inference* and has published numerous articles on the application of Targeted Learning to develop robust real-world evidence. Dr. van der Laan has received many honors for his work, including the COPSS Presidents' Award, the Mortimer Spiegelman Award, and the van Dantzig Award.

Jixian Wang is a statistician in Statistics Methodology group at BMS in Boudry, Switzerland, with over 20 years' pharmaceutical industry experience. Previously, he worked in Novartis and GSK, after spending several years in universities as an academic researcher. His research interests include clinical trial design, epidemiology and drug safety, health technology assessment, and clinical pharmacology. Recently, he has been working on using machine learning methods for adaptive trials, Bayesian approaches to borrowing external controls, and causal inference in analyses using real-world data. He has published about 60 papers and a book on exposure-response modeling by CRC.

Ying Wu is an Associate Professor in the Department of Biostatistics at Southern Medical University in China. Ying has considerable expertise in both theoretical and applied biostatistics, especially in the areas of causal inference and bias analysis for clinical studies. Ying is currently leading a four-year research project funded by the National Natural Science Foundation of China to develop trial design methods and causal inference methods for utilizing external real-world data in hybrid-arm trials to support regulatory decision-making. Ying has been involved in the drafting of dozens of statistical guidelines organized by the Chinese NMPA, including *the Guidance on Using Real-World Evidence to Support Drug Development and Regulatory Evaluation*.

Wai Yin Yeung is a Senior Statistical Scientist at Roche. She was formerly a Research Associate at Lancaster University under the Roche Postdoctoral Fellowship program. She obtained a PhD in Statistics from Queen Mary, University of London in 2013, MSc in Pure Mathematics from Imperial College London in 2006, and BSc in Mathematics and Statistics from Queen Mary, University of London in 2005. Her research interests include RWE for drug development, statistical methodologies for dose-finding, treatment allocation, randomization, and the design and analysis of clinical trials. She has published manuscripts in peer-reviewed journals and has served as a reviewer for journals in medical statistics.

Lilly Q. Yue is Deputy Director, Division of Biostatistics, CDRH/FDA, and has played a key leadership role in advancing regulatory statistics, including a pioneering effort on adapting and advancing propensity score methodology for premarket observational studies and on developing and implementing the novel propensity score-integrated approaches for leveraging real-world evidence to support regulatory decision-making. She is an Editor-in-Chief of *Pharmaceutical Statistics* and served as an Associate Editor (and a Guest Editor) for *Journal of Biopharmaceutical Statistics and Pharmaceutical Statistics*. She is a recipient of

the FDA Scientific Achievement Awards – Excellence in Analytical Science, and a Fellow of the American Statistical Association.

Hongtao Zhang is a biostatistician in Merck's biostatistics methodology research group. Previously, he worked in the statistical innovation group at AbbVie, and early phase biostatistics group at BMS-Celgene as biostatistics lead. He obtained his PhD in Biostatistics from the University of North Carolina at Chapel Hill in 2015. Hongtao's research interests include oncology dose-finding, historical/external control borrowing, dose-response modeling, adaptive designs and survival analysis. He also enjoys statistical consulting and software development. He has facilitated the adoption of many novel statistical methods in his current and previous positions.

Zuoyi Zhang is an Associate Director at Medical Affairs and Health Technology Assessment-Statistics, Data and Statistical Sciences at AbbVie Inc. He received his PhD in Statistics from Purdue University. Prior to joining AbbVie, he worked in the Department of Biostatistics and Health Data Science at Indiana University as Assistant Research Professor. He has more than a decade of experience in the research of the real-world evidence. Dr. Zhang's research focuses on real-world evidence studies using large databases such as electronic health records and claims data. He is specialized in comparative effectiveness research, analysis of observational studies, high-dimensional data, machine learning, and statistical computing. Dr. Zhang is especially interested in evaluating the causal effect of treatments using causal inference methodologies.

Tianyu Zhan is a statistician of Data and Statistical Sciences at AbbVie Inc. He is currently leading several study and compound level activities in Dermatology group and has experience in designing innovative clinical trials in Immunology, Oncology, and Virology. Tianyu has research interests in adaptive clinical trials, Bayesian analysis, machine learning, missing data, multiplicity control, real-world evidence, and survival analysis. He is also an active journal referee and has organized several invited sessions at statistical conferences. Tianyu received his PhD in Biostatistics from the University of Michigan, Ann Arbor in 2017.

Part I
Real-World Data and Evidence to Accelerate Medical Product Development

The Need for Real-World Evidence in Medical Product Development and Future Directions

Weili He, Yixin Fang, Hongwei Wang, and Charles Lee

1 Introduction

Randomized controlled clinical trials (RCTs) have been the gold standard for the evaluation of efficacy and safety of medical interventions. However, the costs, duration, practicality, and limited generalizability have incentivized many to look for alternative ways to optimize it. In recent years, we have seen an increasing usage of real-world data (RWD) and real-world evidence (RWE) in clinical development and life-cycle management. Especially encouraged by legislations and guidance released by regulators and special interest groups in recent years, sponsors have been actively seeking guidance and application use cases. In 2016, the twenty-first Century Cures Act was signed into law [1]. It is designed to help accelerate medical product development and bring new innovations and advances to patients who need them faster and more efficiently. The Food and Drug Administration (FDA) PDUFA (Prescription Drug User Fee Act) VI, released in 2017 for fiscal years 2018–2022, enhances FDA's ability to consider the possibilities of using "real world" (RW) data as an important tool in evaluating drug safety and efficacy [2].

In December 2018, FDA released an FDA's RWE Framework (henceforth called Framework) [3]. The Framework defines RWD as "data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources," and RWE as "the clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD." Examples of RWD in the Framework include data derived from electronic health records (EHR), medical

W. He (✉) · Y. Fang · H. Wang
Medical Affairs and Health Technology Assessment Statistics, Data and Statistical Sciences, AbbVie, North Chicago, IL, USA
e-mail: weili.he@abbvie.com

C. Lee
CVRM Regulatory Affairs, AstraZeneca, Gaithersburg, MD, USA

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_1

claims and billing data, data from product and disease registries, patient-generated data and data from other sources, such as mobile devices. The Framework further indicates that RWD sources can be used for data collection and to develop analysis infrastructure to support many types of study designs to develop RWE, including, but not limited to, randomized trials (e.g., large simple trials, pragmatic clinical trials) and observational studies (prospective or retrospective).

More recently, the PDUFA VII Commitment letter for fiscal years 2023 through 2027 [4] provided further details on the FDA RWE program and indicated the following key aspects:

(a) By no later than December 31, 2022, FDA will establish and communicate publicly a pilot Advancing RWE Program.
(b) The Advancing RWE Program will include, but not be limited to, a list of activities and components, some of which include (1) FDA will solicit applications for RWE programs; (2) FDA will use structured review process to evaluate and rank applications; (3) FDA will accept one to two eligible and appropriate proposals each cycle, and several additional activities FDA will convene following the solicitation and application.
(c) By no later than June 30, 2024, FDA will report aggregate and anonymized information, on at least an annual basis and based on available sources (e.g., information provided by review divisions), describing RWE submissions to CDER and CBER.
(d) By no later than December 31, 2025, FDA will convene a public workshop or meeting to discuss RWE case studies with a particular focus on approaches for generating RWE that can potentially meet regulatory requirements in support of labeling for effectiveness.
(e) By no later than December 31, 2026, experience gained with the Advancing RWE Program, as well as CDER's and CBER's RWE program in general, will be used to update existing RWE-related guidance documents or generate new draft guidance, as appropriate.

Chapter "Overview of Current RWE/RWD Landscape" provides more in-depth information on the regulatory guidance documents in recent years in key regions around the world. Further, there have also been increasing public and private collaborations in RWE research. Examples include the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and International Society for Pharmaceutical Engineering (ISPE) special joint task force on "good practices for RWD studies of treatment and/or comparative effectiveness (CER)" [5], and "reporting to improve reproducibility and facilitate validity assessment for healthcare database studies" [6]. Launched in October 2013, the GetReal was a three-year project of the Innovative Medicines Initiative, a Europe's largest public-private consortium consisting of pharmaceutical companies, academia, Health Technology Assessment (HTA) agencies and regulators, patient organizations, and subject matter experts (SMEs). The efforts resulted in numerous publications including delivery of four work packages [7]. Within the statistical community in the United States, the American Statistical Association (ASA) Biopharmaceutical Section (BIOP) sponsored an RWE Scientific Working Group (SWG) that started in April

2018. The primary goal of the group is to advance the understanding of the RWE research in a precompetitive space, and the membership consists of members from FDA, academia, and industry. The group has produced or submitted six peer-reviewed publications:

- *The Current Landscape in Biostatistics of the use of Real-World Data and Evidence for Medical Product Development: General Considerations* [8]
- *The Current Landscape in Biostatistics of Real-World Data and Evidence: Clinical Study Design and Analysis* [9]
- *The Current Landscape in Biostatistics of Real-World Data and Evidence: Causal Inference Frameworks for Study Design and Analysis* [10]
- *Estimands – From Concepts to Applications in Real-World Setting* [11]
- *Statistical Consideration for Fit-For-Use Real-World Data to Support Regulatory Decision Making in Drug Development* [12]
- *Examples of Applying Causal Inference Roadmap to RWE Clinical Studies* [13]

With the encouragement from regulators and available guidance and literature on the use of RWD and RWE in recent years, we have seen an increased uptake of RWE in various stages of drug development. Figure 1, which is adapted from the figure in [14], depicts the various uses in different stages of drug development and their reliance on RWD in representative types of study design. Together with guidance in the Framework on the usage of RWD, we summarize key usages of RWD in clinical development and life-cycle management as follows, but the list is by no means exhaustive:

- Generate hypothesis for testing in RCTs.
- Identify investigators who provide care for patients with the disease or condition of interest, thereby selecting study sites with appropriate investigators.

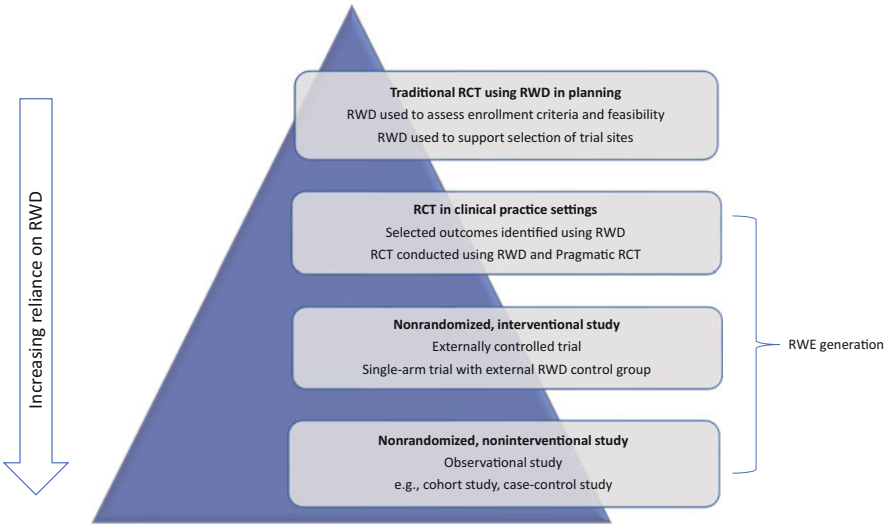

Fig. 1 RWE spectrum

- Assess disease prevalence and sub-population of patients identified by phenotype or genotype, thus assisting with patient selection and enrollment.
- Assess the prevalence of concomitant medications for a disease, along with prevalence of comorbidities of the disease.
- Evaluate biomarker prevalence and discover target for the development of personalized medicine.
- Evaluate indication calibration by assessing unmet medical need and whether the need is consistent across the targeted population.
- Identify outcome measures by ascertaining background event rate in a disease and related population for incidence, duration, severity.
- Fulfil regulatory safety commitment, safety surveillance, and safety label update.
- Enroll patients at point of care and leverage existing RWD, such as EHR and administrative claims, to retrieve historic information and long-term follow-up, thereby employing a so-called hybrid study design to use both existing RWD and prospectively collecting additional study information.
- Use RWD to build external control cohort for single-arm clinical studies or augment concurrent control group of an RCT.
- Describe patient journey, treatment pattern, healthcare utilization to assess unmet medical needs and disease burden and facilitate choice of comparator.

With the above delineation, the use of RWE could lead to support approval of new molecular entities or biologics, accelerate or seek conditional approval, explore new indication or new population, make changes to dosing administration, supplement RCTs information for a regulatory submission, or provide complementary evidence for comparative effectiveness and cost-effectiveness assessment for reimbursement decisions in HTA.

For the rest of the chapter, in Sect. 2, we review the progress to date on the uptake of RWE. Even with these recent progresses, challenges remain. We interpret these challenges as opportunities for further research and development, as described in Sect. 3. The final section provides discussions on future directions and concluding remarks.

2 Where We Are Now with the Use of RWE and RWD

In the last few years, a great stride has been made in advancing the uptake of RWE in drug development. The prevailing environment from regulators is that of encouragement and guidance, along with concrete action plan, as shown by the PDUFA VII Commitment letter for fiscal years 2023 through 2027 [4]. In this section, we will discuss a few major advances as we observed in recent years, focusing primarily on the advancement as described in this book. However, there have been numerous literatures on RWE- or RWD-related publications, such as the work by ASA BIOP RWE SWG, and guidance and publications by regulators and interest groups. It should be noted that the discussion here is by no means thorough, and any further gaps remain to be filled by further observations and research.

2.1 Regulatory Advancement

In the regulatory arena, in rapid succession, FDA released four draft guidance in late 2021 on RWD on assessment EHR, medical claims, and registry data to support regulatory decisions in two guidance documents; data standard for regulatory submission is also delineated in another guidance, along with considerations for the use of RWD and RWE for drug and biologic products in the fourth draft guidance. In October 2021, EMA adopted guideline on registry-base studies. Chapter "Overview of Current RWE/RWD Landscape" provides a good coverage of guidance documents related to RWE from the United States, Japan, China, and the United Kingdom, and discussed similarities and differences between them. The European Medicines Agency (EMA)'s RWE Vision is that, by 2025, the use of RWE will have been enabled and the value will have been established across the spectrum of regulatory use cases [15]. In 2022, EMA established a Coordination Centre for the Data Analysis and Real World Interrogation Network (DARWIN EU®) [16]. DARWIN EU will deliver RWE from across Europe on diseases, populations, and the uses and performance of medicines.

Health Authorities responsible for reimbursement and pricing reviews in HTA submissions have also released draft guidance documents on the use of RWE for HTA submissions. The National Institute for Health and Care Excellence (NICE), the United Kingdom's HTA body, released NICE RWE framework in June 2022 [17]. The key message is that RWD can improve our understanding of health and social care delivery, patient health and experience, and the effects of interventions on patient and system outcomes in routine settings. As described in NICE strategy 2021–2026 [18], NICE wanted to use RWD to resolve gaps in knowledge and drive forward access to innovations for patients. In the rest of the world, French National Authority for Health (HAS) released a guidance in June 2021 on RW studies for the assessment of medicinal products and medical devices [19]. Further, due to limited recommendations to support the appropriate use of RWE, a group of experts from top European Union (EU) academic institutions and HTA bodies in eight countries as part of the EU's Horizon 2020 IMPACT-HTA program published a white paper on the use of nonrandomized evidence to estimate treatment effects in HTA [20]. The key messages are:

- RWE must be relevant for the research question.
- They recommended strategies to study design and analysis.
- The white paper deemed transparency as essential.
- The paper also recommended strengthening infrastructure and investing in resources to design, analyze, and interpret RWE.

Chapter "Use of Real-World Evidence in Health Technology Assessment Submissions" of this book provides more details on the use of RWE in HTA submissions. In summary, these various guidance documents all provided a similar message on fit-for-purpose use of RWE and RWD.

2.2 Advancement in Operational Considerations

Several chapters in this book covered operational considerations in implementation in the use of RWE.

In the RW studies (RWS), key variables such as exposure, treatment, outcome, disease status, or confounders may not be captured in one place, it is therefore important to ascertain these key variables using advance analytics, such as machine learning and nature language processing. Misclassification is also a concern, requiring validations. Chapter "Key Variables Ascertainment and Validation in Real-World Setting" of this book covers these topics and walks through an example study for which the ascertainment of key variables was found to be acceptable from a regulatory standpoint. Once the key variables are in place for an RWD source, assessment of fit-for-use RWD sources is a critical step in the determination of whether an RWD source could be used. Chapter "Assessment of Fit-for-Use Real-World Data Sources and Applications" provides guiding principles in the fit-for-use RWD assessment and illustrates assessment steps with an application. The authors drill down into details on the factors to consider specific to a research question and disease condition and provides sufficient details to allow practitioners to follow in their applications.

The role of health data and interoperability standards is another important element to consider in their harmonization, since lack of harmonization and common data standards would impede the foundation for a vision to achieve large-scale interoperability in supporting technical, methodological, and evidence generation, based on emerging trends. Chapter "Data Standards and Platform Interoperability" presents a discussion on the need for Findable, Accessible, Interoperable, and Reusable (FAIR) data, and the role data standards, in particular those emerging as leading with regards to regulatory decision, and emerging platforms for network, at-scale evidence generation, as unified visions for standards and platforms. It is often of great interest to aggregate and link data from several RWD sources to provide a more comprehensive longitudinal evaluation of treatments from different aspects. Chapter "Privacy-Preserving Data Linkage for Real-World Datasets" reviews privacy framework and different methods in linking data sources, while focusing on patient privacy protection, data pre-processing, linkage, and performance evaluations.

2.3 Advancement in Statistical Methodologies in Causal Inference

Tremendous progress has been made not only in the methodologies of causal inference but also in the applications of these methods in the uptake of RWE generations. Chapter "Causal Inference with Targeted Learning for Producing and Evaluating Real-World Evidence" summarizes a Target Learning roadmap as a

systematic guide to navigate the study design and analysis challenges inherent in real-world studies. ICH E9 (R1) Addendum [21] presents a structured framework to strengthen the dialogue between disciplines involved in the formulation of clinical trial objectives, design, conduct, analysis and interpretation, as well as between sponsor and regulator regarding the treatment effect of interest that a clinical trial should address. Further, the guidance indicates that "The principles outlined in this addendum are relevant whenever a treatment effect is estimated or a hypothesis related to a treatment effect is tested, whether related to efficacy or safety. While the main focus is on randomized clinical trials, the principles are also applicable for single-arm trials and observational studies." Chapter "Framework and Examples of Estimands in Real-World Studies" presents principles for the Estimand Framework for use in RW setting, highlights similarities and differences between RCTs and RWS, and provides a roadmap for choosing appropriate estimand for RWS.

Chapter "Clinical Studies Leveraging Real-World Data Using Propensity Score-Based Methods" provides a comprehensive summary of propensity score-based methods (PSM) to minimize confounding biases in clinical studies leveraging RWD sources. Beyond PSM, chapter "Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods" presents recent statistical developments for comparative effectiveness research using methods, such as G-methods. Chapter "Innovative Hybrid Designs and Analytical Approaches leveraging Real-Word Data and Clinical Trial Data" showcases an innovative hybrid design and analytical approaches leveraging RWD and clinical trial data, while chapter "Statistical Challenges for Causal Inference Using Time-to-Event Real-World Data" highlights statistical challenges for causal inference using time to event RWD. As we know, the lack of randomization in RWD brings the potential for bias into any comparisons between groups or interventions of interest. Commonly used methods such as PSM can account only for confounding variables that are included in the analysis database, but any confounders not contained in the database are 'unmeasured confounders' and may result in a biased treatment effect estimate. Chapter "Sensitivity Analyses for Unmeasured Confounding: This is the Way" focuses on the challenging case of comparative analyses based on RWD and the issue of unmeasured confounding. Further, ICH E9 (R1) discusses the importance of sensitivity analysis [21]. Chapter "Sensitivity Analysis in the Analysis of Real-World Data" guides readers on how to conduct sensitivity analysis to explore the robustness of inference to deviations from the underlying assumptions.

The practice of modern medicine demands personalized medicine (PM) to improve both quality of care and efficiency of the healthcare system. Chapter "Personalized Medicine with Advanced Analytics" dives into the application of advanced analytics in addressing PM research questions, while chapter "Use of Real-World Evidence in Health Technology Assessment Submissions" covers the utility and strengths of well-developed RWE in HTA decision-making in major regions around the world.

2.4 Advancement in Real Case Applications

The concept of causal inference framework and use of causal inference roadmap is crucial in the use of RWD to generate robust RWE. Chapter "Examples of Applying Causal-Inference Roadmap to Real-World Studies" describes a few examples of applying causal inference roadmap to RWSs. Chapter "Applications Using Real-World Evidence to Accelerate Medical Product Development" summarizes six case studies that regulatory agencies considered in recent years in the use of RWE/RWD for regulatory decisions. Some of these use cases succeeded in achieving positive regulatory decisions, while a couple of others didn't meet the principle of adequate and well-controlled study for evidentiary standard. This chapter includes rich details on the analysis of each case study. Finally, chapter "The Use of Real-World Data to Support the Assessment of the Benefit and Risk of a Medicine to Treat Spinal Muscular Atrophy" presents a detailed case study in Spinal muscular atrophy (SMA) and describes how RWD from publications and individual patient data were used to support the development of risdiplam, a medicine to treat SMA.

3 Opportunities for Further Advancement

3.1 Regulatory Context

With the release of numerous regulatory guidance documents in recent years from regions around the world, there is a prevailing need to share more use cases. Through use case studies, practitioners could understand better the regulatory contexts, key regulatory review issues, whether the use of RWE/RWD is pivotal or supplemental for the regulatory decisions, assessment of fit-for-use data sources, statistical methods employed, and whether substantial evidence of effectiveness as stated in Regulations 21CFR314.126 is met for a specific case study. With the encouragement from the Framework and more emerging literature on the changing landscape of regulatory approval processes and case examples as delineated in chapters "Applications Using Real-World Evidence to Accelerate Medical Product Development" and "The Use of Real-World Data to Support the Assessment of the Benefit and Risk of a Medicine to Treat Spinal Muscular Atrophy", we believe that we will see more and more such use cases in the coming years. Further, through a feedback loop between sponsors and regulators, existing RWE-related guidance documents could be updated, or new draft guidance could be developed.

Considering the evolving and diverse regulatory frameworks across jurisdictions, sponsors are encouraged to engage with regulatory agencies and other stakeholders, ideally through joint scientific advice procedures, when applicable, such as EMA/FDA parallel scientific advice [22]. Further, the use of RWE for regulatory submissions and decisions is still relatively new. The FDA draft guidance on data standards for drug and biologic products submissions containing RWD provide

guidance on data standards and data mapping, along with the development of review guide for such submissions [23]. It's helpful for sponsors to gain further experience in these areas and engage regulators for further advice as needed.

3.2 Clinical Context

Up until just a few years ago, RWE has been used primarily to perform post-marketing surveillance to monitor drug safety and detect adverse events or in HTA submissions to understand disease burden, drug effectiveness, or economic modeling. To expand the use to support clinical development and life-cycle management, it is important to consider the clinical contexts regarding the clinical question of interest and whether RWS that generate RWE are sufficient and robust enough for the regulatory question at hand. We believe that RW studies should not be used as a replacement for RCTs, since all the design precautions and/or statistical techniques could still not overcome unquantifiable or poorly recorded data inherent with RWD. However, if used appropriately, RWE could be used to support regulatory decisions in certain situations.

The PRECIS-2 tool [24] is a refined tool of PRECIS (Pragmatic Explanatory Continuum Indicator Summaries) that was intended to help trialists make design decisions consistent with the intended purpose of their trial. PRECIS-2 tool contains nine domains – eligibility criteria, recruitment, setting, organization, flexibility (delivery), flexibility (adherence), follow-up, primary outcome, and primary analysis, scored from 1 (very explanatory) to 5 (very pragmatic). The authors argued that although we often refer to trials as in ideal RCT setting or in RWS, there is no simple threshold to separate the two concepts. Rather than a dichotomy, there is a continuum between the two, by adjusting the factors in either design or study conduct to make a trial more RCT or RWS. Undeniably, clinical context is critically important in determining whether the aim is to answer the question, "Can this intervention work under ideal considerations?" or "Does this intervention work under usual conditions?" The internal and external validity and generalizability can be inferenced from such considerations.

3.3 Study Design and Analysis Context

Chapter "Key Considerations in Forming Research Questions" reviews and identifies key elements of forming sound research questions in RWS. The PROTECT criteria proposed in [25] is discussed in-depth in chapter "Key Considerations in Forming Research Questions". Further, the authors propose a roadmap for revising a research question and/or element of the PROTECT criteria if a question cannot be answered as framed. The authors' way of setting up right research questions is

quite innovative, as they use Estimand framework as "touchstone" to gauge whether a question can be answered or not.

There has been a flurry of literature on the statistical methodologies in analyzing RWD and translating data into robust RWE. Given that rich literature exists on statistical methodologies to handle potential biases and confounding with the use of RWD [9], methodologies are discussed in several chapters in Part III of this book. We believe that it's important for practitioners to understand these approaches, especially sensitivity analysis, to assess the robustness of the findings and apply them appropriately in their RWE projects. We would also like to provide some cautions in methodology development. While many RW study design and/or methodologies have been proposed, some of them might be more of an intellectual interest with less appeal for practical applications. Thus, focusing on those adaptations that are practically feasible will result in the most successful implementations as the research enterprise is collectively gaining experience with this new and evolving field.

3.4 Data Context

Chapter "Assessment of Fit-for-Use Real-World Data Sources and Applications" of this book provides guiding principles for assessing fit-for-use RWD sources in data relevancy and reliability. The authors also illustrate the approach via a hypothetical example. However, further research may still be needed since the actual assessment is very much disease and research question-specific. Further, it may be a good idea for sponsors to engage regulators for discussions on the data source, and rationale and justification on the fit-for-use assessment. As EMA/HMA calls for in [16], it is important to establish and expand catalogues of observational data sources for use in medicines regulation, provide sources of high-quality, validated RWD on the uses for safe and effective medicines, and address specific questions by carrying out high-quality, non-interventional studies, including developing scientific protocols, interrogating relevant data sources, and interpreting and reporting study results. In terms of data sources, technological advancements in health technology and digital wearable devices will become potential sources of RWD. The key to their application in RWE is to ensure that the data generated is of high quality and fit for purpose.

We believe that it will be ideal to establish an industry standard for how an RWD source should be assessed and what criteria constitute a fit-for-use database.

3.5 Governance and Infrastructure Context

In addition to challenges as mentioned previously in this section, there are also additional challenges from resource, logistic, operational, and organizational per-

spectives. Utilization of RWD and RWE involves cross-functional expertise and collaboration, so building these features into an organization's processes, systems, and culture is a prerequisite for uptake. An upfront investment in dedicated resources may be needed, such as building or updating processes in clinical development procedures, developing templates of brand development plans and tools, and providing education on RWD sources and RWE methodologies. Change management may be needed to overcome entrenched decision-making processes that are skeptical about the use of RWE.

Especially, we recommend setting up governance to oversee the data acquisition and usage; developing processes and procedures that facilitate the regulators' requirements for transparency, pre-specification, consistency, reproducibility, and compliance in RWE applications; understanding existing RWD sources and properties and data owner networks; developing data platform to facilitate data flow, data harmonization from diverse sources, and connectivity for research use; and building analytic platform with powerful computational capacity for big data processing and re-usable analytic tools along with centralized coding library to define disease cohorts, exposures, outcome measures, and confounders in a consistent manner.

4 Future Direction and Concluding Remarks

In the past 5 years, we have seen growing international interest among all healthcare stakeholders regarding how to best approach the uptake of RWE and RWD to revolutionize the drug development process. The robust legacy of scientific groundwork as described in this book and regulatory guidance and other literature in recent years has paved the way to the future. What will be the challenges and opportunities for the uptake of RWE over the next 5 years?

In Sect. 3, we discuss opportunities for further advancement in the uptake of RWE from different areas of focus. While the common elements for further advancement have been identified, we expect that the next 5 years will see refinement in the use of specific tools and techniques by regulators around the world. Some agencies may focus on data quality and data platforms, while others may explore novel approaches integrating different sources of RWD for use and further refine guidance documents. Medicine development is a global endeavor. Sponsors therefore will seek a consistent degree of process predictability across target jurisdiction regulatory agencies, and this can come from the use of globally acceptable, standardized, systematic approach to RWE, irrespective of the specific tools and methodologies that each employ in support of its regulatory decisions.

As we have seen more and more collaborations across regions in the world, such as EMA/FDA parallel scientific advice [22] and EUnetHTA21 [26] for an effective and sustainable network for HTA across Europe, a structure process can facilitate work sharing and the potential for joint reviews and improve information sharing with industry partners and other stakeholders. These types of collaborations could also provide a clearer understanding of rationales for different marketing and

labeling decisions in different jurisdictions, such as clinical context and the practice of medicine, and alignment of risk management plans. The next 5 years will also see the growing uptake of RWE in regulatory decisions. Whether interacting with regulatory or HTA agencies, establishing a dialogue with the stakeholders early during medicine development can contribute to effective, ongoing communications with a more consistent understanding and implementation of the expectations from each stakeholder.

As RWE becomes the new information currency in healthcare, decision makers will be challenged using these new types of data sources. Over the next 3–5 years for some therapeutic areas, such as oncology or rare diseases, there may be a shift to the use of integrating RWD into phase II or phase III clinical studies. As development progresses, RWE will enhance the understanding of the product's safety profile and will be used to confirm clinical efficacy and RW effectiveness.

Of course, the use of RWE in drug development will not be without challenges. Great progress has been made on the methodologies to assess the robustness and uncertainty around factors that confound the interpretation of RWE. Further refinement and new methodology development may be called for based on the use cases. RWE collection will need to encompass a global view or, at the least, focus on key markets and jurisdiction experiences. Building a federate model and platform for data and analytic tools sharing may facilitate further leapfrogging in the field. Building high quality RWD and making them widely available may call for standardization of data, such as the use of common data model. Transparency, pre-specification, consistency, documentation, and reproducibility will be the cornerstone to which current and new facilitated regulatory pathways that are designed to accelerate submissions, reviews, and patient access to medicines for serious diseases where there is an unmet medical need will likely be accepted. These new pathways, such as Breakthrough Therapy and Accelerated Approvals in the United States, or Conditional Marketing Authorization in the European Union, may increase the communications and level of commitment between the sponsors and the agencies.

Finally, we want to emphasize that these opportunities to incorporate RWE in drug development should be used with care. The last thing we want to do is to treat opportunities offered haphazardly, which will result in rejected submissions and lead to mistrust in the use. The latter will delay broad acceptance of properly designed and executed studies and submissions incorporating RWE.

References

1. US Congress, "21st Century Cures Act. H.R. 34, 114th Congress," 2016. https://www.gpo.gov/fdsys/pkg/BILLS-114hr34enr/pdf/BILLS-114hr34enr.pdf (accessed Jan. 26, 2021).
2. J. Darrow, J. Avorn, and A. Kesselheim, "Speed, Safety, and Industry Funding-From PDUFA I to PDUFA VI," *N. Engl. J. Med.*, vol. 377, no. 23, pp. 2278–2286, 2017.
3. FDA, "Framework for FDA's Real-world Evidence Program," 2018. https://www.fda.gov/media/120060/download (accessed Nov. 03, 2019).

4. FDA PDUFA VII, "Commitment Letter," 2022. https://www.fda.gov/industry/prescription-drug-user-fee-amendments/pdufa-vii-fiscal-years-2023-2027
5. M. Berger, H. Sox, R.J. Willke, D.L. Brixner, H. Eichler, W. Goettsch, 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," *Value Health*, vol. 20, no. 8, pp. 1003–1008, 2017.
6. S. V. Wang, S. Schneeweiss, M. L. Berger, J. Brown, F. de Vries, I. Douglas, et al., "Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1. 0," *Value Health*, vol. 20, no. 8, pp. 1009–1022, 2017.
7. IMI, "Innovative Medicines Initiative GetReal," 2013. https://www.imi-getreal.eu/
8. M. Levenson, W. He, J. Chen, Y. Fang, D. Faries, B. A. Goldstein, et al., "Biostatistical Considerations when using RWD and RWE in Clinical Studies for Regulatory Purposes: A Landscape Assessment," *Stat. Biopharm. Res.*, pp. 1–11, 2021.
9. J. Chen, M. Ho, K. Lee, Y. Song, Y. Fang, B. A. Goldstein, et al., "The Current Landscape in Biostatistics of Real-World Data and Evidence: Clinical Study Design and Analysis," *Stat. Biopharm. Res.*, pp. 1–14, 2021.
10. M. Ho, M. van der Laan, H. Lee, J. Chen, K. Kee, Y. Fang, et al., "The Current Landscape in Biostatistics of Real-World Data and Evidence: Causal Inference Frameworks for Study Design and Analysis," *Stat. Biopharm. Res.*, pp. 1–14, 2021.
11. J. Chen, D. Scharfstein, H. Wang, B. Yu, Y. Song, W. He, et al., "Estimand in real-world evidence studies," *Submitted to Stat. Biopharm. Res.*, 2022.
12. M. Levenson, W. He, L. Chen, S. Dharmarajan, R. Izem, Z. Meng, et al., "Statistical Consideration for Fit-for-Use Real-World Data to Support Regulatory Decision Making in Drug Development," *Accepted by Stat. Biopharm. Res.*, 2022. https://doi.org/10.1080/19466315.2022.2120533
13. M. Ho, S. Gruber, Y. Fang, D. E. Faris, P. Mishra-Kalyani, D. Benkeser, Mark van der Laan., "Examples of Applying RWE Causal Inference Roadmap to Clinical Studies," *Accepted by Stat. Biopharm. Res.*, 2023.
14. J. Concato and J. Corrigan-Curay, "Real-World Evidence-Where Are We Now?," *N. Engl. J. Med.*, vol. 386, no. 18, pp. 1680–1682, 2022.
15. P. Arlett, J. Kjær, K. Broich, and E. Cooke, "Real-World Evidence in EU Medicines Regulation: Enabling Use and Establishing Value," *Clin. Pharmacol. Ther.*, vol. 111, no. 1, p. 21, 2022.
16. EMA/HMA, "Big Data Workplan 2022–2025," 2022. https://www.ema.europa.eu/en/documents/work-programme/workplan-2022-2025-hma/ema-joint-big-data-steering-group_en.pdf
17. NICE, "Real-World Evidence Framework.," 2022. https://www.nice.org.uk/corporate/ecd9/chapter/overview
18. NICE, "Strategy 2021 to 2026.," 2021. https://static.nice.org.uk/NICE%20strategy%202021%20to%202026%20-%20Dynamic,%20Collaborative,%20Excellent.pdf
19. HAS, "Real-World Studies for the Assessment of Medicinal Products and Medical Devices," 2021. https://www.has-sante.fr/upload/docs/application/pdf/2021-06/real-world_studies_for_the_assessment_of_medicinal_products_and_medical_devices.pdf
20. S. Kent, M. Salcher-Konard, S. Boccia, J. C. Bouvy, C. de Waure, J. Espin, et al., "The Use of Nonrandomized Evidence to Estimate Treatment Effects in Health Technology Assessment," *J. Comp. Eff. Res.*, vol. 10, no. 14, pp. 1035–1043, 2021.
21. ICH, "ICH E9(R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials," 2020. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf
22. EMA/FDA, "Parallel Scientific Advice," 2021. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/pilot-programme-european-medicines-agency-food-drug-administration-parallel-scientific-advice-hybrid/complex-generic-products-general-principles_en.pdf

23. FDA, "Data Standards for Drug and Biological Product Submissions Containing Real-World Data," 2021. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-drug-and-biological-product-submissions-containing-real-world-data
24. K. Loudon, S. Treweek, F. Sullivan, P. Donnan, K. E. Thorpe, and M. Zwarenstein, "The PRECIS-2 Tool: Designing Trials that are Fit for Purpose," *BMJ*, vol. 350, 2015.
25. Y. Fang, H. Wang, and W. He, "A Statistical Roadmap for Journey from Real-World Data to Real-World Evidence," *Ther. Innov. Regul. Sci.*, vol. 54, no. 4, pp. 749–757, 2020.
26. EUnetHTA, "EUnetHTA21." 2021. [Online]. Available. https://www.eunethta.eu/about-eunethta/

Overview of the Current Real-World Evidence Regulatory Landscape

Rima Izem, Ruthanna Davi, Jingyu Julia Luan, and Margaret Gamalo

1 Introduction

Understanding the regulatory landscape in real-world evidence (RWE) is strategically important in therapeutic development, as it can help better plan studies or data collection to inform relevant regulatory questions and it can help with fair communication of benefit–risk information relevant to patient, their doctors, and the healthcare system [1, 2].

As we discuss in this chapter, RWE has the great potential to fill knowledge gaps in the planning or in the life cycle of the development program of new therapies to inform regulatory approval or payer decisions. While randomized controlled trials are the gold standard for evaluating new medical treatments and providing high quality internally valid evidence for judging medical product efficacy and safety in a controlled setting, their use for all regulatory decision-making regarding marketing and reimbursement has some limitations. Some research questions cannot or generally are not answered with clinical trial data and real-world data (RWD) may offer insights not otherwise possible. In addition, leveraging existing data may increase efficiency in evidence generation and accelerate patient access to safe and effective therapies. Statistical thinking can facilitate the leveraging of

R. Izem (✉)
Statistical Methodology, Novartis Pharma AG, Basel, Switzerland
e-mail: rima.izem@novartis.com

R. Davi
Data Science, Medidata AI, New York, NY, USA

J. J. Luan
Regulatory Affairs, AstraZeneca, Gaithersburg, MD, USA

M. Gamalo
Global Biometrics, Pfizer, Collegeville, PA, USA

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_2

RWD, including data gathered as part of the delivery of health care, or other existing external data such as registries or prior clinical trials, to generate actionable evidence.

The guidance documents we summarize in Sect. 2 draw a balance between leveraging RWD while keeping the standards high for regulatory decision-making. The concept of "fitness-for-purpose" is therefore central to the regulatory theory and practice, worldwide, around RWE. Demonstrating fitness-for-purpose starts with a clearly stated purpose or the regulatory context for using RWD. These include supporting a new indication, evaluating the benefit–risk in a novel subgroup, revising the label, and more generally updating the benefit–risk profile. Then, one needs to demonstrate that the data are of sufficient quality, reliability, and validity and that the methodological approaches for using the data are of sufficient rigor.

In Sect. 3, we illustrate how the key concepts and principles outlined in the earlier section have been applied in practice for different purposes with a few examples. Each subsection focuses on a particular purpose, explains the underlying motivation, and illustrates examples of use, data sources, and statistical methodology supporting that use.

2 Key Concepts in Real-World Evidence Worldwide

This section will define some key concepts in the use of RWE in drug development and discuss their similarities and differences worldwide.

We focus our source documents (Fig. 1) on those that use the terminology RWD and RWE in drugs and biologic therapeutic development. Our sources include important publicly available RWD/RWE guidance documents for drugs from the United States Food and Drug Administration (FDA) [3–7], the European Medicines Agency (EMA) [8, 9], the Medicines and Healthcare Products Regulatory Agency (MHRA) in the United Kingdom [10, 11]. Our sources also include personal communications relating to documents from the Japan Pharmaceutical and Medical Device Agency (PMDA) and the Chinese Center for Drug Evaluation (CDE).

We acknowledge that while the regulatory framework of RWE has been recently formalized worldwide, and the terminology or concepts are sometimes novel, the use of RWD in a regulatory setting is long-standing. Thus, sources in Fig. 1 exclude earlier guidance documents related to RWE. For example, regulators used adverse events reporting systems for decades to evaluate post-market safety [12, 13]; they used non-interventional studies to inform the risks of existing products and how to mitigate them, and they used literature review or historical clinical trials to contextualize treatment effects in single arm studies [14]. While these examples will be discussed in the next section of this chapter, the guidance documents are not used as source documents in this section.

There is no consensus international conference of harmonization (ICH) document regarding RWE, but the guidance documents published by different countries generally share a similar thinking and philosophy. The next subsections review the key concepts, the similarities, and the differences worldwide.

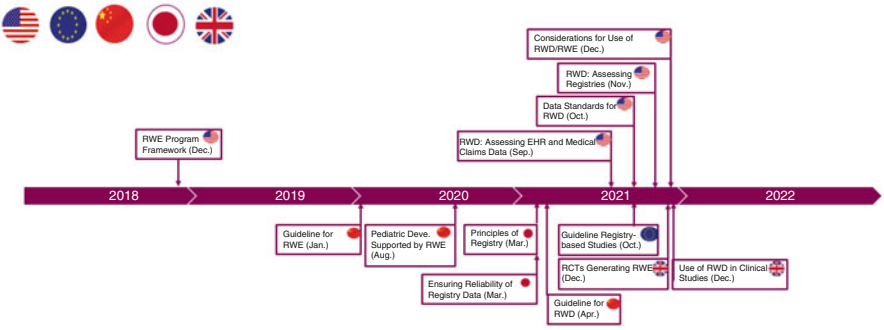

Fig. 1 Timeline of important guidance documents in five major health authorities. This figure shows the main milestone in regulatory guidance documents across the world, with flags from left to right representing the US-FDA, The European EMA, the Chinese CDE, the Japanese PMDA, and the United Kingdom's MHRA. Refer to text for discussion of specific documents

2.1 Sources of Real-World Data and Real-World Evidence

All guidance documents agree that RWE is the evidence generated from RWD. While the definition of RWE is similar worldwide, the needs of different regulatory bodies vary. In the United States, decisions are based on benefit and risk evaluations of medical products, whereas in other parts of the world, they include reimbursement decisions. Similarly, while the definition of RWD is generally similar worldwide, there are differences in the existing data sources.

The definitions for RWD are broad worldwide. For example, the US FDA 2018 RWE Framework guidance states that "Real-World Data (RWD) are data relating to patient health status and/or delivery of health care routinely collected from a variety of sources." While the PMDA 2021 guidance states that "RWD are data on patient's health conditions and/or provided medical practices routinely collected from various data sources." These definitions are broad because they include prospective data collection in a registry or retrospective data collection in an electronic healthcare database. They additionally include data from interventional studies, such as randomized pragmatic studies, or data from non-interventional observational studies such as prospective cohorts. Thus, the data could have been collected with a primary intent to answer the research question, for primary use, or the data could have been collected for other purposes than research, such as clinical care or fee-for-service reimbursement and be a secondary use source. Finally, the definitions are specific to the setting and type of information that is collected, but not to the sample of patients or time frame in which the information is collected.

The sources of RWD can vary worldwide due to differences between countries or healthcare infrastructures with regard to how routine care is provided and recorded. For example, data from traditional Chinese medicine is a more readily available source of RWD in China than in other countries. Similarly, some healthcare systems in the United States or Germany are federated with no expectations of a single source

having a full longitudinal picture for a particular patient. Other healthcare systems, for example, in the Nordic countries or in the United Kingdom, are centralized with many existing linkages and typically have good longitudinal data capture. Finally, given the current advances in digital technology and broad use of various digital devices in clinical care, the sources of RWD have been evolving and expanding.

The FDA requires submission of clinical data in marketing applications. That requirement extends to RWD [7]. Although the document recognizes the challenges involved in standardizing study data derived from RWD sources for inclusion in applicable drug submissions, it emphasizes the importance of documenting processes for managing RWD.

2.2 Regulatory Acceptability and Demonstrating Fitness-for-Purpose

All documents indicate that evidentiary standards for approval will remain high and require substantial evidence and adequate and well-controlled studies. Assessing fitness-for-purpose is thereby ensuring that for a given purpose, the RWD and the methods used to generate the evidence are relevant and adequate. The RWD attributes relative to a purpose include the following: quality, reliability, relevance, provenance, security, and protection of personal information. Thus, rather than the one-size-fits all approach, the RWD attributes and the proposed methods, including the design and analysis, are tailored to a particular purpose and the regulatory evaluation is specific to that purpose. We refer the reader to chapter "Assessment of Fit-For-Use Real-World Data Sources and Applications" in this book for definitions and discussion of these attributes.

In addition, the design and analyses need to minimize any potential biases in the evaluation of treatment effectiveness or safety. Thus, any deviation from randomization or blinding in the design must be justified, as they may increase bias. Even though it is not explicitly expressed in all guidelines, transparency of the study design and pre-specification of analytical methods are very important. For example, the PMDA's 2021 basic principle guidance document states that "For reliability assurance of the results, it is therefore important to demonstrate the reliability of the data and transparency of the study design and analysis." Moreover, the MHRA's 2021 guidance states that "From a regulatory perspective whether the study data is all from the real-world setting or the result of a hybrid or traditional RCT is not critical. The important thing is that the trial is designed in a way which allows it to provide the evidence required to answer the regulatory question and a well-designed and conducted prospective randomized controlled trial provides a high level of evidence irrespective of the categorization of the data source."

3 Regulatory Precedent Examples of Fit-for-Purpose Real-World Evidence

This section reviews regulatory precedents in RWE in the development life cycle and provides for each example of use, or scientific purpose, the motivation behind using RWD. The stated purposes below are similar to those outlined by the US FDA in their review paper [1]. When relevant, we also share new developments in statistical methodology supporting these uses. We ordered these uses from more established to less established uses in the regulatory setting.

3.1 Scientific Purpose of Supporting Planning of Clinical Trials

Exploratory analyses of data from clinical practice are increasingly informing planning of design of clinical trials for new therapies. The RWD sources, their sizes and their types, the standardization of the data structure through common data models [15], and the sophistication and scalability of queries on these data have evolved over the last few years. For example, although selection of inclusion and exclusion criteria and clinical sites for a clinical trial traditionally relied on prior experience or investigator assessment, these selections are increasingly informed by queries of large electronic healthcare data using computable phenotypes or algorithms (see [16] and chapter "Key Variables Ascertainment and Validation in Real-World Setting" of this book). The query process leverages the common data model across a data network of electronic healthcare data or insurance claims to run validated algorithms and provide a snapshot of the population at a given time. This may inform eligibility criteria development by examining associations with key trial outcomes. Efficient use of the data to identify and follow participants in clinical trials was deployed in a large scale to test new or existing COVID-19 therapies [17–19].

Another growing area of use of RWD to inform clinical development of new therapies is qualification of novel biomarkers supported by results of data mining and machine learning methods. For example, the EMA qualification of Islet Autoantibodies in Type 1 Diabetes relied on machine learning methods and data mining of RWD to identify prognostic factors and validate the biomarker for use in clinical trials [20].

3.2 Scientific Purpose of Supporting Safety and Effectiveness Evaluation

Over the past decades, post-market safety assessments often relied on what we call today RWD and is more traditionally called safety surveillance using spontaneous safety reporting databases or prospectively designed epidemiological studies. Thus, using RWD to support evaluation of safety post-marketing is a relatively mature field in pharmacoepidemiology with several existing guidance documents, all finalized before the terminology of RWD or RWE were first coined [12, 13].

Several review papers give an overview of the regulatory landscape as well as the statistical considerations in the post-market safety setting. Those include, for example, the following review publications [21–24]. As these papers and the examples discussed in them illustrate, all general RWE principles outlined in Sect. 2 hold also for RWE to support safety. In fact, many of the RWE principles were probably inspired by the regulatory experience of using RWD in safety in the past decades. However, a few elements are safety-specific and highly impact the fitness-for-purpose assessment. These include the rarity and/or the importance of long-term longitudinality in some safety events. Rarity of events results in lack of power of smaller databases to detect the risk and lack of feasibility of some analytical methods. Also, poor or incomplete capture of long-term or long progression safety outcomes renders many RWD inadequate to assess these outcomes.

In terms of statistical methodology, the use of increasingly large spontaneous reporting databases has spurred the development of several statistical methods in disproportionality analyses [21, 25]. The main methodological challenges with these databases are handling reporting bias and lack of information on the universe of drug utilization (aka, no denominator). Similarly, the use of increasingly large distributed claims databases in post-market safety spurred the use and implementation of different cohort selection and causal inference methods in pharmacoepidemiology, including, for example, the propensity score methods discussed in these review publications [26, 27]. The main methodological challenges are assessing and handling selection bias and confounding in the causal inference. The rarity of the outcomes further challenges model fitting and adjustment. Although most safety outcomes are binary, the choice of summary measures is simplified with rare outcomes because hazard ratios, relative risks, and odds ratios are all asymptotically equivalent.

When safety assessments using RWD is not sufficient, some therapeutic areas require the design of a dedicated randomized clinical trial starting in the pre-market setting. For example, randomized and controlled clinical trials were carried out for the examination of the risk of asthma-related hospitalization, intubation, and death associated with long-acting beta agonist use [28]. Also, cardiovascular outcomes trials are often required for antidiabetic therapies [29]. These large studies require extensive resources but were deemed necessary for multiple reasons, including the importance of balancing baseline composition of all covariates through randomization.

Randomized simple pragmatic trials with data collection through electronic healthcare data may provide similar benefits as the traditional randomized trial approach, while conserving resources and making very large trials more feasible. These designs were recommended for cardiovascular outcome studies in antidiabetic drugs [30]. Statistical considerations in these designs include the degree of pragmatism and its implication on the design, analysis, and regulatory acceptability as well as the interoperability between data collected solely for the clinical trial with data retrieved from the electronic healthcare system [31, 32].

In the case of the long-acting beta agonist example mentioned above, the value-added of randomizing treatment in the clinical trial was challenged by multiple researchers [33], who argued that large scale multinational (non-randomized) cohort studies could have accomplished the same goal. Similarly, one of the main findings in the RCT DUPLICATE initiative funded by the FDA was that well designed cohort studies could replicate the findings from pragmatic randomized clinical trials [34].

3.3 Scientific Purpose of Serving as External Control to a Clinical Study

For many years, approval of new therapies in some rare diseases or oncology indications relied customarily on single arm trials and comparisons to a well-documented natural history of the disease or outcomes of active treatment in an external comparable population. This practice is reserved for diseases with high and predictable mortality and circumstances where the effect of the drug is self-evident since the historical control is often not as well assessed as a concurrent control population [14]. More recently, with the availability of a large volume of patient-level data and using statistical methods for balancing baseline composition, there is potential to improve the quality of these external comparisons, thereby allowing a more nuanced inference [3, 35, 36].

The review paper [37] outlines some recent case studies using RWD as an external control in oncology and rare disease single arm studies in the marketing submission to the FDA. Other examples are also discussed in chapter "Applications Using Real-World Evidence to Accelerate Medical Product Development" of this book. In all examples, ensuring fitness-for-purpose of the RWD and the methods were critical in the regulatory reviews. These examples include the study of overall survival associated with selinexor in patients with penta-exposed, triple-class refractory multiple myeloma using a Phase 2 single arm trial and an external control cohort created from electronic medical records. The issues with assessing fitness-for-purpose were discussed at an FDA advisory committee meeting preceding the accelerated approval of the product [38]. The examples also include the use of natural history studies to contextualize the findings of a single arm study in a rare pediatric lysosomal disorder to support the approval and labeling of cerliponase alfa. The application discussed validity of the endpoint capture in the natural

history study, and corrected for confounding using design and analytical strategies [39]. Another example is evaluating the treatment effect of Tecartus on objective response rate and overall survival in relapsed/refractory adult B-cell precursor acute lymphoblastic leukemia that was examined with a single arm trial and an external control derived from individual patient-level data sampled from historical clinical trials [40]. The historical data was made comparable to the single arm study using propensity score methods. This was followed by a positive recommendation for the product from the EMA [41].

Beyond the application of external controls to single arm trials, external data can be used to augment a randomized control instead of replacing it [42]. For example, such a design has been proposed in a Phase 3 registrational trial in recurrent gliobastoma as a (3:1) randomized trial, augmented with external control patients to form a fully powered Phase 3 registrational trial [43]. Additionally, the recently published Phase 2 study [44] randomized fewer patients to placebo and augmented the control with data from historical clinical trials. In both studies, the external data came from previously conducted clinical trials. In the former, frequentist propensity score methods were applied and, in the latter, Bayesian borrowing approaches were used in the analyses. As the last example demonstrates, the use of external controls to replace or augment controls in early phase development (e.g., continued development after Phase 2 to Phase 3) is promising.

3.4 Scientific Purpose of Supporting Extrapolating Efficacy or Safety Findings

Extrapolating efficacy or safety findings to patients outside the controlled setting and restrictive eligibility criteria typical of clinical trials can be challenging. Patients with comorbidities, such as older age or chronic kidney disease, at increased risk for serious adverse reactions, or those with concomitant treatments that may confound assessment of efficacy or that may modify exposure to the drug are often intentionally excluded from clinical trials. However, these patients are likely to be treated with the medical products when approved and available in clinical practice. In addition, racial and other groups are often underrepresented in clinical trials, possibly owing to lesser access to or willingness to participate in clinical trials. Thus, RWD may present opportunities to fill these knowledge gaps or improve our understanding of the therapy's effects after an early period of marketing and use in the clinical setting. Estimates of the so-called, real-world effectiveness and safety, including these lesser studied or rare populations and exploration of treatment effects by demographic, medical history, and disease characteristics, or socioeconomic status, could be highly impactful for the treatment and care of patients.

Because extrapolation, as a concept, was already in use in the development of new therapies in pediatrics, RWD as a useful data source was embraced by the

recently released international pediatric guidance discussing extrapolation [45]. For example, Blinatumomab was approved for precursor B-cell acute lymphoblastic leukemia in children based on efficacy data from an open-label Phase 1/2 trial and safety data from a single-arm, open-label (observational) expanded access study [46]. Of note, given that most pediatric trials are short, real-world data can inform our understanding of the long-term effects of exposure to medicines [47, 48].

Beyond pediatrics, this extrapolation thinking was successful in supporting expanding the indication of palbociclib to male cancer patients after approval of the drug for female cancer patients based on a successful randomized clinical study [49].

4 Conclusions and Discussion

This chapter summarizes the regulatory landscape of the use of RWD to generate RWE. As we discuss in Sect. 2, fitness-for-purpose is a central concept in the generation of RWE. As the regulatory precedent examples in Sect. 3 illustrate, arguments for fitness-for-purpose are tailored to the purpose in each example.

We believe that concepts in RWE are bound to evolve in sophistication in the next few years, as new RWD sources emerge and as some novel designs blur the distinction between traditional clinical trials and clinical care. For example, pragmatic studies are often randomized, but embedded in routine care and can lead to using or collecting RWD. Similarly, decentralized studies incorporate data from wearable digital technology, a growing source of RWD, into clinical trials.

Our summary of purposes in Sect. 3 illustrates that use of fit-for-purpose RWD to generate RWE is well established in those situations where RWD is filling a gap that would be difficult or impossible to fill by a typical clinical trial. Those include using fit-for-purpose RWD to help plan a randomized study, to support post-market safety, and to evaluate comparative effectiveness.

In other situations, it is challenging to balance the use of RWD without raising concerns of lowering the regulatory standards. More specifically, the situations where findings from some RWD sources are intended to replace findings from randomized clinical trials are still being defined. While Sect. 3 had examples using RWD as external controls in rare diseases and oncology clinical development, using RWD to expand indication(s) for an already approved product with a well characterized safety profile, and in using RWD to support extrapolation of a treatment effect from one group to another, it is unclear how generalizable these examples would be to other therapies or populations. With this evolving landscape, a continuous open dialog and early consultation with regulators can benefit every research program considering the use of RWD.

References

1. Concato, J. and J. Corrigan-Curay, *Real-World Evidence—Where Are We Now?* New England Journal of Medicine, 2022. **386**(18): p. 1680–1682.
2. Arlett, P., et al., *Real-World Evidence in EU Medicines Regulation: Enabling Use and Establishing Value.* Clinical Pharmacology & Therapeutics, 2022. **111**(1): p. 21–23.
3. The US Food and Drug Administration. *Framework for FDA's Real-World Evidence Program,*. 2018 December 2018; Available from: https://www.fda.gov/media/120060/download.
4. The US Food and Drug Administration. *Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drug and Biological Products,*. 2022 [cited 2022; Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/submitting-documents-using-real-world-data-and-real-world-evidence-fda-drug-and-biological-products.
5. The US Food and Drug Administration. *Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products,*. 2021 [cited 2022 September]; Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory.
6. The US Food and Drug Administration. *Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry,*. 2021 [cited 2022 September]; Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-registries-support-regulatory-decision-making-drug-and-biological-products.
7. The US Food and Drug Administration. *Data Standards for Drug and Biological Product Submissions Containing Real-World Data,*. 2021 [cited 2022 September]; Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-drug-and-biological-product-submissions-containing-real-world-data.
8. The European Medicines Agency. *A vision for use or real-world evidence in EU medicines regulation,*. 2021 [cited 2022 September]; Available from: https://www.ema.europa.eu/en/news/vision-use-real-world-evidence-eu-medicines-regulation.
9. The European Medicines Agency. *Guideline on registry-based studies,*. 2021 [cited 2022 September]; Available from: https://www.ema.europa.eu/en/guideline-registry-based-studies-0.
10. The Medicines & Healthcare Products Regulatory Agency, *MHRA guideline on randomized controlled trials using real-world data to support regulatory decisions*. 2021.
11. The Medicines & Healthcare Products Regulatory Agency, *MHRA guidance on the use of real-world data in clinical studies to support regulatory decisions,*. 2021.
12. The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). *Guide on Methodological Standards in Pharmacoepidemiology (Revision 10),*. 2010 [cited 2022 September 2022]; Available from: http://www.encepp.eu/standards_and_guidance.
13. The US Food and Drug Administration, *Guidance for industry and FDA staff. Best practices for conducting and reporting pharmacoepidemiologic safety studies using electronic healthcare data. May 2013*. 2015.
14. International Conference of Harmonization. *ICH E10 Choice of control group in clinical trials,*. 2001 [cited 2022 September]; Available from: https://www.ema.europa.eu/en/ich-e10-choice-control-group-clinical-trials.
15. Kush, R.D., et al., *FAIR data sharing: The roles of common data elements and harmonization.* Journal of Biomedical Informatics, 2020. **107**: p. 103421.
16. Nelson, S.J., et al., *EHR-based cohort assessment for multicenter RCTs: a fast and flexible model for identifying potential study sites.* Journal of the American Medical Informatics Association, 2022. **29**(4): p. 652–659.

17. Brown, J.S., L. Bastarache, and M.G. Weiner, *Aggregating Electronic Health Record Data for COVID-19 Research—Caveat Emptor.* JAMA Network Open, 2021. **4**(7): p. e2117175–e2117175.
18. Normand, S.-L.T., *The RECOVERY Platform.* New England Journal of Medicine, 2020. **384**(8): p. 757–758.
19. Haendel, M.A., et al., *The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment.* Journal of the American Medical Informatics Association, 2021. **28**(3): p. 427–443.
20. Podichetty, J.T., et al., *Leveraging Real-World Data for EMA Qualification of a Model-Based Biomarker Tool to Optimize Type-1 Diabetes Prevention Studies.* Clinical Pharmacology & Therapeutics, 2022. **111**(5): p. 1133–1141.
21. Izem, R., et al., *Sources of Safety Data and Statistical Strategies for Design and Analysis:Postmarket Surveillance.* Therapeutic Innovation & Regulatory Science, 2018. **52**(2): p. 159–169.
22. Ma, H., et al., *Sources of Safety Data and Statistical Strategies for Design and Analysis:Transforming Data Into Evidence.* Therapeutic Innovation & Regulatory Science, 2018. **52**(2): p. 187–198.
23. Chakravarty, A.G., et al., *The role of quantitative safety evaluation in regulatory decision making of drugs.* Journal of Biopharmaceutical Statistics, 2016. **26**(1): p. 17–29.
24. Wang, W., et al., *Quantitative Drug Safety and Benefit Risk Evaluation: Practical and Cross-Disciplinary Approaches*. 2021: CRC Press.
25. Arnaud, M., et al., *Methods for safety signal detection in healthcare databases: a literature review.* Expert Opinion on Drug Safety, 2017. **16**(6): p. 721–732.
26. Gagne, J.J., et al., *Design considerations in an active medical product safety monitoring system.* Pharmacoepidemiol Drug Saf, 2012. **21 Suppl 1**: p. 32–40.
27. Franklin, J.M., et al., *Comparing the performance of propensity score methods in healthcare database studies with rare outcomes.* Stat Med, 2017. **36**(12): p. 1946–1963.
28. Seymour, S.M., et al., *Inhaled Corticosteroids and LABAs — Removal of the FDA's Boxed Warning.* New England Journal of Medicine, 2018. **378**(26): p. 2461–2463.
29. The US Food and Drug Administration, *Guidance for industry: diabetes-mellitus- evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes*. 2008.
30. Sharma, A., et al., *Impact of Regulatory Guidance on Evaluating Cardiovascular Risk of New Glucose-Lowering Therapies to Treat Type 2 Diabetes Mellitus.* Circulation, 2020. **141**(10): p. 843–862.
31. Loudon, K., et al., *The PRECIS-2 tool: designing trials that are fit for purpose*. BMJ: British Medical Journal, 2015. **350**: p. h2147.
32. Rockhold, F.W., et al., *Design and analytic considerations for using patient-reported health data in pragmatic clinical trials: report from an NIH Collaboratory roundtable.* J Am Med Inform Assoc, 2020. **27**(4): p. 634–638.
33. Suissa, S., et al., *Food and Drug Administration-mandated Trials of Long-Acting β-Agonist Safety in Asthma. Bang for the Buck?* Am J Respir Crit Care Med, 2018. **197**(8): p. 987–990.
34. Franklin, J.M., et al., *Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies.* Circulation, 2021. **143**(10): p. 1002–1013.
35. Davi, R., et al., *Informing single-arm clinical trials with external controls.* Nat Rev Drug Discov, 2020. **19**(12): p. 821–822.
36. Burcu, M., et al., *Real-world evidence to support regulatory decision-making for medicines: Considerations for external control arms.* Pharmacoepidemiol Drug Saf, 2020. **29**(10): p. 1228–1235.
37. Izem, R., et al., *Real-World Data as External Controls: Practical Experience from Notable Marketing Applications of New Therapies.* Therapeutic Innovation & Regulatory Science, 2022: p. 1–13.
38. US Food and Drug Administration. *Meeting of the Oncologic Drugs Advisory Committee Meeting,*. 2019 [cited 2022 September]; Available from: https://www.fda.gov/advisory-committees/advisory-committee-calendar/february-26-2019-meeting-oncologic-drugs-advisory-committee-meeting-announcement-02262019-02262019.

39. Schulz, A., et al., *Study of intraventricular cerliponase alfa for CLN2 disease.* New England Journal of Medicine, 2018. **378**(20): p. 1898–1907.
40. Shah, B.D., et al., *The Comparison of Kte-X19 to Current Standards of Care: A Pre-Specified Synthetic Control Study Utilizing Individual Patient Level Data from Historic Clinical Trials (SCHOLAR-3).* Blood, 2021. **138**(Supplement 1): p. 3844–3844.
41. Kitepharma. *Press Release: Kite's CAR T-Cell therapy Tecartus receives positive CHMP opinion in relapsed or refractory acute Lymphoblastic Leukemia,*. 2022 [cited 2022 September]; Available from: https://rsconnect-prod.dit.eu.novartis.net/content/013dd42c-fac2-4ad8-9ae2-f395bba5d3d2/matching-adjusted-indirect-comparisons.html.
42. Schmidli, H., et al., *Beyond randomized clinical trials: Use of external controls.* Clinical Pharmacology & Therapeutics, 2020. **107**(4): p. 806–816.
43. Majumdar, A., et al., *Building an External Control Arm for Development of a New Molecular Entity: An Application in a Recurrent Glioblastoma Trial for MDNA55.* Statistics in Biosciences, 2022: p. 1–19.
44. Richeldi, L., et al., *Trial of a Preferential Phosphodiesterase 4B Inhibitor for Idiopathic Pulmonary Fibrosis.* New England Journal of Medicine, 2022. **386**(23): p. 2178–2187.
45. International Conference of Harmonization. *ICH guideline E11A on pediatric extrapolation,*. 2022 [cited 2022 September]; Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/draft-ich-guideline-e11a-pediatric-extrapolation-step-2b_en.pdf.
46. Locatelli, F., et al., *Blinatumomab in pediatric patients with relapsed/refractory acute lymphoblastic leukemia: results of the RIALTO trial, an expanded access study.* Blood Cancer Journal, 2020. **10**(7): p. 77.
47. McMahon, A.W. and G. Dal Pan, *Assessing Drug Safety in Children - The Role of Real-World Data.* N Engl J Med, 2018. **378**(23): p. 2155–2157.
48. Lasky, T., et al., *Real-World Evidence to Assess Medication Safety or Effectiveness in Children: Systematic Review.* Drugs Real World Outcomes, 2020. **7**(2): p. 97–107.
49. Wedam, S., et al., *FDA Approval Summary: Palbociclib for Male Patients with Metastatic Breast Cancer.* Clin Cancer Res, 2020. **26**(6): p. 1208–1212.

Key Considerations in Forming Research Questions and Conducting Research in Real-World Setting

Yixin Fang and Weili He

1 Introduction

To accelerate medical product development to help patients who are in need, Food and Drug Administration (FDA) and European Medicines Agency (EMA) develop guidance documents on real-world data (RWD) and real-world evidence (RWE) for industry [1–6]. These guidance documents provide recommendations to sponsors on the use of RWE to support approval of a new indication for a medical product that has already been approved or to help support postapproval study requirements, along with other objectives such as investigating disease burdens and treatment patterns. These guidance documents center on how to derive robust RWE from the analysis of RWD and the use of RWE in regulatory decision-making.

In a statistical roadmap for journey from RWD to RWE [7], the point of departure is forming a sound research question. As raised in ICH E9(R1) [8], "central questions for drug development and licensing are to establish the existence, and to estimate the magnitude, of treatment effects: how the outcome of treatment compares to what would have happened to the same subjects under alternative treatment (i.e., had they not received the treatment, or had they received a different treatment)." Note that ICH I9(R1) asks the central questions in terms of potential outcomes. The potential outcomes framework has been used in the community of causal inference since Neyman proposed it in his 1923 Master's thesis and Rubin in 1974 extended it into a general framework for causal inference in both interventional studies and non-interventional settings [9, 10]. "What would have happened" is also called counterfactual outcome, and human's ability of imagining counterfactual outcomes plays the most crucial role in forming research questions [13].

Y. Fang (✉) · W. He
Data and Statistical Sciences, AbbVie, North Chicago, IL, USA
e-mail: yixin.fang@abbvie.com

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_3

According to the definitions in FDA guidance document [4], an interventional study (a.k.a., a clinical trial) is "a study in which participants, either healthy volunteers or volunteers with the disease being studied, are assigned to one or more interventions, according to a study protocol, to evaluate the effects of those interventions on subsequent health-related biomedical or behavioral outcomes." A non-interventional study (a.k.a., an observational study) is "a type of study in which patients received the marketed drug of interest during routine medical practice and are not assigned to an intervention according to a protocol." Both traditional randomized controlled trials (RCTs) and pragmatic clinical trials (PCTs) are examples of interventional studies, and non-interventional study designs include cross-sectional studies, observational cohort studies, and case–control studies. In addition, single-arm trials with external controls from RWD can be considered as a hybrid of interventional and non-interventional studies.

To form a well-built clinical question, the PICO criteria, as far as we know, were first proposed in 1995 [11], which proposed that "the question must be focused and well articulated for all 4 parts of its 'anatomy': (1) the patient or problem being addressed; (2) the intervention or exposure being considered; (3) the comparison intervention or exposure, when relevant; (4) the clinical outcomes of interest." Since then, many versions of the PICO criteria have also been proposed for forming a sound clinical question, including the PICOT criteria [12], which added the fifth part, Time.

Although the PICOT criteria are still useful for forming research questions in clinical setting, for real-world setting, we need to revise them to take into account the real-world features. In this chapter, we will discuss the PROTECT criteria [7], aligning with the recent FDA guidance documents.

The remaining of the chapter is organized as follows. In Sect. 2, we discuss the steps that we may take before we form a research question, including gathering knowledge and evidence gap, and specifying assumptions and causal model. In Sect. 3, we discuss five key elements of the PROTECT criteria in real-world setting. In Sect. 4, we discuss how to enhance the assumptions or how to revise the elements of the research question if the question cannot be answered. In Sect. 5, we discuss the key considerations in the planning of real-world studies to answer the research question in real-world setting. We conclude the chapter with some discussion in Sect. 6.

2 Gathering Knowledge

"Knowledge" includes scientific experience we have had in the past, clinical evidence we have gathered so far, or systematic literature review we have conducted up to date. The knowledge includes two main parts, the knowledge on what we have known (evidence) and the knowledge on what we have not known (evidence gap). The hidden part (i.e., the unknown) is out of the scope. To fill the evidence gap is the motivation of forming any research question.

Let us consider an example. Assume that, after conducting a placebo-controlled RCT, we have gathered the clinical evidence of efficacy and safety of an investigative treatment compared with placebo in clinical setting. If the treatment has been approved by regulatory agency, now we are interested in investigating the effectiveness and long-term safety of the treatment compared with the standard of care (SOC) in real-world setting.

Since the central questions for drug development and licensing are phrased in terms of potential outcomes in ICH E9(R1), we turn to one book by Pearl, *The Book of Why: The New Science of Cause and Effect* [13], for guidance on how to summarize the knowledge we have gathered into a causal model. As pointed out in [13], the knowledge remains implicit in the mind of investigators before investigators make them explicit by specifying a list of assumptions based on the available knowledge.

Continue the above example. Based on the available knowledge, the investigators may specify two assumptions: (1) one set of variables (say, income, education, disease severity) are associated with the outcome variable and the decision of taking the investigative treatment or SOC and (2) another set of variables (say, age, gender, race) are associated with the outcome variables but not the treatment decision. The first set of variables are examples of confounders and the second set of variables are examples of effect modifiers. Defined in FDA guidance document [1], a confounder is a variable that can be used to decrease confounding bias when properly adjusted for in an analysis and an effect modifier is a factor that biologically, clinically, socially, or otherwise alters the effects of another factor under study.

Following the thought in [13], these explicit assumptions can then be encapsulated in a causal model. A causal model can be defined in various formats, including causal diagrams and structural equations [14]. Continue the above example. Denote the treatment variable as A (1 for investigative treatment and 0 for SOC) and the outcome variable as Y. Denote the first set of covariates as C and the second set of covariates as M. Figure 1 is an example of causal diagram encapsulating the above two assumptions.

From here we may move forward to form research questions. But sometimes we may want to examine the extent to which some preliminary data are compatible with the causal model. Continue the above example. The causal model in Fig. 1 implies that M and A are independent, and we can use some preliminary data, if available, to test whether M and A are associated. These testable results may lead to revising the causal model. Besides these testable implications, there are untestable assumptions, for example, the assumption that there is no unmeasured confounder. Figure 2 shows an example of causal diagram where there is unmeasured confounder U.

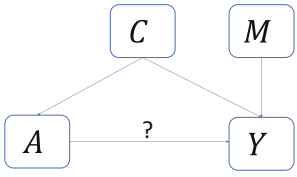

Fig. 1 An example of causal diagram

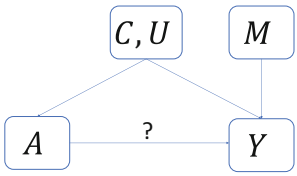

Fig. 2 With unmeasured confounder

3 Forming Research Question

After knowledge gathering, we have a causal model. Now we are ready to form sound research questions. Aligning with the recent FDA guidance documents, we discuss the PROTECT criteria [7] for forming research questions in real-world setting, with five elements discussed in the next five subsections, respectively. The five elements of the PROTECT criteria are summarized in Table 1.

After thinking through these five elements, we are able to articulate research questions. Here are some examples:

- What is the average treatment effect (ATE) of 4 weeks of treatment A on the outcome Y at 12 months after treatment completion compared to 6 weeks of treatment B, among the defined population, after adjusting for confounders (disease severity at the treatment initiation, age, and gender)?
- What is the average treatment effect among the treated (ATT) of one-time treatment C on the outcome Z at 6 months after treatment initiation compared to SOC, among the population of patients who are treated by treatment C in real-world setting?
- What is the 5-year long-term safety of treatment D after treatment initiation, among the population of patients who are treated by treatment D in real-world setting?
- What are the treatment patterns of treatment E from the treatment initiation up to 5 years, among the population of patients who are treated by treatment E initially in real-world setting?

The first two examples have all the five key elements of the PROTECT criteria, which are two specific versions of the central questions raised in ICH E9(R1). The third example also has those five key elements, with the outcome being safety instead of effectiveness. In the fourth example, there is no clinical outcome variable, indicating that other research questions may be formed in real-world setting besides the central questions.

3.1 Population

The target population is the population in which we are interested and for which we will draw conclusions after a study is conducted. As defined in FDA guidance

Table 1 Five elements in the PROTECT criteria

Symbol	Element	Explanation
P	Population	Study population defined via I/E criteria[a]
R/O	Response/outcome	Dependent variable
T/E	Treatment/exposure	Primary independent variable
C	Covariates	Including confounders and effect modifiers[b]
T	Time	When variables are measured?

[a] In general, the 'C' element means counterfactual thinking
[b] I/E criteria are inclusion/exclusion criteria

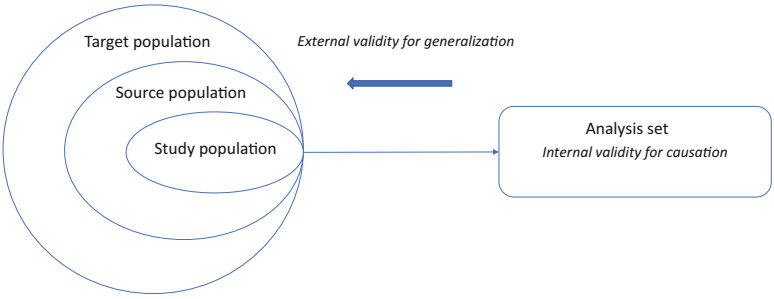

Fig. 3 Three populations and one sample

document [4], the source population is "the population from which the study population is drawn" and the study population is "the population for which analyses are conducted." In addition, the analysis set is a sample of the study population.

The difference between the target population and the source population becomes important in real-world setting when databases are used. For example, if medical claims data are used, the source population is only limited to those whose claims data are collected, because the purpose of medical claims data is to support payment for care. There is further difference between the source population and the study population because the study population is often defined via some data entry criterion that requires that a certain set of variables including treatment variable and/or outcome variable are collected. The third layer of difference is that the analysis set may not be representative of the study population without random sampling. All these three layers of difference lead to selection bias that we should acknowledge in forming research questions.

Figure 3 shows the above three layers of potential selection bias. Ideally, we should collect data on variables that differentiate the analysis set from the populations. There are other sources of selection bias that we should distinguish from confounding bias [15], with one example showed in Fig. 4. In Fig. 4, there are arrow from treatment variable A to S and arrow from outcome variable Y to S such that S becomes a collider [14]. Therefore, conditioning on collider $S = 1$ introduces spurious association between A and Y, introducing selection bias (a.k.a., Berkson's bias [16]). To avoid such selection bias, in the definition of study population, we

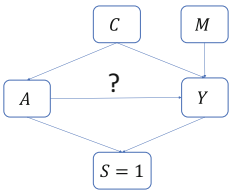

Fig. 4 Selection bias because of $S = 1$

should not have inclusion criteria that require the availability of both treatment data and outcome data. It is internally valid to have inclusion criteria that only require the availability of treatment data, as in cohort studies. It is also internally valid to have inclusion criteria that only require the availability of outcome data, as in case–control studies.

3.2 Response/Outcome

Response variable and outcome variable are two interchangeably used names for dependent variable. ICH E9(R1) simply calls it "variable" when describing the five attributes of an estimand: "The variable (or endpoint) to be obtained for each patient that is required to address the clinical question." However, ICH E9(R1) uses terms "response" and "outcome" in multiple places, such as "patient's outcome" and "response to treatment."

Unlike in traditional clinical setting, in real-world setting when databases are used, we often do not have protocol-defined follow-up visits to ascertain outcome variable. FDA guidance document [1] points out that "a crucial step in selecting a data source is determining whether it captures the clinical outcome of interest." Chapters "Assessment of Fit-for-Use Real-World Data Sources and Applications" and "Key Variables Ascertainment and Validation in RW Setting" of this book will discuss outcome variable ascertainment in more detail. Here we propose to utilize these two terms, response variable and outcome variable, to distinguish two different types of dependent variables.

One type of dependent variable is outcome variable that is defined in the protocol and is to be collected in real-world studies such as pragmatic clinical trials and observational cohort study or outcome variable that is captured in the data source. For example, electronic health records (EHRs) data capture outcomes that are brought to the attention of a health care professional and documented in the medical record. This type of dependent variable also includes outcome variable that is not captured in the given data source but can be ascertained from another data source via data linkage.

The other type of dependent variable is response variable that can be derived (via the definition, construction, and validation process) from the other outcome variables and can be used to measure the patient's response to the investigative

medical product. For example, if patient report outcomes (PROs) or physician report outcomes are captured in the data source, then we can derive some response variable based on these captured subjective outcomes. In some scenarios, response variable may be derived from free texts such as doctors' notes via machine learning and natural language processing (NLP) techniques.

3.3 Treatment/Exposure

In the PICOT criteria, the "I" element stands for intervention, which is inappropriate for non-interventional real-world setting. The "C" element in the PICOT criteria stands for comparator, which is inappropriate for real-world setting where there is no comparator. Therefore, these two elements are replaced by "T/E" in the PROTECT criteria, which stands for treatment/exposure, noting that treatment variable and exposure variable are the two names for primary independent variable that are interchangeably used in real-world setting.

ICH E9(R1) states that "the **treatment** condition of interest and, as appropriate, the alternative treatment condition to which comparison will be made" (referred to as "treatment" through the remainder of this document). FDA guidance document [1] stats that "the term **exposure** applies to the medical product or regimen of interest being evaluated in the proposed study. The product of interest is referred to as the treatment, and may be compared to no treatment, a placebo, standard of care, another treatment, or a combination of the above."

3.4 Covariates (Counterfactual Thinking)

The "C" element has two versions, tangible version and abstract version. Covariates are not included as an element in the PICOT criteria because randomization and blinding are usually applied in the traditional clinical setting. The tangible version of the "C" element in the PROTECT criteria stands for key covariables, which include (1) confounders which will be used to maintain the internal validity of causation and (2) effect modifiers which will be used to better understand heterogeneity of treatment effect and will be potentially used to achieve external validity of generalization. Refer to Fig. 1 for an example of confounders and effect modifiers. Refer to Fig. 3 for the concept of internal validity of causation and external validity of generalization.

The abstract version of the "C" element stands for counterfactual thinking. To form the central questions of ICH E9(R1), we need to imagine counterfactual outcomes, i.e., what would have happened to all the subjects in a certain population under alternative treatment conditions. The covariates ensure that such counterfactual thinking is possible under those assumptions we make in the knowledge gathering stage; for example, under the exchangeability assumption that the counterfactual outcome and the treatment decision are independent given covariates [17].

3.5 Time

In clinical setting, baseline and follow-up period are prespecified in the protocol. In real-world setting, the "T" element in the PROTECT criteria plays a crucial role in understanding all the above four elements.

In the "P" element, we should specify the time period in the definition of the source population. We may consider the time frame as one of the inclusion criteria in the definition of study population drawn from the source population.

In the "R/O" element, we should first identify if the variables are ascertained at a specific time (cross-sectional), retrospectively, prospectively, or in a hybrid fashion. Then, we should clearly define the time periods when the response/outcome variables are measured and collected. In scenarios where baseline and follow-up periods are needed, we should also clearly define the baseline (e.g., treatment initiation, disease diagnosis, or patient enrollment) and follow-up periods (e.g., 6 months after baseline, 12 months after baseline, along with predetermined time windows).

In the "T/E" element, we should first identify whether the treatment/exposure is one-time medical product or other products that may be intended for use over a period of time. If the medical product is one-time, likely the time when the treatment is applied is considered as baseline. If the medical product is intended for use over a period of time, likely the baseline is defined as the treatment initiation, and we should make sure the source data capture the treatment/exposure duration as well, along with data on treatment discontinuation and, if possible, data on the switched treatments.

In the "C" element, we should first distinguish time-independent covariates and time-dependent (a.k.a., time-varying) covariates. For time-dependent covariates, we should describe whether and how frequently the data on these covariates can be captured. Without collecting time-dependent confounders, it is impossible to adjust for time-dependent confounding.

4 Revising Research Question

After a research question is formed, before we answer it, we should evaluate whether or not the research question can be answered. Like in [13], we rely on estimand construction to verify whether the research question can be answered or not. ICH E9(R1) defines an estimand as "a precise description of the treatment effect reflecting the clinical question posed by the trial objective." We can generalize this definition to cover both clinical setting and real-world setting. An estimand is a statistical quantity to be estimated that provides a precise description of the treatment effect reflecting the research question. If we can construct an estimand reflecting the research question, we are able to answer the question. If we cannot, we should either enhance the assumptions or revise some of the PROTECT elements of the question.

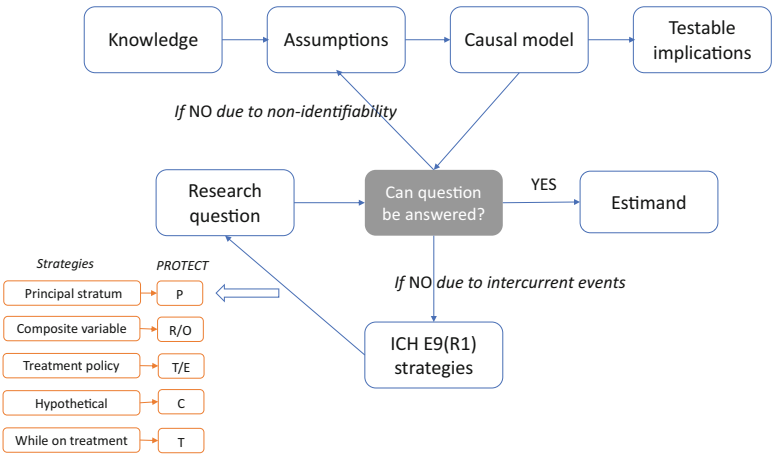

Fig. 5 The roadmap for forming a research question, revising the question if it cannot be answered, and constructing an estimand if the question can be answered

Chapter "Estimand in Real-World Evidence Study: From Frameworks to Application" of this book will discuss estimand in great detail. Here, we only discuss how to revise the question being asked if the question cannot be answered due to lack of identifiability or potential existence of intercurrent events. Figure 5 is motivated by Figure I.1 in the book of why [13]. Figure 5 provides a roadmap of how we check if one research question can be answered and, if not, how we make it answerable.

If the question cannot be answered due to lack of identifiability, we should enhance the assumptions, leading to a revised causal model, such that the question can be answered by the revised causal model. For example, in the construction of estimand, we realize that the list of confounders includes unmeasured confounder U, as displayed in Fig. 2, and then we should either identify a data source to capture the data of U or go back to the knowledge gathering stage to enhance the identifiability assumptions by assuming that U is not a confounder.

If the question cannot be answered due to intercurrent events, we should revise some of the PROTECT elements of the question to address the intercurrent events. ICH E9(R1) defines intercurrent events as "events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest. It is necessary to address intercurrent events when describing the clinical question of interest in order to precisely define the treatment effect that is to be estimated."

Therefore, according to ICH E9(R1), we should discuss how to address intercurrent events when describing the question, instead of waiting until we conduct research to answer the question. ICH E9(R1) proposes five strategies for how to address intercurrent events, with each strategy corresponding to one of the five elements of the PROTECT criteria. This provides another insight that the PROTECT criterion is valid and comprehensive in protecting the quality of the question we form.

If we want to apply the treatment policy strategy to address intercurrent events, we should revise the "T/E" element. According to ICH E9(R1), if the treatment policy strategy is applied, "the intercurrent event is considered to be part of the treatments being compared." That means, we need to revise the definition of treatment or exposure to include the intercurrent event as part of it. For example, if the use of additional medication is considered as an intercurrent event, then this additional medication is considered as part of the revised treatment.

If we want to apply the hypothetical strategy to address intercurrent events, we should revise the "C" element. According to ICH E9(R1), if the hypothetical strategy is applied, "a scenario is envisaged in which the intercurrent event would not occur." That means, we need to revise our counterfactual thinking to imagine what would have happened if the intercurrent event would not occur. Following the tangible version of "C" element, we need to capture data on covariates conditional on which the occurrence of the intercurrent event and the counterfactual outcome can be assumed to be independent.

If we want to apply the composite variable strategy to address intercurrent events, we should revise the "R/O" element. According to ICH E9(R1), if the composite variable strategy is applied, "an intercurrent event is considered in itself to be informative about the patient's outcome and is therefore incorporated into the definition of the variable." That means, we need to revise the definition of response or outcome variable to include the intercurrent event as part of it. For example, if the outcome variable was already success or failure, discontinuation of treatment would simply be considered another mode of failure.

If we want to apply the while on treatment strategy to address intercurrent events, we should revise the "T" element. According to ICH E9(R1), if the while on treatment strategy is applied, "response to treatment prior to the occurrence of the intercurrent event is of interest." That means, we need to revise the definition of the timepoint when the response or outcome variable is measured.

If we want to apply the principal stratum strategy to address intercurrent events, we should revise the "P" element. According to ICH E9(R1), if the principal stratum strategy is applied, "the target population might be taken to be the principal stratum in which an intercurrent event would occur. Alternatively, the target population might be taken to be the principal stratum in which an intercurrent event would not occur." That means, we need to revise the definition of the population as some prespecified principal stratum.

To summarize, according to ICH E9(R1), we should address intercurrent events when describing the research question of interest in order to precisely define the treatment effect that is to be estimated. Five ICH E9(R1) strategies are corresponding to five elements of the PROTECT criteria. If a combination of several strategies is applied for intercurrent events, then we should revise the corresponding combination of elements of the PROTECT criteria as well.

5 Answering Research Question

After an answerable research question is formed and an estimand is defined accordingly to reflect the question, we can move forward to conduct research to answer the question. ICH E9(R1) develops a framework to align five stages of any real-world study with estimand: planning, design, conduct, analysis, and interpretation. The Part II and Part III of this book mainly focus on these five stages, so here we only provide an overview of these five stages, as displayed in Fig. 6, and some key considerations in the planning of real-world studies.

In the planning stage, the first step is to choose appropriate real-world study designs. There are a variety of real-world study designs. Many real-world study designs do not follow the traditional sequence: design a study, enroll subjects, and generate data. Instead, they often intertwine real-world study, real-world data, and/or clinical trial components, so we categorize them into three major categories according to three scenarios. In scenario one, we design a real-world study to prospectively generate real-world data for research purpose. In scenario two, based on one or several existing real-world data sources, we design a retrospective real-world study. In scenarios three, we utilize real-world data in the design, conduct, and analysis of clinical trials.

Category one includes pragmatic clinical trials and observational studies that generate real-world data for research purpose. We may apply some sampling techniques to enroll participants for such studies. We may also utilize RWD to identify potential participants for such studies.

Category two includes study designs that are exclusively based on databases such as EHR, claims, and registries data. FDA guidance documents [1] and [3] assess EHR/claims data and registries data to support regulatory decision-making for drug and biological products, respectively.

Category three includes study designs that are using RWD to augment traditional clinical trials. FDA guidance document [4] provides several such study designs: (1)

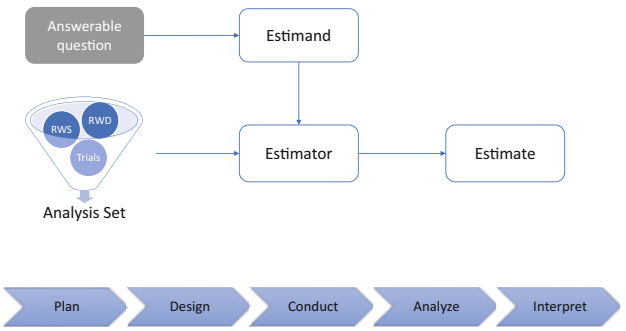

Fig. 6 Alignment of real-world study planning, design, conduct, analysis, and interpretation with estimand; a study may have intertwined real-world study (RWS), real-world data (RWD), and/or clinical trial components

to utilize RWD to identify potential participants for an RCT, (2) to utilize RWD to ascertain response or outcome variable for an RCT, and (3) to use RWD and data of historical trials to augment an RCT or to construct an external control arm for a single-arm clinical trial.

6 Discussion

In medical product development, in clinical setting, we are evaluating efficacy and safety, asking research questions such as "Can it work?". In real-world setting, often we are evaluating effectiveness and safety, asking research questions such as "Does it work?". The PROTECT criteria consist of five elements, helping us to articulate sound research questions.

There is a rich literature on how to form a research question, but the literature on how to revise the research question if it is not answerable is lacking. In this chapter, we propose that we can enhance the causal model assumptions if it is due to lack of identifiability or revise some of the PROTECT elements of the question. ICH E9(R1) emphasizes that "it is necessary to address intercurrent events when describing the clinical question of interest in order to precisely define the treatment effect that is to be estimated." There is also a rich literature on how to address intercurrent events, but no one else argues that it is because we need to revise the question to make it answerable given the existence of intercurrent events. We further demonstrate that each of the five strategies is according to revising one of the five PROTECT elements. This important finding supports our claim that the PROTECT criteria are useful for forming a research question and revising the research question if it is not answerable.

In this chapter, we also argue that estimand is the "touchstone," by which we can verify whether or not a research question is answerable. If an estimand reflecting the question can be constructed, then the question is answerable and the same estimand will guide through the five stages (planning, design, conduct, analysis, and interpretation) of research to answer the question.

References

1. FDA Guidance Document (2021): Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products, https://www.fda.gov/media/152503/download
2. FDA Guidance Document (2021): Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry, https://www.fda.gov/media/153341/download
3. FDA Guidance Document (2021): Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry, https://www.fda.gov/media/154449/download

4. FDA Guidance Document (2021): Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products, https://www.fda.gov/media/154714/download
5. EMA Guidance Document (2021): Guideline on Registry-based Studies, https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-registry-based-studies_en-0.pdf
6. EMA Guidance Document (2021): European Medicines Regulatory Network Data Standardisation Strategy, https://www.ema.europa.eu/en/documents/other/european-medicines-regulatory-network-data-standardisation-strategy_en.pdf
7. Fang, Y., Wang, H., He, W.: A statistical roadmap for journey from real-world data to real-world evidence. *Therapeutic Innovation & Regulatory Science*. **54**, 749–757 (2020)
8. FDA Guidance Document (2021): E9(R1) Statistical Principles for Clinical Trials; Addendum: Estimand and Sensitivity Analysis in Clinical Trials, https://www.fda.gov/media/148473/download
9. Rubin, D.: Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. **66**, 688–701 (1974)
10. Imbens, G., Rubin, D.: Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, (2015)
11. Richardson, W., Wilson, M., Nishikawa, J., Hayward, R.: The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*. **123**, A12–A13 (1995)
12. Haynes, R.: Forming research questions. *Journal of Clinical Epidemiology*. **59**, 881–886 (2006)
13. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic books, New York (2018)
14. Pearl, J.: Causality. Cambridge university press (2009)
15. Haneuse, S.: Distinguishing selection bias and confounding bias in comparative effectiveness research. *Medical Care*. **54**, e23 (2016)
16. Berkson, J.: Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*. **2**, 47-53 (1946)
17. Hernan, M., Robins, J: Causal Inference: What If. Boca Raton: Chapman & Hall/CRC (2020).

Part II
Fit-for-Use RWD Assessment and Data Standards

Assessment of Fit-for-Use Real-World Data Sources and Applications

Weili He, Zuoyi Zhang, and Sai Dharmarajan

1 Introduction

In December 2018, FDA released an FDA's RWE framework (henceforth called Framework) [1]. The framework defines RWD as "data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources," and RWE as "the clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD." Examples of RWD in the Framework include data derived from EHR, medical claims and billing data, data from product and disease registries, patient-generated data, and data from other sources, such as mobile devices. The Framework further indicates that RWD sources can be used for data collection and to develop analysis infrastructure to support many types of study designs to derive RWE, including, but not limited to, randomized trials (e.g., large simple trials, pragmatic clinical trials) and observational studies (prospective or retrospective).

As mentioned in chapter "Key Considerations in Forming Research Questions and Conducting Research in Real-World Setting", RWD can be prospectively generated by designing a prospective RW study (RWS). In addition, if there are relevant existing RWD sources, one can design a retrospective RWS to utilize such existing RWD sources. These different RWD sources come with different strengths and limitations. For example, the scope of claims data may contain broad information from all doctors and providers caring for a patient, whereas EHR may only be limited to the portion of care provided by doctors using the specific EHR

W. He (✉) · Z. Zhang
Medical Affairs and Health Technology Assessment Statistics, Data and Statistical Sciences, AbbVie, North Chicago, IL, USA
e-mail: weili.he@abbvie.com

S. Dharmarajan
FDA CDER, Silver Spring, MD, USA

W. He et al. (eds.), *Real-World Evidence in Medical Product Development* ,
https://doi.org/10.1007/978-3-031-26328-6_4

of a provider organization. On the other hand, claims only contain information as necessary for reimbursement (diagnoses, procedures, treatments), while EHRs comprise more complete medical picture (diagnoses, laboratory results, vital signs, doctors' notes). While prospective observational studies come with obvious inherent selection or information biases along with confounding bias due to non-randomized nature, pragmatic randomized controlled trials (RCTs) may still result in bias due to intercurrent events and missing data during a study even given randomization [2, 3]. Therefore, we believe that while assessment of fit-for-use RWD sources may be disease and research question-specific, the general approach for such assessment should be applicable to common RWD sources.

Numerous literature [4–10] have discussed different aspects and criteria for the fitness of RWD sources. In the white papers published by Duke Margolis [6–8], they summarized data relevancy as an "assessment of whether the data can adequately address the regulatory questions, in part or whole," and indicated relevancy as including the aspects of representativeness of the population of interest and whether the RWD source contains key variables and covariates. Data reliability "considers whether the data adequately represent the underlying medical concepts they are intended to represent," and speaks to data completeness, conformance, and plausibility. Many of these concepts brought up by the various authors as referenced above could be summarized in general as data relevancy or data reliability following Duke Margolis's white papers. Following the principles as suggested in the Duke Margolis papers, He et al. [11] further summarized key elements of fit-for-use data sources as including the following aspects: (1) relevant patient population supporting relevant clinical questions; (2) adequacy in recording and validation of key exposure and outcome variables along with confounders in terms of accuracy and correctness of data types, ranges of values, consistencies between independent values that measure similar attributes, (3) availability of complete exposure window, (4) longitudinality of data, (5) sufficient number of subjects, (6) data completeness, (7) availability of key data elements of patients for linkage of different data assets, (8) provenance in terms of transparent and traceable data sources, (9) extent of data curation and processing, and (10) data conversion and storage that adheres to appropriate standards.

Levenson et al. [12] provided an in-depth discussion on issues related to data integrity, principles and approaches to ascertain key variables from RWD, principles and approaches to validating outcomes and addressing bias from RWD, and key considerations in determining fit-for-use RWD. Their discussions align well with the recent draft guidance documents from FDA on assessing electronic health records and medical claims data or registries to support regulatory decision-making for drug and biological products [13, 14]. They further proposed a stepwise semi-quantitative approach to assess fit-for-use RWD sources with the use of quantitative measures for relevancy and reliability [12]. The idea is to first assess relevancy as the first dimension that includes variables related to disease population, response/outcome, treatment/exposure, confounders, time frame, and generalizability as to the representativeness of the underlying disease population. If the relevancy assessment yielded major gaps in data relevancy to answer a specific research question, then

there is no point in assessing the second dimension for reliability that includes quality relating to validity of the data elements, logical plausibility and consistency, and completeness of the data, including amount of missing data for key variables.

Since such assessment of fit-for-use RWD sources is disease and research question-specific, Levenson et al. [12] did not apply the principle and conceptual approach to a real RWD source. In this chapter, we attempt to apply and operationalize the semi-quantitative approach as proposed by Levenson et al. [12] to a hypothetical research question using a real RWD source. We share our learnings and best practices on such application.

The FDA/Harvard RCT DUPLICATE Initiative [15] proposed a structured process to assess the ability of using existing RWD sources, collected for other purposes such as Claim databases, to duplicate results with those from RCTs. Although the purpose of the duplicate project may be different from our own in this chapter, the findings and conclusions from that project could provide insights on the type of RCTs and type of outcomes that may be suitable for fit-for-use RWD sources. In Sect. 2, we review the duplicate project and provide our analysis of the learnings and insights. Further, we review a few typical RWD sources, including EHR, claims, registry, survey, NCI Surveillance, Epidemiology, and End Results (SEER) Program registries, and CDC National Health and Nutrition Examination Survey (NHANES) to provide additional insights on different RWD sources fitting different research goals. Section 3 is devoted to the application of the semi-quantitative approach to real RWD sources. The final section provides discussions and concluding remarks.

2 Gaining Insights on Aligning Research Questions with RWD Sources

2.1 Learning from RCT DUPLICATE Initiative

Regulators are evaluating the use of non-interventional RWS to assess effectiveness of medical products. The RCT DUPLICATE initiative (Randomized Controlled Trials Duplicated Using Prospective Longitudinal Insurance Claims: Applying Techniques of Epidemiology) [15] uses a structured process to design RWS emulating RCTs and compares results. The Initiative was funded by the US FDA to Brigham and Women's Hospital. They initially identified 40 RCTs that were conducted to support regulatory decision-making and estimated that 30 attempted replications would be completed after feasibility analyses. They used Optum Clinformatics Data Mart beginning in 2004, IBM MarketScan beginning in 2003, and Medicare Parts A, B, and D, across varying time ranges for select therapeutic areas for the replication project. To identify RCTs for replication, they cited Hernan [16] and considered the following design elements in consideration of the RWD

sources, and exclude RCTs in which some of these key design features cannot be replicated in RWD sources:

- Large, randomized trials with relatively simple treatment protocols, which are more likely to be replicable with RWS.
- The primary outcomes that are objectively measured are likely to be captured in the claims, such as myocardial infarction or stroke. Endpoints that are surrogate or symptomatic in nature are less likely to be captured in claims.
- For RCTs that include major inclusion/exclusion criteria that cannot be discerned from claims databases, they are excluded.
- While randomization cannot be replicated in claim databases, it is important to identify and ensure that important potential predictors of outcomes, including demographics, disease severity and history, concomitant medication, and intensity of healthcare utilizations are ascertained in the claims databases, so that they can be balanced in design or analysis stages.

First results from the RCT DUPLICATE Initiative were published in 2021 [17]. Results of replication for three active-controlled and seven placebo-controlled RCTs were reported. To assess RCT–RWE agreement, the authors used three binary agreement metrics: (1) "regulatory agreement" was defined as the ability of the RWE study to replicate the direction and statistical significance of the RCT finding, (2) "estimate agreement" was defined as the RWE hazard ratio estimate that was within the 95% confidence interval for the RCT estimate, and (3) hypothesis testing to evaluate whether there was a difference in findings by calculating the standardized difference between the RCT and RWE effect estimates.

The authors [17] found that although identifying the magnitude and direction of residual bias attributable to the nonrandomized study design is the key objective of calibrating RWE against RCTs, limitation of available RWD in emulation of other design features remain and need to be minimized. Further, even though attempts were made to emulate the features of each RCT as closely as possible, including inclusion and exclusion criteria, exposures, and outcomes, the constraints of the healthcare databases still made exact emulation impossible. In addition, close emulation of placebo is impossible via RWD and, hence, any placebo controlled RCTs for emulation for this reason. Further noted is that adherence to medications used in routine care is often very poor, and the analysis for the replication project used on-treatment approach that censor patients at treatment discontinuation whereas RCTs often use intention-to-treat approach for analysis. As a result, the follow-up time and the opportunity to capture longer term outcomes differ. The authors further indicated that the use of claims data, which lack clinical details but provide longitudinal data across the care continuum, affected the agreement between RCT and RWE findings. Other RWD sources, such as EHR and patient registries, would almost certainly have led to different results, as they often have detailed clinical information that may improve confounding adjustment. Patorno et al. [18], who used RWD to predict findings of an ongoing phase IV cardiovascular outcome trial, made a few similar points.

Although our purpose is not to replicate any RCTs using RWD but to assess fit-for-use RWD source, the insights the investigators of the RCT DUPLCATE Initiative provided are very helpful in focusing our attention to important design and data elements for assessment to address specific research questions. In Sect. 2.2, we review a few specific databases and provide our assessment of the type of research questions these RWD sources could be determined fit-for-use to answer.

2.2 Further Insights on Aligning Research Question with RWD Sources

The data owner of the ConcertAI database is one of the leaders in enterprise Artificial Intelligence (AI) and RWD solutions for life sciences and health care. ConcertAI has the largest network of over 400 oncology centers across the United States and its database contains de-identified EHR of more than 4.5 million patients treated by 1100 hematologist or oncologists. Although clinical details, such as biomarker or pathology, may be unstructured and requires curation, the patient clinical charts in oncology EHR are used by clinics to track patient care and therefore are gold standard for clinical details in oncology. Across multiple EHR systems, ConcertAI has business associate relationships with clinics, which enable ConcertAI comprehensive access to EHR and unstructured data to better standardize and curate clinical data. This makes ConcertAI database a representative data source of cancer care in the US population. Further, approximately 50 percentage of patients' information in the ConcertAI EHR data are also linked to claims data, which enriches ConcertAI as a RWD source.

In 2021, ConcertAI began a five-year collaborative research program with the US FDA. This collaboration will derive RWE across a number of clinical and regulatory use cases through utilizing ConcertAI's oncology RWD and advanced AI technology solutions. ConcertAI's oncology RWD contains millions of patients' EHR from a variety of academic and community cancer care settings. In this chapter, we will use acute myeloid leukemia (AML) and their treatments to showcase the assessment of fit-for-use RWD in session 3.

Optum Clinformatics Data Mart (Optum) is one of the largest claims databases. It is de-identified and derived from a database of administrative health claims for members of a large healthcare company affiliated with Optum. The population is geographically diverse and spans all 50 states. In addition to medical claims and pharmacy claims, Optum claims data include information with member eligibility and inpatients confinements, along with standard pricing for all outpatient claims, pharmacy claims, and inpatient confinements. This database comprises both commercial and Medicare Advantage health plan data, and therefore is useful for healthcare research institutes to address healthcare challenges, such as cost of care and healthcare utilization. Optum is also widely used by pharmaceutical companies to conduct scientific research to evaluate the clinical and economic value of their

products and medical devices. For instance, Optum data could be used to evaluate the patients' unmet need for certain conditions, and then investigate if their products could fill the gap in care. Optum data is a good option to understand the treatment patterns and medication adherence, discontinuation, and switching. It is also useful to assess comparative effectiveness for certain medications. All the research of treatment patterns and medication adherence, discontinuation, switching, and comparative effectiveness in real-world setting is important to understand and improve the gaps in care.

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) is an important source of information for cancer incidence and survival in the United States. SEER already includes cancer incidence and survival data from various cancer registries covering approximately 48% of the US population. The data about cancer patient demographics, primary tumor site, tumor morphology and stage diagnosis, first line of treatment, and follow-up of vital status is regularly collected into the SEER program registries. The SEER data includes the comprehensive information of stage of cancer at the time of diagnosis and survival data and is associated by age, sex, race, year of diagnosis, and geographic areas. Many research activities are developed based on the SEER data, such as cancer prevention and control, pattern of care and quality of care studies.

The National Health and Nutrition Examination Survey (NHANES) is a major program developed by the National Center for Health Statistics (NCHS), part of the Centers for Disease Control and Prevention (CDC). This program is designed to assess the health and nutritional status of adults and children in the United States and unique since it comprises interviews and physical examinations. The NHANES program has been conducted as a series of surveys for various population groups or health topics since the early 1960s. To meet the emerging needs, the survey has become a continuous program and focused on different health and nutrition measurements since 1999. Each year, the survey inspects a representative sample of approximately 5000 persons from different counties across the country and NHANES will visit 15 of the counties each year. The NHANES interview comprises demographic, socioeconomic, dietary, and health-related questions. The medical, dental, physiological measurements, and laboratory tests were collected during the NHANES examination. The survey contents include the data on the prevalence of chronic conditions and risk factors that may increase the chances of developing a certain condition. The survey also collects smoking, alcohol consumption, sexual practices, drug use, physical fitness and activity, weight, and dietary intake. Seventeen diseases, medical conditions, and health indicators are studied in NHANES, including anemia, cardiovascular disease, diabetes, environmental exposures, eye diseases, hearing loss, infectious diseases, kidney disease, nutrition, obesity, oral health, osteoporosis, physical fitness history and sexual behavior, respiratory disease (asthma, chronic bronchitis, emphysema), sexually transmitted diseases, vision. Due to the comprehensive health information, the data from the survey has been widely used to determine the prevalence of major diseases and risk factors for disease and assess the nutritional status and its association with health promotion and disease prevention. The NHANES data has been used to evaluate the relationship between

environmental chemical exposures and adverse health outcomes. For instance, the association between US population levels of chemicals in blood and/or urine and biochemical indicator is extensively investigated using NHANES.

3 Semi-Quantitative Approach for Fit-for-Use RWD Assessment – Application of a Case Study

3.1 Estimand Related to Fit-for-Use RWD Assessment

As ICH E9 (R1) [2] indicated, the "central questions for drug development and licensing are to establish the existence, and to estimate the magnitude, of treatment effects: how the outcome of treatment compares to what would have happened to the same subjects under alternative treatment (i.e., had they not received the treatment or had they received a different treatment)". The estimand framework in E9 (R1) includes five attributes as follows:

1. *Population*. Patients targeted by the clinical question
2. *Treatment*. The treatment condition of primary interest (e.g., new drug) and, as appropriate, the alternative treatment condition to which comparison will be made (i.e., comparator)
3. *Variable* (or endpoint). The endpoint obtained from each patient to be used to address the clinical question
4. *Intercurrent events* (ICEs). Events occurring after treatment initiation that affect either (1) the interpretation or (2) the existence of measurements of endpoints associated with the clinical question of interest
5. *Population-level summary*. A summary measure for the endpoint that provides a basis for comparison between treatment conditions

As the fit-for-use RWD assessment is very much disease and research question-specific, the focus of our assessment should be aligned with the estimand framework in addressing certain research questions as discussed in chapter "Estimand in Real-World Evidence Study: From Frameworks to Application" of this book.

3.2 Evaluation of Key Variables as Determined by Research Questions

Levenson et al. proposed a stepwise semi-quantitative approach to assess fit-for-use RWD sources with the use of quantitative scores for relevancy and reliability [12]. As the fit-for-use RWD assessment is very much disease and research question specific, Levenson et al. developed a set of principles (see Table 1 of their paper).

Using a global multiple sclerosis (MS) cohort study RWD source, Kalincik et al. [9] proposed error rate, data density score, and generalizability score using the MS database (MSBase). The data density score was calculated across six domains (follow-up, demography, visits, MS relapses, paraclinical data, and therapy). The error rate evaluated syntactic accuracy and consistency of data. The generalizability score evaluated believability of the demographic and treatment information. Correlations among the three scores and the number of patients per center were evaluated. The authors believe that this evaluation process will facilitate further improvement of data quality in MSBase and its collaborating centers. It will also enable quality-driven selection of research data and will enhance quality and generalizability of the generated evidence.

In Sect. 3.3, we apply the set of principles in evaluating relevancy and reliability of an RWD source as proposed by Levenson et al., along with the implementation example by Kalincik et al. to a case study. We will use a hypothetical research question as described in Sect. 3.3, but which could also be of real clinical research interest.

3.3 Hypothetical Research Question and Quantitative Assessment Algorithms

For our hypothetical research question, we have the following research plan as shown below. For the purpose of providing an implementation exercise, we blinded the actual treatment patients received in our application, but these treatments are real treatments in the RWD source we use.

Research question

- To assess the long-term effectiveness of AML patients treated with drug A vs. drug B in overall survival at 2 years

Research objectives

In patients treated with drug A vs. drug B:

- To assess the long-term effectiveness of drug A vs drug B in AML patients

Hypothesis

- Treatment with drug A will result in improved overall survival as compared to treatment with drug B in patients with AML after 2 years

Study Design

In this implementation exercise, AML patients are identified from ConcertAI using ICD 9 and ICD 10 codes. The study cohort is defined as the AML patients taking on drug A or drug B as first-line treatment.

See Table 1 for more detailed study design features. To make the assessment for fit-for-use RWD source to answer our research questions, we developed the

Table 1 Study design

Index date	Initial treatment date of drug A or drug B occurred at least 1 year after the patient's first encounter in the database to allow larger number of patients to be included in this study cohort
Baseline	6 months prior to the index date
Follow-up period	Maximum of 2 years post the index date
Treatment	Drug A and drug B
Outcome	Overall survival
Censoring rules for a subject in the specified order on the right	1. Switching between drug A and drug B within 2 years' follow-up 2. Loss to follow up in the database within 2 years follow-up 3. Two years' of follow-up in the database without outcome event of death
Inclusion criteria	Diagnosis of AML 18+ at diagnosis date Entered ConcertAI at least 1 year before the index date Started drug A or B as first line of treatment
Exclusion criteria	AML in relapse
Study variables	*AML* ICD 9: 205.0205.00205.01206.0206.00206.01207.0207.00207.01207.21 ICD 10: C92.0 C92.00 C92.01 C92.4 C92.40 C92.41 C92.5 C92.50 C92.51 C92.6 C92.60 C92.61 C92.62 C92.A C92.A0 C92.A1 C93.0 C93.00 C93.01 C94.0 C94.00 C94.01 C94.2 C94.20 C94.21 C94.4 C94.40 C94.41 *Confounding variables*[a] Age Gender Race Body mass index (BMI) The eastern cooperative oncology group (ECOG) performance status scale Concomitant treatments: Posaconazole Comorbidities: Diabetes, CHF[b]
Sample size	The study with approximately 2360 patients, 1180 in each group, will have 80% power to detect a 6% improvement in survival rate between drug A and drug B at 2-year, alpha = 0.05, 2-sided. This assumes that the treatment with drug B has an overall survival rate of 60% at 2-year, and 15% patients may be lost to follow-up or switched during the 2-year follow-up period.

[a] *ECOG* Eastern Cooperative Oncology Group (ECOG) performance status
[b] *CHF* Congestive Heart Failure

following algorithms to derive the assessment measures, as shown in Tables 2 and 3, respectively. Relevancy assessment as shown in Table 2 is to specifically address the adequacy of data elements as defined in Estimand for a specific research question, such as population, treatment, outcome, confounders, and time, as indicated by the hypothetical research question and study design in Table 1. The denominator for Disease Population in Table 2 includes all the patients in ConcertAI that we have access to, whereas the denominators for all other measures are specific to the study

Table 2 Assessment of Relevancy of Fit-for-Use RWD Source

Dimensions	Variable	Assessment score
1st assessment dimension: Relevancy based on a specific research question		
Disease population	ICD9 ICD10	%patients (pt) in the population = (# of patients meeting the disease condition / # of patients in the database[a])*100
Response/outcome	Overall survival	%pt with events = # of patients with outcome event of death within 2 years / # of patients in the study cohort)*100 %pt with switch = (# of patients switching drugs within 2 years / # of patients in the study cohort)*100 %pt censored = (# of patients censored with follow-up <= 2 years / # of patients in the study cohort)*100 %pt censored after 2 yrs = (# of patients censored with follow-up >2 years / # of patients in the study cohort)*100
Treatment/exposure	Drug A, drug B	For each treatment, we calculate the score before switching treatment separately Score A = (# of patients receiving drug A / # of total patients in the study cohort)*100 Score B = (# of patients receiving drug B / # of total patients in the study cohort)*100
Confounders	Age, gender, race, BMI, ECOG, Posaconazole, diabetes, CHF	For each identified key confounder, we checked on whether potential confounding variables were collected in the study cohort
Time	Description of follow up time from study entry to censoring in the database	The intent is to describe time duration for treatment/exposure, response/outcome, and any time varying confounders and whether it is sufficient to address the research questions
Generalizability score	Describe the representativeness of the disease population in the RWD source. Information on demographics and disease-specific indicators may be used	Male to female ratio based on the epidemiology of AML Age range of AML prevalence Reported prevalence of general AML population as compared to the measure in the dataset

[a]The denominator is the total number of patients in the ConcertAI database with access

cohort for the hypothetical research question. Further, since most research based on existing RWD sources are based on non-randomized groups, if effectiveness comparison is the focus, ascertainment of potential confounding factors is critical. As intercurrent events are more prevalent in RWD sources, reliability assessments in Table 3 assess data quality and consistency relevant to the relevancy parameters in Table 2.

For the assessment in Table 3, the assessment is specific to the study cohort at baseline and key time points during follow-up period.

Table 3 Assessment of reliability of fit-for-use RWD source

2nd assessment dimension: Reliability based on a specific research question		
Quality	Assess the validity of the data elements, checking the logical plausibility of the data (e.g., a lab result is within the limits of biological possibility), and examining the data consistency for each patient (within related data fields and over time) as well as the conformance of the data to any applicable internal standards or external data models	*Completeness* = proportion of missing data for key variables such as confounding variables as listed above *Syntactic accuracy* = proportion of critical variables with values corresponding to their range *Consistency* = proportion of the recorded variables congruent with other recorded variables (we could think of a few such variables, such as BMI = 40 when weight and height are not extreme)
Data density score	Assess the amount of information as represented by data density, such as follow-up, clinical visit, and symptoms or outcome ascertainment, standardized as patient-year of follow-up over the planned study follow-up duration	*Cumulative follow-up* = median follow up in year as a standardized measure *Clinical visit* = (sum of #visits) / (sum of all exposure time of each patient in year)*2 as a standardized measure *Disease symptoms* = (sum of #reported symptoms) / (sum of all exposure time of each patient in year)*2 *Laboratory test* = (sum of #lab tests) / (sum of all exposure time of each patient in year)*2

3.4 Results

Based on ICD 9/10 codes of AML, we first extracted a subset of the dataset with 19,064 AML patients. Two treatments of AML: drug A and drug B, were selected for assessment of data relevancy and data reliability in this study. The data include AML patients with diagnosis date from 1985 to 2021. Only the patients who entered ConcertAI at least 1 year before the initiation of drug A or drug B as the index date and whose age was $\geq$18 on the index date were included in this study and the patients with AML in relapse were excluded. Finally, 2366 AML patients were identified in the final study cohort, of which 737 AML patients were treated with drug A and 1629 AML patients were treated with drug B. The maximum of 2 years post the index date was the follow-up period. The objective was to evaluate the overall survival, as the outcome, of AML patients with the treatment of drug A vs. drug B as the first-line treatment. The censoring rules for this study included (1) switching between drug A and drug B within 2 years' follow-up; (2) loss to follow up in the database within 2 years' follow-up; and (3) 2 years of follow-up in the database without outcome event of death. We will assess the relevancy

and reliability of the subset of the patients' data from ConcertAI that fit with the inclusion/exclusion criteria, as stated in Table 1.

3.4.1 Relevancy

Assessment of Parameters Related to Estimand Attributes

For patients in ConcertAI with AML, 12.41% (2366/19064) of them fit the study design inclusion criteria and received drug A or drug B treatments. Within the 2-year follow-up period, 56.26% (1331/2366) AML patients were deceased, 0.21% (5/2366) AML patients switched between drug A and drug B (for simplicity issue, we only considered switching between Drug A and B), and 13.06% (309/2366) AML patients were censored for loss to follow-up. In addition, 30.47% (721/2366) AML patients were censored after 2 years' follow-up.

In this study cohort of patients, 31.15% (737/2366) received drug A as first line treatment and 68.85% (1629/2366) drug B as first line treatment, respectively. Eight confounders with the assessment scores were identified for the cohort from ConcertAI. Patients with missing confounding variables were assessed, and the results are as follows (% shown as available data):

- Five patient characteristics: age (100%), gender (100%), race (80.81%), BMI (79.37%), ECOG (17.58%)
- One concomitant treatment: Posaconazole (12.85%)
- Two comorbidities: diabetes (13.23%) and CHF (6%)

Generalizability

Acute myeloid leukemia (AML) is a malignant disorder of the bone marrow which is characterized by the clonal expansion and differentiation arrest of myeloid progenitor cells. The age-adjusted incidence of AML is 4.3 per 100,000 annually in the United States (US). Incidence increases with age with a median age at diagnosis of 68 years in the United States. Differences in patient outcomes are influenced by disease characteristics, access to care, including active therapies and supportive care, and other factors. AML is the most common form of acute leukemia in adults and has the shortest survival (5-year survival = 24%) [19].

In this study cohort, the male to female ratio is 1.24 (1308/1058), which is very close to the estimated ratio 1.25 in the United States [20]. However, the average age of AML diagnosis is younger (63 ± 14 years) in the study cohort than that (68 years) in the United States. It is uncommon that the people are diagnosed with AML before age 45. However, 11.75% (278/2366) patients were diagnosed with AML before age 45 in the study cohort.

In addition, 0.8% (19,064/2,192,910) AML prevalence of cancers in ConcertAI is close to the 1% AML prevalence of cancers in the United States. The 2-year

survival rate for the AML patients since the treatment initiation in this study cohort is 40.3%, which is slightly higher than the estimated 34.1% (since diagnosis) for AML patients of all ages as posted on the SEER website at NCI [21].

3.4.2 Reliability

The quality and data density of this study cohort were evaluated for reliability based on the definition in Table 3. Note that data quality, syntactic accuracy, and data consistency are research question and RWD source-dependent, and we provide a few illustrative assessment calculations specific to the hypothetical study design in Table 1.

Quality

Three quality metrics were evaluated for data quality:

- Completeness

  Age, BMI, ECOG, and Posaconazole treatment (concomitant treatment) were considered as the critical variables to evaluate the data completeness. For confounders in Table 2, all the AML patients had age values and decent number of patients (79.37%) had BMI value in the study cohort. But only a small number of patients in this cohort had ECOG scores (17.6%) or Posaconazole concomitant treatment (14.0%). Based on the specific research questions, practitioners could identify other or additional key confounding variables to check on magnitude of missingness.
- Syntactic accuracy

  In RWD, some patients had the death date prior to the last observation date in ConcertAI. This scenario is common with date in RWD. The reasons could be incorrect input of death date or the delayed data entries for prescription fill, laboratory test, diagnosis in EHR system, or the incorrect patient linkage. In this study cohort, 57.44% (1359/2366) AML patients had death date prior to the last observation date and the death date was set as the last observation date for these patients. Based on the specific research questions, practitioners could identify additional key study variables and check on syntactic accuracy, such as plausible range of values for these variables.
- Consistency

  To evaluate consistency, we chose BMI and related height and weight in deriving BMI. The height and weight of patients in the study cohort were extracted from the database and then we calculated the BMI based on the formula BMI = kg/m2. Among the 2366 AML patients in the study cohort, 71.4% (1689/2366) AML patients had height value(s) and 92.6% (2190/2366) AML patients had weight value(s) at the baseline. The height and weight closest to the index date were selected and BMI (calculated BMI) were obtained for

1687 patients. Among the 1878 AML patients with recorded BMI in the study cohort, only 1348 AML patients had both the calculated BMI. Among 1348 AML patients, 25% (337/1348) had the BMI difference greater than 1 between the calculated BMI and recorded BMI.

Data Density

Data density was evaluated across four domains in this study: cumulative follow-up, clinical visit, laboratory test, and disease symptoms, where fatigue, fever, and weight loss were considered the disease symptoms. This is a way to assess the time factor in data relevancy, how often patients' information is captured in the RWD source, and the ability of using such an RWD source to answer a research question.

Based on the censoring rules, the last visit was defined as the death date if the patient was deceased within the 2 years' follow-up period, or the date switching drug A and drug B within the 2 years' follow-up period, or censoring date within the 2 years' follow-up period, or the date of 2 years post the index date if the patient was deceased after 2 years post the index date or the last observation date was after 2 years post the index date.

- Cumulative follow-up

  The cumulative follow-up for each patient was defined as from index date to the censoring date in the database. The median follow-up time is 1.02 years for this study cohort.
- Clinical visit

  All encounter visits during the follow-up for each patient were collected. Based on the algorithm in Table 2, the standardized average number of clinical visit per patient is 31 visits in 2 years in this study cohort.
- Disease symptoms

  We selected fatigue, fever, and weight loss as disease symptoms for each patient in this study. Based on the algorithm in Table 2, the standardized score is 2.6, which may be underestimated than expected. The reason might be because the disease symptoms are usually not captured as structured data in clinical care setting but described in the clinical notes by physicians. To have more complete disease symptoms, natural language processing may be utilized to identify the disease symptoms from clinical notes.
- Laboratory test

  Due to the complications of laboratory test, we reviewed the laboratory test names and categorized the laboratory test as blood count, fluid test, urine test, prothrombin time (PT), specimen, stool, and other. The data density for laboratory test was assessed based on the categorized data of laboratory test. Based on the algorithm in Table 2, the standardized data density score for lab test is 12, i.e., on average, each patient from the study cohort had about 12 lab tests in 2 years.

4 Discussions and Conclusion

Following the semi-quantitative approach as developed by Levenson et al. [12] and the implementation example by Kalincik et al. [9], we have applied the concept and implemented the same in a case study using ConcertAI database.

In assessing relevancy, we started with assessing the number of relevant diseased patients in the database. If there is insufficient number of subjects fitting the disease conditions in the database as required in addressing the research questions and the associated sample size needed, then there is no point in moving forward with additional assessment for this data source. This is what we coined as a stepwise semi-quantitative approach. Next, we checked on whether certain key variables for relevancy assessment were collected in the database and how much missingness for such key variables. For this hypothetical research question, we found that there are a couple of key confounding variables with a large amount of missing data, such as ECOG status. For posaconazole use or diabetes status, often as the convention in RWD source, clinicians would assume that not ascertaining posaconazole use or diabetes status meant that patients were not using posaconazole or their diabetes status is no. However, the assumption needs to be carefully assessed. Further, to assess whether the results from this research based on ConcertAI could be generalized to the AML patient population at large, we selected a few key epidemiologic factors, such as male to female ratio, incidence, and death rate in this cohort of patients. As can be seen, the assessment of generalizability is greatly dependent on the research question and disease under study. Practitioners should make your own judgement on how this dimension could be assessed.

In assessing reliability, from completeness, syntactic accuracy, and consistency perspectives, we also selected a few key variables that are relevant to the research questions at hand. Practitioners should select variables they deemed important to answer their research questions. This is what we did, by selecting a few key variables to demonstrate completeness, syntactic accuracy, and consistency, with the understanding that it may be not possible to check all the variables that were collected in a database. We modified the concept of data density, as originally proposed by Kalincik et al. [9], to fit with our research question. We believe that the concept of data density is a very important one, to gage on the richness of the database, longitudinality of the follow up, and frequency of key information ascertained. Practitioners could identify fit-for-use data density measures of their own based on the principles as discussed in this section.

In conclusion, we implemented a case study to assess a fit-for-use RWD source to answer a hypothetical research question. The assessment revealed that this RWD source may not be fit-for-use for the research question due to the following data relevancy and/or reliability issues we identified:

- A few key confounding variables have a large amount of missing data.
- The median follow-up time is only 1.02 years for a 2-year study. It means that 50% of patients in this cohort has discontinued at 1 year, making assessment of overall survival at 2 years quite inaccurate.

However, we believe that such assessment may have an element of subjectivity to it. Sponsors may need to engage regulators to have such a discussion on the data source, providing rationale and justification based on the guiding principles and assessment details, as delineated in this chapter.

Further research and additional implementation case studies may still be needed to guide practitioners to fully understand the development of assessment algorithms based on specific research questions, the rationales and justification for assessment, and any available regulatory feedback on RWD sources that were deemed fit-for-use or otherwise. In chapter "Applications Using Real-World Evidence to Accelerate Medical Product Development", six case studies are presented. Of a few case studies in that chapter, FDA reviewers deemed the RWD sources as adequate and fit-for-use. However, additional details on the justification of making such a conclusion are not available to the chapter authors for further delineation in the chapter.

Disclaimer This chapter reflects the views of the authors and should not be construed to represent FDA's views or policies.

References

1. FDA, *Framework for FDA's Real-world Evidence Program.* 2018.
2. ICH, *ICH E9(R1) Addendum to Statistical Principles for Clinical Trials on Choosing Appropriate Estimands and Defining Sensitivity Analyses in Clinical Trials.* 2020.
3. Westreich, D., *Berkson's Bias, Selection Bias, and Missing Data.* Epidemiology, 2012. **23**(1): p. 159–164.
4. BIO, *Incorporating Real-world Evidence Within the Label of an FDA approved Drug: Perspectives from BIO Membership.* 2019.
5. N. A. Dreyer, P. Velentgas, K. Westrich, and R. Dubois, *The GRACE Checklist for Rating the Quality of Observational Studies of Comparative Effectiveness: A Tale of Hope and Caution.* Journal of Managed Care Pharmacy, 2014. **20**(3): p. 301–308.
6. Duke Margolis, *Characterizing RWD Quality and Relevancy for Regulatory Purposes.* Duke-Margolis Center for Health Policy, 2018.
7. Duke Margolis, *Determining Real-world Data's Fitness for Use and the Role of Reliability.* Duke-Margolis Center for Health Policy, 2019a.
8. Duke Margolis, *Understanding the Need for Non-interventional Studies Using Secondary Data to Generate Real-world Evidence for Regulatory Decision Making, and Demonstrating Their Credibility.* Duke-Margolis Center for Health Policy, 2019b.
9. T. Kalincik, J. Kuhle, E. Pucci, J. I. Rojas, M. Tsolaki, C-A. Sirbu, et al., *Data quality evaluation for observational multiple sclerosis registries.* Multiple Sclerosis Journal, 2017. **23**(5): p. 647–655.
10. S. V. Wang, S. Schneeweiss, M. L. Berger, J. Brown, F. de Vries, I. Douglas, et al., *Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0.* Pharmacoepidemiology and Drug Safety, 2017. **26**(9): p. 1018–1032.
11. W. He, Y. Fang, H. Wang, I. Chan, *Applying Quantitative Approaches in the Use of RWE in Clinical Development and Life-Cycle Management.* Statistics in Biopharmaceutical Research, 2021. **0**(0): p. 1–12.
12. M. Levenson, W. He, L. Chen, S. Dharmarajan, R. Izem, Z. Meng, et al., *Statistical consideration for fit-for-use real-world data to support regulatory decision making in drug development.* Submitted to SBR, 2022. https://doi.org/10.1080/19466315.2022.2120533

13. FDA, *Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products.* 2021a.
14. FDA, *Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products.* 2021b.
15. J. M. Franklin, A. Pawar, D. Martin, R. J. Glynn, M. Levenson, R. Temple, et al., *Nonrandomized Real-World Evidence to Support Regulatory Decision Making: Process for a Randomized Trial Replication Project.* Clinical Pharmacology & Therapeutics, 2020. **107**(4): p. 817–826.
16. Hernan, M.A. and J.M. Robins, *Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available.* American Journal of Epidemiology, 2016. **183**(8): p. 758–764.
17. J. M. Franklin, E. Patorno, R. J. Desai, R. J. Glynn, D. Martin, K. Quinto, et al., *Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies.* Circulation, 2021. **143**(10): p. 1002–1013.
18. E. Patorno, S. Schneeweiss, C. Gopalakrishnan, D. Martin, and J. M. Franklin, *Using Real-World Data to Predict Findings of an Ongoing Phase IV Cardiovascular Outcome Trial: Cardiovascular Safety of Linagliptin Versus Glimepiride.* Diabetes Care, 2019. **42**(12): p. 2204–2210.
19. R. M. Shallis, R. Wang, A. Davidoff, X. Ma, and A. M. Zeidan, *Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges.* Blood Reviews, 2019. **36**: p. 70–87.
20. Cancer.Net, *Leukemia – Acute Myeloid – AML: Statistics.* 2022.
21. NCI SEER, *Acute Leukemia Myeloid (AML) SEER Relative Survival Rates by Time since diagnosis, 2000-2018.* 2022.

Key Variables Ascertainment and Validation in RW Setting

Sai Dharmarajan and Tae Hyun Jung

1 Introduction

In real-world data, it is important to assess whether the outcome of interest is being correctly captured and captured consistently in a way that it is accessible. For example, in imaging data, the frequency of imaging evaluation should be adequate to provide a reasonably precise measure and to enhance the consistency of image assessment [1]. At the same time, these real-world data should be accurately ascertained. Ascertainment is difficult not only because there are multiple types of outcomes but also their methods for ascertainment vary by data sources.

More generally, studies in real world data sources must first include a conceptual definition for key variables that define the inclusion and exclusion criteria of study population, exposure, outcome, and key confounders. The conceptual definition should reflect the current clinical or scientific thinking about the variable. For example, this could be the clinical criteria to determine if a patient has a condition that defines the study population, outcome, or key covariate or the measurement of drug intake that defines the exposure. Based on the conceptual definition, an operational definition should then be developed to extract the most complete and accurate data from the data source. An operational definition essentially translates a theoretical, conceptual variable of interest into a set of specific operations or procedures that define the variable's meaning within a specific study and available data sources. In many studies using electronic health record (EHR) or medical claims data, the operational definition will usually be an algorithm constructed using structured data elements such as codes indicating the presence of a diagnosis, medical procedure or medication dispensation. For example, for identifying the

S. Dharmarajan · T. H. Jung (✉)
Division of Biometrics VII, Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, Food and Drug Administration, White Oak, MD, USA
e-mail: TaeHyun.Jung@fda.hhs.gov

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_5

presence of diabetes, an operational definition may be an ICD code for the diagnosis of diabetes. In some other instances, the algorithm may be constructed using relevant information derived from unstructured data occurring as free-text such as clinical progress or discharge notes in combination with structured data elements. Additional linked data, such as a patient survey, may also be used to specify an operational definition.

The operational definition is often called the phenotype definition, with the underlying clinical characteristic or concept being the phenotype. When a clinical characteristic can be ascertained using an operational definition solely from the data, either using structure or unstructured elements, in EHRs or any other clinical data repository (including disease registries, claims data) it is called a computable phenotype [2]. The word computable stemming from the fact that these can be ascertained using a phenotype definition composed of data elements and logic expressions (AND, OR, NOT) that can be interpreted and executed by a computer, without the need for human intervention in the form of a chart review. Computable phenotypes along with their definitions are important as they can be standardized to facilitate identification of similar patient populations and enable efficient selection of populations for large-scale clinical studies across multiple health care systems and data sources [2].

The development of phenotype definitions is discussed in detail in the next section, but it is important to note that computable phenotype definition should include metadata and supporting information about the definition, its intended use, the clinical rationale or research justification for the definition, and data assessing validation in various health care settings [3]. In terms of regulatory considerations, the computable phenotype definition should be described in the protocol and study report and should also be available in a computer-processable format. Clinical validation of the computable phenotype definition should be described in the protocol and study report [4].

In the subsequent sections of the chapter we first provide an overview of available and commonly used methods for ascertainment of key variables, followed by a discussion of the importance and role of validation. We then lay out some special considerations for three types of key variables: exposure, outcome, and confounders. This is followed by a detailed discussion of a published example of RWE in the post-market setting. Specifically, we walk through the ascertainment and validation of key variables in studies conducted using RWD sources to successfully fulfill a post-marketing requirement. Finally, we present a discussion of the key takeaways and important learnings.

2 Methods for Ascertainment

Identifying patients with certain clinical characteristics of interest (outcome, exposure or other key variable used for cohort definitions) in real world data sources require looking for patterns throughout the patient's record suggestive of those

characteristics. Here, we describe the methods used for ascertainment of clinical characteristics in data sources where direct ascertainment of the characteristics through a single variable is not possible, most notably in EHR and administrative claims databases.

2.1 Rule-Based Methods

The traditional approach to ascertainment has involved specifying inclusion and exclusion criteria or rules based on structured data elements such as diagnosis codes, medications, procedures, and lab values using criteria often drawn from consensus guidelines around diagnosis and treatment [5]. These methods are often termed as rule-based methods. A well-established example is the identification of Type 2 Diabetes for which the requirement may include at least one mention of the diagnosis code, evidence of at least one hypoglycemic medication, or an HbA1c above a certain threshold [6]. Oftentimes, multiple instances or mentions of diagnoses, or the occurrence of a diagnosis along with a medication or lab value are required to ensure that "rule-out" diagnoses that are recorded for further confirmation aren't incorrectly identified as true diagnoses. Rule-based methods tend to do well when there are clear, reliable diagnosis and procedure codes that are used often or when there's a reliable surrogate or proxy. There has also been some concerted effort to improve the quality of rule-based phenotypes. Collaborations such as the eMERGE (Electronic Medical Records and Genomics) network [7] have developed a large catalog of generalizable EHR phenotypes, including hypothyroidism, type 2 diabetes, atrial fibrillation, and multiple sclerosis, and have created, PheKB (Phenotype Knowledgebase; available at http://phekb.org) [8], a repository which facilitates the sharing and validation of phenotypes in different health care settings and across different coding libraries (see chapter "Privacy-Preserving Record Linkage for Real-World Data" for more details on coding libraries). But the scope of rule-based approaches is limited in capturing more complex phenotypes or when working in less standardized datasets. For example, Kern et al. [9] found that rule-based queries for chronic kidney disease among diabetic patients had poor sensitivity with a maximum of 42% when using seven alternative ICD-9 diagnosis codes. In another instance, Wei et al. [10] showed that a rule for capturing type 2 diabetes did not identify many true positives when used at only a single site because patient data were often fragmented across inpatient and outpatient data repositories.

2.2 Machine Learning (ML)-Based Methods

An improvement on rule-based methods has been made recently by leveraging machine learning to combine numerous structured and unstructured data elements

into an algorithm for classifying patients with the clinical characteristic or phenotype of interest. The ML-based approaches to build a phenotype can broadly be categorized into supervised, semi-supervised, or weakly supervised, based on the requirement of gold standard-labeled data, a fraction of gold standard-labeled data, and silver standard-labeled data, respectively.

In supervised approaches, the data consist of a 'gold-standard' label of the presence or absence of the phenotype. These labels are usually annotated from a manual review of patient records but sometimes can also be derived from lab values or registry data. With these labels, an ML algorithm (e.g., random forest, Support Vector Machine, artificial neural net) is trained to classify patients with the phenotype using all the relevant data elements, usually spanning in the hundreds, as features in the model. For example, Gibson et al. [11] developed an algorithm for identifying Rhabdomyolysis cases in the IBM Watson EHR database, where laboratory data was leveraged to come up with gold standard labels. The best performing algorithm which combined information from diagnosis codes, procedure codes, and medication using a neural net had an AUC of 0.88. In another example, Carrell [12] developed a phenotype for Anaphylaxis using manually abstracted medical chart data as training data.

As manual chart abstraction is resource intensive and often infeasible, and other sources of gold-standard labels involve challenges of their own, including poor validity, other ML based methods such as weakly supervised or semi-supervised methods either completely do away with the requirement of gold standard labels or require only a limited amount of labeled data, respectively. To minimize the burden of chart review, semi-supervised methods train ML algorithms with a large amount of unlabeled data (e.g., unreviewed medical records), together with a small amount of labeled data. With phenotyping hypertension as an example, Henderson et al. [13] showed that these methods may slightly underperform compared to supervised learning methods, but may require only a fraction of the number of reviewed charts (e.g., AUROCsemi-supervised 0.66, AUROCsuperivsed 0.69 for hypertension).

In a weakly supervised method, a "silver-standard" or noisy label can be easily extracted from all available records in place of doing a chart review. This silver standard label is usually a highly predictive but imperfect proxy for the gold-standard, that is, they have a high positive predictive value, but weak sensitivity. For example, in a study of systemic lupus erythematosus, the silver standard label for patients having the condition was four or more disease-specific ICD-9 codes were present in their record [14]. Well-known examples of weakly supervised phenotyping methods include PheNorm [15] and the Automated PHenotype Routine for Observational Definition, Identification, Training, and Evaluation (APHRODITE) [16]. These two methods differ in their approach for constructing an ML algorithm using the silver standard labels. In methods like PheNorm, it is assumed that the silver-standard label follows a mixture model representing actual cases and controls. PheNorm specifically uses Gaussian mixture-modeling and denoising self-regression with silver standard labels based on counts of relevant billing codes such as diagnosis and procedure codes for the condition of interest and free-text mentions of the condition of interest in clinical notes [15]. With such a method, the authors showed

that PheNorm achieved comparable accuracy to penalized logistic regression trained with 100–300 gold-standard labels for four phenotypes [15]. In contrast, the anchor and learn framework of methods like APHRODITE uses supervised learning methods trained with a silver standard whose presence unambiguously indicates the presence of the condition, whereas the absence is uninformative. APHRODITE, which uses a logistic regression trained using the silver standard label, was applied to ten phenotypes across three Observational Health Data Sciences and Informatics (OHDSI) sites in the United States and Korea, and obtained mean recall (Positive Predictive Value; PPV) and precision (sensitivity) of 0.54 and 0.73 in the United States, and, 0.46 and 0.24 in Korea [16]. Regional difference in the quality of silver standard labels likely determined the difference in quality of model performance [16].

2.3 Text Processing for Phenotyping

As mentioned earlier, a certain amount of key clinical data in real-world data sources such as EHR databases occur in the form of free text (e.g., clinical notes) or as other non-standardized (e.g., images or radiology reports). Recent technological advances in the field of artificial intelligence, including natural language processing and deep learning, have enabled the extraction and use of this unstructured information to identify and ascertain clinical characteristics. The most common way to process text data has been to use an openly available NLP software or pipeline [17] to map clinical notes, say a discharge summary, into a bunch of medical concepts within the Unified Medical Language System (UMLS) metathesaurus [18]. These extracted medical concepts are then engineered into features (e.g., as the number of positive mentions of the phenotype in a patient's discharge summary) to be fed into an ML-based phenotyping approach mentioned in the previous section [19]. Another approach to process clinical notes gaining popularity recently is the use of word embeddings [20]. Word embeddings typically serve as the input layer to deep learning models, such as convolutional neural nets, for identifying a phenotype [21]. This approach has shown to have some advantages over the previously mentioned approach of extracting medical concepts and using them as features [19].

2.4 Ascertainment Through Linkage and Using Proxy

Data linkage of one RWD source to additional sources can be used to increase the amount of information available on individual patients, improving the capture of key variables of interest and providing additional data for validation purposes. For example, Zhang et al. [22] improved the capture of mortality for cancer patients in an EHR database through a linkage with obituary information. They then validated this composite by linking to the national death index data and showed sensitivity

above 80%, specificity above 93%, PPV above 96%, and negative predictive value above 75.0% across multiple cancer types. Data linkage such as performed in this study is deterministic, as in the linked records have an exact match to a unique or set of common identifiers, and the match status can be determined using a single or multiple step process. Different types of data linkage and other details are discussed in detail in chapter "Causal Inference with Targeted Learning for Producing and Evaluating Real-World Evidence" of this book.

Data linkage is also one way to address the problem of missing data. When data on a key variable are truly missing, it may be possible to identify a variable that is a proxy for this variable of interest. For example, low-income subsidy status under the Medicare Part D prescription drug program may serve as a proxy for a patient's socioeconomic status. Another example of the use of proxy measures is for the identification of a tumor burden endpoints such as an achievement objective response in real world data sources, where information on standardized clinical trial criteria such as Response Evaluation Criteria in Solid Tumors (RECIST) are not available. Griffith et al. [23] compared radiology-anchored and clinician-anchored approaches to RECIST-based methodology in an EHR data source and found the latter to be infeasible. This proxy has been used in RWE to support the approval of Ibrance for the indication of male metastatic breast cancer [24], a case which is discussed in chapter "The Use of Real-World Data to Support the Assessment of the Benefit and Risk of a Medicine to Treat Spinal Muscular Atrophy".

Regardless of the method used, it is important to ensure the validity [25] of the derived phenotype or operational definition in external data and the portability [26] of phenotypes across health systems and time periods before wide adoption. It must also be ensured that any algorithm used is not amplifying existing disparities in the healthcare system [27].

3 Validation

As operational definitions are usually imperfect in the sense that they will not accurately capture the variable or condition of interest for every subject in the data, steps should be taken to confirm their validity. The aim of doing so is to minimize the bias the mismeasurement and misclassification of key variables may cause in the findings of the study. In order to determine what steps need to be taken and for which variables, it is important to understand the implications of potential misclassification of a variable of interest. Thus, it is important to consider (1) the magnitude or degree of classification or measurement error; (2) whether the error is differential or non-differential (e.g., misclassification of outcome may occur unequally or equally by exposure), and, independent or dependent (e.g., misclassifications of exposure and outcome may be correlated when both are self-reported in the same survey); and (3) the direction toward which the results might be biased because of the error.

The most thorough way to minimize misclassification error, for example, is to conduct a complete verification of the variable by checking the variable for

Table 1 Performance measures of an operational definition for a binary variable

Condition based on operational definition	Condition based on reference standard/conceptual definition			
	Yes	No	Total	
Yes	a (true positive)	b (false positive)	a + b	PPV = a/(a + b)
No	c (false negative)	d (true negative)	c + d	NPV = d/(c + d)
Total	a + c	b + d	N	
	Sensitivity = a/(a + c)	Specificity = d/(b + d)		

each subject using a reference standard of choice and assigning an accurate value. For example, medical record review may be conducted for all subjects in a study using EHR data to determine if they met the conceptual definition for having a clinical condition of interest. However, this may often be infeasible due to lack of resources. In such scenarios, validation studies need to be conducted to measure the performance of an operational definition. For the binary classification case, validation studies focus on measuring performance in terms of sensitivity, specificity, positive and negative predictive value (Table 1).

As the performance of an operational definition may depend on the data source, study population, time frame and the reference standard, a validation is ideally carried out in an adequately large sample of the same study population as a part of the proposed RWE study. For example, to validate a myocardial infarction algorithm in the US FDA sentinel system, medical chart reviews and adjudication was done on a random sample of 143 individuals identified as having an event by the algorithm [28]. The positive predictive value (PPV), defined as confirmation of occurrence of the event by adjudication, was 86% in this random sample. In another example, Desai et al. [29] and Zhang et al. [22] consider the misclassification of cause-specific mortality outcome due to the information not being well captured in a medical claims database. Specifically, due to lack of cause of death information in the study data, the outcome cardiovascular death (CV death) was operationally defined as any death within 30 days of a major CV event recorded in the database. The authors assessed the bias implications of misclassification for this variable with a validation dataset, from the National Death Index data, where information on the cause of death was available and concluded that there was a possibility for substantial bias in the estimated treatment effects with their operational definition.

Validation studies can be used in combination with methods for correcting and adjusting bias due to misclassification and measurement error. Keogh et al. [30] discuss the complex nature of assessing and correcting for information bias in inference and present two methods, regression calibration and simulation extrapolation, to adjust for measurement error, when there is some availability of quantitative information regarding the measurement error. More complex methodologies to assess and correct biases for complex cases is presented in Shaw et al. [31]. Lian et al. [32] propose a Bayesian modeling strategy to correct for exposure

misclassification. They apply their methods to correct for misclassification in patient self-reported smoking status in a retrospective real-world study of diabetic nephropathy patients. Using an external validation study to estimate the potential bias and to assess sensitivity and specificity, they provide bias adjustment in the comparative analysis. Even if not correcting or adjusting for biases, quantitative bias assessments, which are a set of sensitivity analyses to assess the impact of potential biases on a study inference [33], are recommended to demonstrate whether and how misclassification might affect study results.

More generally, for outcomes or other binary variables of interest, the trade-off between false-positive and false-negative cases when selecting an operational definition should be considered and a proper outcome validation approach to support internal validity of the study should be identified. From a regulatory perspective, the recently published Draft FDA Guidance for Industry on [4] Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision Making for Drug and Biological Products offers the following advice for sponsors submitting RWE:

> Regarding outcome validation, sponsors should justify the proposed validation approach, such as validating the outcome variable for all potential cases or non-cases, versus assessing the performance of the proposed operational definition; if the latter will be done, justify what performance measures will be assessed. The protocol should include a detailed description of the outcome validation design, methods, and processes, as well as sampling strategy (if applicable). If a previously assessed operational definition is proposed, additional information should be provided, including (1) data source and study population; (2) during what time frame validation was performed; (3) performance characteristics; (4) the reference standard against which the performance was assessed; and (5) a discussion of whether prior validation data are applicable to the proposed study.

4 Special Consideration for Key Variables

4.1 Exposure

The definition of medication exposure should include dose, formulation, strength, route, timing, frequency, and duration. The data source used must be able to identify the product of interest. This can be done through patient or physician reports, billing, or procedure codes. Correspondingly, the operational definition used must reflect the resources available such as the coding system in EHR or claims databases, with an understanding of prescription, delivery, and reimbursement characteristics of the drug. For example, in some data, the same billing or diagnostic code may be used to indicate administering of multiple vaccine, making it impossible to identify the specific vaccine formulation. Commonly, in EHR and medical claims data sources, operational definitions for ascertaining exposures are based on structured data elements which contain codes for the medication dispensed (e.g., National Drug Codes associated with prescription fills in claims data) or procedure performed (e.g., HCPCS J code for inpatient administration of injectables). It is also possible to

combine information from unstructured data, using medical chart review of notes in combination with dispensing and prescribing data to confirm patient's use of medication after dispensation. Bustamente et al. [34] ascertain aspirin exposure in a retrospective cohort study of veterans undergoing usual care colonoscopy using such a method.

As some medications such as vaccines are designed as one-time exposures and other medications are intended to be used over time, the ability of the data source to capture the relevant duration of the exposure should be considered. In terms of ascertainment, the operational definition should address how medication use will be measured, how potential gaps in therapy and how refill stockpiling will be addressed, especially in data sources ascertaining exposure through prescription fills or dispensations.

It is important to note that RWD sources often capture only the prescription fills or dispensations of drugs, but not the actual exposure to drug, as the latter depends on patients obtaining and using the prescribed medication. As such, exposures in these settings are ascertained through a proxy. Thus, validation, where the exposure classification is compared to a reference standard to produce estimates of misclassification that can be used in sensitivity analysis or adjusted for is important. While validating, attention must be paid to all characteristics of exposure, including duration, dose, and switching. Validation can be done by performing additional studies in the same population such as by undertaking a survey of study participants to assess drug intake. In some cases, prior studies such as published reports of number of people taking vaccines may be relied on to estimate misclassification rates. Apart from misclassification, other sources of bias stemming from lack of information such as the unavailability of information on nonprescription drug usage must also be considered.

4.2 Outcome

As noted earlier, the conceptual definition of an outcome should reflect the current medical and scientific understanding and may vary by study. A description of the conceptual definition in the study protocol should include the signs, symptoms, and laboratory and radiology results needed to confirm the presence of the condition. The conceptual definition for anaphylaxis, for example, may include sudden onset, rapid progression of signs and symptoms, $\geq$1 major dermatological criterion, and $\geq$1 major cardiovascular or respiratory criterion. The conceptual definition may be operationalized using diagnosis (e.g., ICD-9-CM, ICD-10) or procedure codes (HCPCS), laboratory tests (e.g., identified using LOINC codes) and values or unstructured data (e.g., physician notes, radiology and pathology reports). The operational definition description should include the coding system, if any, used in the data source, the rationale and limitations of the definition and the impact on misclassification.

The general considerations on validation presented in Sect. 3 apply while validating the outcome. Perhaps, most importantly, the trade-off between false

positive and false negative cases must be considered and inform the approach for validation. For rare disease outcomes, it might be prudent to select an operational definition with high sensitivity (and consequently, low PPV) and perform complete verification (e.g., through chart review) of cases to maximize the possibility that all true cases are captured, and false positives are minimized. In other situations with a more common outcome (e.g., disease-specific hospitalization), misclassification through false-positive and false-negative may both happen at a considerable rate. In these situations, measuring PPV alone would not be enough to inform bias due to misclassification.

4.3 Confounders

Depending on the type of variable, the specific principles and considerations described in the above two subsections may apply to a key confounder of interest. For example, covariates that are medical events such as comorbidities or procedure utilizations are similar in nature to outcomes, whereas covariates such as concurrent or past medication uses are similar to an exposure variable in terms of ascertainment and validation. Sometimes covariates such as family history, lifestyle factors may need to be ascertained or validated through data linkage to provider or patient surveys.

5 A Case Study from Myrbetriq® Postmarketing Requirement

This section introduces an example of rule-based ascertainments of outcome and exposure in a post-marketing safety study using real-world data. On June 2012, the FDA approved Myrbetriq® (mirabegron) to treat overactive bladder (OAB) with symptoms of urge urinary incontinence, urgency, and urinary frequency. During the premarket clinical development, a number of cardiovascular (CV) and malignant events were observed in the mirabegron arm compared to the placebo arm. Thus, the FDA required the Applicant to conduct two postmarketing safety studies to evaluate the incidence of the adverse outcomes of interest among OAB medication users [35]. One postmarketing study (PMR 1898-3) primarily focused on the incidence of CV outcomes during current exposure in patients administered mirabegron. This study adapted real-world data identified from five data sources in US and European electronic healthcare databases with appropriate linkage: Danish National Patient Registry (NPR), Swedish NPR, Clinical Practice Research Datalink (CPRD; UK), Optum Research Database (ORD; USA), and Humana Database (USA). The CPRD database included CPRD-linked and CPRD-unlinked. The study design and analysis results are published elsewhere [36]. In this example, we introduce how the research partners for each real-world database ascertained and identified the study outcome and exposure.

As the study obtained outcomes and exposures from various data sources, methods applied for identification and ascertainment were different for each database. However, the study clarified that research partners followed the common protocol and statistical analysis plan along with site-specific protocols [36]. For outcomes, the study ascertained the cases through "direct linkage to registries, medical record review, or physician questionnaires" [36]. Particularly, ORD and Humana ascertained mortality outcomes through linkage to the national death index (NDI) [36, 37]. For exposure, the study classified a person-day as current exposure to the medication if it falls under the days of supply of the prescription or dispensing with an additional grace period of 50% [36]. This grace period accommodated patient's varying adherence to medications beyond the days of supply to adjust missed scheduled dose or changes in dosing schedule [36]. The total days of supply were estimated by different methods based on the available information in each database. Either switching to a newly prescribed medication group from other treatment group or reaching the end of days' supply (after applying the grace period) terminated the current exposure status of a given person–time for treatment group [36]. To avoid overlap in days of supply between the prescriptions/dispensing, the authors truncated the first prescription/dispensing on the day before the subsequent prescription [36]. Brief introductions of each database and their ascertainments of outcome and exposure are described as follows:

The Danish National Patient Registry (NPR) is a population-based administrative registry that contains clinical and administrative data from all Danish hospitals since 1977 [38]. In this register, diagnosis codes are entered upon discharge according to the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) and adapted for use in the Danish healthcare system [36, 38]. The authors stated that the outcomes were identified through a direct linkage to patient registers using ICD-10 diagnosis codes [36]. The Danish NPR did not have direct information on the days of supply, thus an alternative approach was used to estimate the exposure, days of supply. A waiting-time distribution was used to identify maximum interval between two prescriptions deemed as belonging to the same treatment episode [39, 40]. If any interval was larger than the identified maximum length, it was considered as a gap in treatment [36]. The accuracy of AMI and stroke diagnoses was validated through several studies and demonstrated fairly high PPVs of AMI (81.9–100%) [41–43] and stroke (79.3–97%) [44–47], respectively.

Similarly, in Sweden, the National Patient Register (NPR) was used to identify the CV outcomes using ICD-10 diagnosis codes. The Swedish National Inpatient Register (NIPR) is a part of the NPR launched in 1964 and has completed national coverage since 1987 [48, 49]. It is known that more than 99% of all somatic (including surgery) and psychiatric hospital discharges are registered in the current NIPR. The Swedish NPRs also did not have a direct linkage to the days of supply information. Thus, the days of supply were estimated by dividing the number of prescribed or dispensed tablets from the number of daily recommended tablets [36]. The PPV of AMI ranged between 86–98% [50, 51] and the PPV of stroke demonstrated 94% [52].

The CPRD collects deidentified patient data from a network of general practitioners (GPs)' practices across the UK [53–55]. Its primary care data are linked to a range of other health-related data to provide a longitudinal, representative UK population health dataset. The CPRD-linked includes the data from general practices that permitted hospital and mortality data linkage and the CV outcomes were identified through a direct linkage to patient registers using ICD-10 codes [55]. As the CPRD-unlinked does not have this linkage, the potential CV outcomes were identified using the Read codes [56] and adjudicated by the GPs who provided care for patients [36]. As the CPRD rarely record days of supply, the total days of supply was estimated by using a combination of available information such as recorded number of days of supply, quantity of tablets prescribed, daily dose, and tablet strength [36]. For explicit records on days of supply, the researchers assessed the plausibility of the prescription record values against the corresponding values for the quantity of prescribed tablets and the daily dose for that prescription. If the value of recorded days of supply did not match with value of the quantity of tablets prescribed divided by the daily dose, the calculated value was used instead [36]. Physician questionnaire was used to validate the AMI and stroke outcomes [36, 57].

The ORD and Humana database used claims data to identify potential acute myocardial infarction and stroke outcomes. These outcomes were based on ICD-9 or ICD-10 Clinical Modification (ICD-9-CM or ICD-10-CM) diagnosis codes in the principal diagnosis position on at least one facility inpatient claim for hospitalization [2]. For validation purpose, medical record reviews were used to adjudicate these outcomes for both databases. To identify all-cause and CV mortality, both database used external linkage to the National Death Index, which is a central computerized index of death record information on file in the state vital statistics offices [37]. The primary and underlying causes of death were recorded using ICD-10-CM diagnosis codes. Unlike the European database described above, the ORD and Humana database were able to directly capture the days of supply associated with outpatient dispensing. The researchers calculated the total days of supply by summing days of supply for all consecutive prescriptions or dispensing of a given medication [36].

The post-marketing study was performed using real-world data collected from five different sources. As each source had different operational structure for collecting data, no universal method was applicable to ascertain and validate the outcome and exposure variables. Thus, each data source used its own methods for ascertainment and validation.

6 Discussions and Concluding Remarks

Ascertainment of key variables and their validation are the most important steps in designing a study in RWD once it has been identified that the database is fit-for-purpose to answer the research question at hand. In this chapter, we covered the challenges of ascertaining outcomes, specifically, identifying an operational definition or a computable phenotype in RWD sources, and went through the many

methods which can be employed. Particularly, with the help of rapid technological advances, we discussed how all the data available in RWD sources can be leveraged to identify complex clinical characteristics. While such approaches are encouraging for the future of RWE, it must be stressed, as done here, that ensuring the internal and external validity of any approach is of paramount importance. Validation does not just imply calculating relevant metrics such as PPV and sensitivity but involves the consideration and assessment of the bias that can be caused by the imperfection of the operational definition.

We also walked through an example where RWE was used to satisfy post-marketing safety requirements for an approved drug. The studies conducted by the applicant used rule-based methods and data linkage to identify outcomes and exposures in five different RWD sources to answer the same safety question. Through the example, we highlight how a study can be designed to thoroughly address the question of ascertainment and validation.

Disclaimer This chapter reflects the views of the authors and should not be construed to represent FDA's views or policies.

References

1. FDA. Clinical Trial Imaging Endpoint Process Standards. https://www.fda.gov/files/drugs/published/Clinical-Trial-Imaging-Endpoint-Process-Standards-Guidance-for-Industry.pdf 2018.
2. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. J Am Med Inform Assoc. 2013;20(e2):e226–31. https://doi.org/10.1136/amiajnl-2013-001926.
3. Richesson RL, Smerek MM, Blake Cameron C. A Framework to Support the Sharing and Reuse of Computable Phenotype Definitions Across Health Care Delivery and Clinical Research Applications. EGEMS (Wash DC). 2016;4(3):1232. https://doi.org/10.13063/2327-9214.1232.
4. FDA. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products (Guidance for Industry, Draft Guidance). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory 2021.
5. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annu Rev Biomed Data Sci. 2018;1:53–68. https://doi.org/10.1146/annurev-biodatasci-080917-013315.
6. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc. 2012;19(2):212–8. https://doi.org/10.1136/amiajnl-2011-000439.
7. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med. 2013;15(10):761–71. https://doi.org/10.1038/gim.2013.72.
8. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am

Med Inform Assoc. 2016;23(6):1046–52. https://doi.org/10.1093/jamia/ocv202.
9. Kern EF, Maney M, Miller DR, Tseng CL, Tiwari A, Rajan M, et al. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. Health Serv Res. 2006;41(2):564–80. https://doi.org/10.1111/j.1475-6773.2005.00482.x.
10. Wei WQ, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc. 2012;19(2):219–24. https://doi.org/10.1136/amiajnl-2011-000597.
11. Gibson TB, Nguyen MD, Burrell T, Yoon F, Wong J, Dharmarajan S, et al. Electronic phenotyping of health outcomes of interest using a linked claims-electronic health record database: Findings from a machine learning pilot project. J Am Med Inform Assoc. 2021;28(7):1507–17. https://doi.org/10.1093/jamia/ocab036.
12. Carrell DS, Gruber S, Floyd JS, Bann M, Cushing-Haugen K, Johnson R, et al. Improving methods of identifying anaphylaxis for medical product safety surveillance using natural language processing and machine learning. PHARMACOEPIDEMIOLOGY AND DRUG SAFETY: WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA; 2021. p. 16–7.
13. Henderson J, He H, Malin BA, Denny JC, Kho AN, Ghosh J, et al. Phenotyping through Semi-Supervised Tensor Factorization (PSST). AMIA Annu Symp Proc. 2018;2018:564–73.
14. Murray SG, Avati A, Schmajuk G, Yazdany J. Automated and flexible identification of complex disease: building a model for systemic lupus erythematosus using noisy labeling. J Am Med Inform Assoc. 2019;26(1):61–5. https://doi.org/10.1093/jamia/ocy154.
15. Yu S, Ma Y, Gronsbell J, Cai T, Ananthakrishnan AN, Gainer VS, et al. Enabling phenotypic big data with PheNorm. J Am Med Inform Assoc. 2018;25(1):54–60. https://doi.org/10.1093/jamia/ocx111.
16. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc. 2017;2017:48–57.
17. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507–13. https://doi.org/10.1136/jamia.2009.001560.
18. Chen L, Gu Y, Ji X, Sun Z, Li H, Gao Y, et al. Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning. J Am Med Inform Assoc. 2020;27(1):56–64. https://doi.org/10.1093/jamia/ocz141.
19. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS One. 2018;13(2):e0192360. https://doi.org/10.1371/journal.pone.0192360.
20. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. J Biomed Inform. 2019;100S:100057. https://doi.org/10.1016/j.yjbinx.2019.100057.
21. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc. 2020;27(3):457–70. https://doi.org/10.1093/jamia/ocz200.
22. Zhang Q, Gossai A, Monroe S, Nussbaum NC, Parrinello CM. Validation analysis of a composite real-world mortality endpoint for patients with cancer in the United States. Health Serv Res. 2021;56(6):1281–7. https://doi.org/10.1111/1475-6773.13669.
23. Griffith SD, Tucker M, Bowser B, Calkins G, Chang CJ, Guardino E, et al. Generating Real-World Tumor Burden Endpoints from Electronic Health Record Data: Comparison of RECIST, Radiology-Anchored, and Clinician-Anchored Approaches for Abstracting Real-World Progression in Non-Small Cell Lung Cancer. Adv Ther. 2019;36(8):2122–36. https://doi.org/10.1007/s12325-019-00970-1.
24. Wedam S, Fashoyin-Aje L, Bloomquist E, Tang S, Sridhara R, Goldberg KB, et al. FDA Approval Summary: Palbociclib for Male Patients with Metastatic Breast Cancer. Clin Cancer Res. 2020;26(6):1208–12. https://doi.org/10.1158/1078-0432.CCR-19-2580.

25. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc. 2013;20(e1):e147–54. doi:https://doi.org/10.1136/amiajnl-2012-000896.
26. Pacheco JA, Rasmussen LV, Kiefer RC, Campion TR, Speltz P, Carroll RJ, et al. A case study evaluating the portability of an executable computable phenotype algorithm across multiple institutions and electronic health record environments. J Am Med Inform Assoc. 2018;25(11):1540–6. https://doi.org/10.1093/jamia/ocy101.
27. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Intern Med. 2018;178(11):1544–7. doi:https://doi.org/10.1001/jamainternmed.2018.3763.
28. Cutrona SL, Toh S, Iyer A, Foy S, Daniel GW, Nair VP, et al. Validation of acute myocardial infarction in the Food and Drug Administration's Mini-Sentinel program. Pharmacoepidemiol Drug Saf. 2013;22(1):40–54. https://doi.org/10.1002/pds.3310.
29. Desai RJ, Levin R, Lin KJ, Patorno E. Bias Implications of Outcome Misclassification in Observational Studies Evaluating Association Between Treatments and All-Cause or Cardiovascular Mortality Using Administrative Claims. J Am Heart Assoc. 2020;9(17):e016906. https://doi.org/10.1161/JAHA.120.016906.
30. Keogh RH, Shaw PA, Gustafson P, Carroll RJ, Deffner V, Dodd KW, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1-Basic theory and simple methods of adjustment. Stat Med. 2020;39(16):2197–231. https://doi.org/10.1002/sim.8532.
31. Shaw PA, Gustafson P, Carroll RJ, Deffner V, Dodd KW, Keogh RH, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2-More complex methods of adjustment and advanced topics. Stat Med. 2020;39(16):2232–63. https://doi.org/10.1002/sim.8531.
32. Lian Q, Hodges JS, MacLehose R, Chu H. A Bayesian approach for correcting exposure misclassification in meta-analysis. Stat Med. 2019;38(1):115–30. https://doi.org/10.1002/sim.7969.
33. Lash TL, Fox MP, Cooney D, Lu Y, Forshee RA. Quantitative Bias Analysis in Regulatory Settings. Am J Public Health. 2016;106(7):1227–30. https://doi.org/10.2105/AJPH.2016.303199.
34. Bustamante, R, A Earles, JD Murphy, AK Bryant, OV Patterson, AJ Gawron, T Kaltenbach, MA Whooley, DA Fisher, SD Saini, S Gupta, and L Liu, 2019, Ascertainment of Aspirin Exposure Using Structured and Unstructured Large-scale Electronic Health Record Data, Med Care, 57:e60–e64.
35. FDA. FDA Approval Letter of Myrbetriq (mirabegron). https://www.accessdata.fda.gov/drugsatfda_docs/appletter/2012/202611Orig1s000ltr.pdf 2012.
36. Hoffman V, Hallas J, Linder M, Margulis AV, Suehs BT, Arana A, et al. Cardiovascular Risk in Users of Mirabegron Compared with Users of Antimuscarinic Treatments for Overactive Bladder: Findings from a Non-Interventional, Multinational, Cohort Study. Drug Saf. 2021;44(8):899–915. https://doi.org/10.1007/s40264-021-01095-7.
37. National Centers for Health Statistics, https://www.cdc.gov/nchs/ndi/index.htm. 2017. Accessed 20 July, 2022.
38. Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sorensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol. 2015;7:449–90. https://doi.org/10.2147/CLEP.S91125.
39. Hallas J, Gaist D, Bjerrum L. The waiting time distribution as a graphical approach to epidemiologic measures of drug utilization. Epidemiology. 1997;8(6):666–70. https://doi.org/10.1097/00001648-199710000-00009.
40. Pottegård A, Hallas J. Assigning exposure duration to single prescriptions by use of the waiting time distribution. Pharmacoepidemiology and drug safety. 2013;22(8):803–9.
41. Joensen AM, Jensen MK, Overvad K, Dethlefsen C, Schmidt E, Rasmussen L, et al. Predictive values of acute coronary syndrome discharge diagnoses differed in the Danish National Patient Registry. J Clin Epidemiol. 2009;62(2):188–94. doi:https://doi.org/10.1016/j.jclinepi.2008.03.005.

42. Madsen M, Davidsen M, Rasmussen S, Abildstrom SZ, Osler M. The validity of the diagnosis of acute myocardial infarction in routine statistics: a comparison of mortality and hospital discharge data with the Danish MONICA registry. J Clin Epidemiol. 2003;56(2):124–30. https://doi.org/10.1016/s0895-4356(02)00591-7.
43. Coloma PM, Valkhoff VE, Mazzaglia G, Nielsson MS, Pedersen L, Molokhia M, et al. Identification of acute myocardial infarction from electronic healthcare records using different disease coding systems: a validation study in three European countries. BMJ Open. 2013;3(6). https://doi.org/10.1136/bmjopen-2013-002862.
44. Wildenschild C, Mehnert F, Thomsen RW, Iversen HK, Vestergaard K, Ingeman A, et al. Registration of acute stroke: validity in the Danish Stroke Registry and the Danish National Registry of Patients. Clin Epidemiol. 2014;6:27–36. https://doi.org/10.2147/CLEP.S50449.
45. Johnsen SP, Overvad K, Sorensen HT, Tjonneland A, Husted SE. Predictive value of stroke and transient ischemic attack discharge diagnoses in The Danish National Registry of Patients. J Clin Epidemiol. 2002;55(6):602–7. https://doi.org/10.1016/s0895-4356(02)00391-8.
46. Frost L, Andersen LV, Vestergaard P, Husted S, Mortensen LS. Trend in mortality after stroke with atrial fibrillation. Am J Med. 2007;120(1):47–53. https://doi.org/10.1016/j.amjmed.2005.12.027.
47. Krarup LH, Boysen G, Janjua H, Prescott E, Truelsen T. Validity of stroke diagnoses in a National Register of Patients. Neuroepidemiology. 2007;28(3):150–4. https://doi.org/10.1159/000102143.
48. Swedish National Patient Register. Accessed July 20 2022.
49. Ludvigsson JF, Andersson E, Ekbom A, Feychting M, Kim J-L, Reuterwall C, et al. External review and validation of the Swedish national inpatient register. BMC public health. 2011;11(1):1–16.
50. Hammar N, Alfredsson L, Rosen M, Spetz CL, Kahan T, Ysberg AS. A national record linkage to study acute myocardial infarction incidence and case fatality in Sweden. Int J Epidemiol. 2001;30 Suppl 1:S30–4. https://doi.org/10.1093/ije/30.suppl_1.s30.
51. Linnersjo A, Hammar N, Gustavsson A, Reuterwall C. Recent time trends in acute myocardial infarction in Stockholm, Sweden. Int J Cardiol. 2000;76(1):17–21. https://doi.org/10.1016/s0167-5273(00)00366-1.
52. Lindblad U, Rastam L, Ranstam J, Peterson M. Validity of register data on acute myocardial infarction and acute stroke: the Skaraborg Hypertension Project. Scand J Soc Med. 1993;21(1):3–9. https://doi.org/10.1177/140349489302100102.
53. Clinical Practice Research Datalink. Accessed 20 July 2022.
54. Ghosh RE, Crellin E, Beatty S, Donegan K, Myles P, Williams R. How Clinical Practice Research Datalink data are used to support pharmacovigilance. Therapeutic advances in drug safety. 2019;10: https://doi.org/10.1177/2042098619854010.
55. Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. European journal of epidemiology. 2019;34(1):91–9.
56. Digital N: Read Codes. https://digital.nhs.uk/services/terminology-and-classifications/read-codes (2020). Accessed 15 July 2022.
57. Arana A, Margulis AV, Varas-Lorenzo C, Bui CL, Gilsenan A, McQuay LJ, et al. Validation of cardiovascular outcomes and risk factors in the Clinical Practice Research Datalink in the United Kingdom. Pharmacoepidemiol Drug Saf. 2021;30(2):237–47. doi:https://doi.org/10.1002/pds.5150.

Data Standards and Platform Interoperability

Nigel Hughes and Dipak Kalra

1 Why We Need to Scale Up the Generation and Use of Real-World Evidence

In recent decades the proportion of health and care information that is captured within electronic health record systems is steadily growing [1], giving rise to a rich but fragmented resource of "real-world data" (routinely collected health, care, and wellness information) from which all stakeholders can discover vital insights [2].

Health systems urgently need to improve their capability to learn from the data they hold, in order to optimize care pathways, to achieve the best possible outcomes for patients, make the best use of resources and improve patient safety. This need for evidence-based improvement includes the increasing societal expectation of equity of care standards across and between health systems, and between different population groups (for example equity on the basis of ethnicity, as recently highlighted by Brown et al. [3]). From a regulatory science perspective, the need for timely, assured qualifications and approval of ever more complex therapeutic interventions, especially utilizing both clinical trial data and real-world evidence is of paramount importance.

The need for public health systems to learn from data has never been more acutely highlighted than in the COVID-19 pandemic when there was an urgent need for disease and treatment understanding regarding this new infective threat [4]. The academic and industry research sectors also need to leverage large-scale data in order to understand the fine differences between disease sub-populations,

N. Hughes (✉)
Epidemiology, Global R&D, Janssen, Beerse, Belgium
e-mail: nhughes@its.jnj.com

D. Kalra
Institute for Innovation through Health Data (i~HD), Ghent, Belgium

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_6

Table 1 Examples of health system, population health, academic and industry research areas needing to make use of big health data

Health systems and population health purposes	Academic and industry research purposes
Healthcare provider performance and planning	Epidemiology
Health services and resource planning	Disease understanding and stratification
Quality and safety, care pathway optimization	Digital innovation: devices, sensors, apps
Population health needs assessment	AI development
Personalized medicine services	Personalized medicine and bio-marker research
Pharmacovigilance	Diagnostics development
Public health surveillance	Drug development
Prevention and wellness programs	Clinical trial planning and optimization
Public health strategy	Comparative effectiveness research

for precision medicine [5], and to develop a wide range of personalized therapies [6], diagnostics, monitoring, and medical devices [7]. Artificial intelligence learning needs very large data sets in order to deliver precise, accurate, and safe recommendations [8], as well as for training algorithms. Large data sets are also invaluable for the training of clinical and research personnel. Table 1 lists some examples of knowledge discovery purposes for which health data is needed, to improve care and to accelerate research.

The case for combining health data from heterogeneous sources in order to maximize this learning opportunity has never been more compelling. The opportunities are now vast, with electronic health records (EHRs) becoming more sophisticated in hospitals, specialty care, and primary care, and with a greater proportion of that data needing to be structured and coded. There is increasing adoption by patients of home monitoring devices for long-term condition management and other apps that support them with wellness and prevention [9]. Countries continue to invest in an increasing number of disease and procedure registries that provide great value for research, especially in rare diseases [10]. Over the past decade, healthcare funders and ministries have substantially invested in national scale eHealth infrastructures and clinical research infrastructures, for example, in Germany and France [11]. There are also important multicountry data infrastructures already operational such as the European Union Innovative Medicines Initiative (IMI) European Health Data and Evidence Network (EHDEN) [12], forthcoming such as the European Medicines Agency (EMA) DARWIN EU® initiative [13] and the European Commission's proposals for a European Health Data Space (EHDS) [14]. Globally, the Observational Health Data Sciences and Informatics (OHDSI) open science collaborative network has been established to support rapid network studies internationally.

There is already a wealth of valuable research generated through big health data ecosystems, demonstrating the utility and societal value of leverage of this knowledge [15]. The European Institute for Innovation through Health Data is starting to publish summary case studies of health data use, especially for research,

in order to help communicate these beneficial uses of data to the public and other stakeholders [16].

Two recent examples of research findings that illustrate the value of large-scale data access have been published by partners of the OHDSI and EHDEN networks. A paper published in the Lancet in 2019 by Suchard et al. reported on small but statistically significant advantages of thiazide diuretics over angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, dihydropyridine or non-dihydropyridine calcium channel blockers in the reduction of risk from complications of hypertension, in particular myocardial infarction, stroke, and heart failure [17]. These very small effects were detected by examining the records of 4.9 million patients treated for essential hypertension across four countries, studying historical data going back several years. The authors estimated that this research might have required 22,000 conventionally sized randomized clinical trials and would have taken many years to generate results as opposed to the months that they took. In 2022, Li et al. reported on a large data study of patients vaccinated and unvaccinated for COVID-19, examining the incidence of rare neurological complications [18]. This study involved over eight million people who had received at least one inoculation with a COVID-19 vaccination, around three quarters of a million unvaccinated individuals with COVID-19 infection and over 14 million general population controls. The study found no increase in incidence of the purported rare neurological conditions in vaccinated individuals, but did find a small increase in those complications in individuals who had contracted COVID-19.

In both of these OHDSI and EHDEN supported studies, the large volumes of health records utilized were not extracted from multiple data sources and consolidated within a single data repository. Instead, they adopted a well-recognized federated architecture, in which research queries are cascaded from a central research point to multiple data sources across countries, to be executed locally on each data source as a distributed query (often termed, 'data visiting'). Through this architecture, only the query results, almost always a numeric frequency distribution or cross tabulation, is returned to the central point, and further synthesized via meta-analysis for overall conclusion. This distributed query methodology avoids the need to transport patient-level data between sites and between countries, which greatly reduces the risks from a data protection and information security perspective. The nature of this architecture, and its interoperability requirements, are discussed later in this chapter.

2 Enabling Health Information Interoperability

It is well recognized that health data is collected through very different hospital, General Practitioner (GP), patient facing, and other applications, stored in different Information and Communications Technologies (ICT) products that utilize different ways of representing health information. However, when conducting evidence generating research questions across multiple data sources, it is necessary to

harmonize these data representations first, in order to ensure that the data are always correctly interpreted. Irrespective of whether data for research is combined into large data sets and databases, or whether distributed querying (through federated architecture) is adopted, a common representation of the data is an overarching requirement. Clinical data standards are therefore essential for enabling the scaling up of learning from data, for the benefits of care, strategic decision-making, and research.

The task of representing health data is far from straightforward. This is partly because of the inherent complexity of health data, which covers many different categories of information ranging from health history and examination findings through to laboratory and radiology and genomic investigation results, sophisticated physical and psychological assessment methods, a diversity of diagnostic and treatment data types, and monitoring information. Furthermore, health is focusing more strongly now on wellness and prevention, which not only requires the collection and analysis of health-related factors but also other influences such as lifestyle and environmental considerations, which have their own data categories and representations.

The individual data items that make up these different categories of health information are themselves somewhat complex to represent, because the individual data values are held within a rich context that includes the structural organization of multidimensional clinical observations, accompanying interpretation context such as whether a finding is present or absent, certain or uncertain, etc., when and where the information was acquired, its provenance, and visualization management. This context information may radically alter the meaning of a simple-looking clinical term, as illustrated by these examples in Fig. 1, which lists many different interpretations that might apply to a clinical term for chronic obstructive pulmonary disease (COPD) in an electronic health record.

The EHR will also need to represent provenance information, which is sometimes important when clinical findings are being interpreted for the generation of real-world evidence. Interoperability standards should therefore aim to incorporate most of the information indicated in Fig. 2.

Despite this complexity, health information interoperability standards are relatively mature, capable of representing structure, content, and context faithfully and therefore to enable the meaningful exchange of information between systems for continuity of care and the accurate combining of information for knowledge discovery.

However, the various international standards development organizations that are active in the health domain, and the standards that they have developed, have grown in response to particular needs and drivers for interoperability, giving rise to standards for representing specific kinds of data (such as laboratory findings, medicines, clinical observations) which have been developed by different organizations at different times and do not necessarily align well when they are used in combination. This can lead to standards adoption uncertainty and complexity when eHealth or research infrastructures are being developed, which will need to cover a wide range of health data types and to represent these using standards.

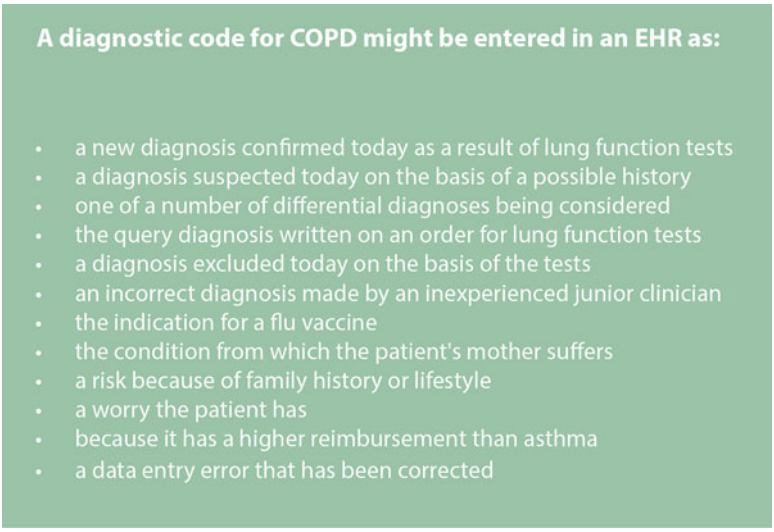

Fig. 1 Different possible interpretations of a diagnostic code for chronic obstructive pulmonary disease that could be conveyed through context information within an EHR

Fig. 2 Provenance, traceability and security context information usually represented within an EHR system

It is important for those responsible for making standard adoption decisions to be aware of the different kinds of standard, and standards development organization, within the health data ecosystem in order to make wise adoption decisions. The next section summarizes some of the major organizations that develop health data standards and highlight some of the main standards.

It is first helpful to distinguish

(a) Standards that have been developed and are largely used for the point-to-point communication of patient-level data, for example, to support continuity of care for individual patients.
(b) Standards that are used for patient-level data, but exclusively in a clinical research (clinical trial) context and not used in routine healthcare.
(c) Standards that specify the representation of data for analysis purposes, which still represent patient-level real-world data, but are optimized for population level use of the data in generating real-world evidence.

For each of these interoperability use cases, the standards themselves may focus on representing the data from one or more of these well recognized perspectives.

- Technical or structural (syntactic) interoperability, which focuses on the organizational structure of a health record or a clinical data set, the relationships between parts of complex data structures and the detailed organizational structure of health data types such as measured quantities.
- Semantic interoperability which represents the meaning of data items and their observed values, which itself comprises some different layers
  - Terminology systems which represent part or all of the clinical meaning landscape with particular emphasis on textual data
  - Measurement units and other term lists that specify the interpretation of quantities and complex multimedia data types
  - Detailed clinical models that specified the aggregation of data items to represent the complete documentation pattern for an EHR entry, such as a prescribed drug that combined several individual data items such as the drug name, dose, frequency, etc.

The above list focuses on the representation of the health data. There are many other standards that specify how information should be stored, or telecommunicated, and others that specify how the information should be protected from an information security perspective.

The complete implementation of an RWE generation ecosystem will need to utilize standards from all of these areas. It is beyond the scope of this chapter to go into detail on all of them. The section below summarizes the standards development organizations, and example standards that they publish and support, that are most widely used for the representation of various kinds of health data. Although this chapter focuses on Real-World Evidence generation, for which the standards that represent data for analysis would be the most relevant, it is important to recognize that the standards used for healthcare interoperability are also relevant because they

are likely to be supported (e.g., as export formats) by the systems such as EHR systems from which the health data will originate (as Real-World Data).

3 The Main Standards Used to Support Continuity of Health Care

The following standards are primarily used for the representation and communication of routinely collected clinical information, often within and between electronic health record systems.

3.1 Health Level Seven (HL7)

HL7 [19] is an international community of health care subject matter experts and information scientists who work together to create accredited standards for the exchange, management, and integration of electronic health care information. The HL7 community is organized in the form of a global organization (Health Level Seven, Inc.) and country-specific affiliate organizations. HL7 is supported by more than 1600 members from over 50 countries, including 500+ corporate members representing health care providers, government stakeholders, payers, pharmaceutical companies, vendors/suppliers, and consulting firms. HL7's standards are accredited by the US ANSI organization and many HL7 standards have also been adopted as ISO standards.

Its early standards were for the representation of messages to communicate information about a patient's admission to a hospital, discharge or transfer between care providers, laboratory information, treatment information, and some specialized health information exchanges. In the mid-1990s, HL7 initiated a family set of standards based on a common Reference Information Model (HL7 RIM). A wide range of message models were developed during the 1990s and have had varied success in the marketplace. One particular model that has been taken up by many health systems worldwide is the Clinical Document Architecture (CDA). In more recent years, HL7 has developed and is now rapidly promoting the use of smaller building block models known as Fast Healthcare Interoperability Resources (FHIR), which are proving more popular with industry and with national health programs because of their flexibility and lower cost of adoption.

Most data elements exchanged by HL7 standards are encoded in a terminology created and supported by other standards organizations such as SNOMED, LOINC, or WHO. HL7 also actively collaborates with other accredited healthcare international and country-specific standards groups that address information domains outside of HL7's.

3.2 The International Organization for Standardization (ISO)

Technical Committee 215 of the International Standards Organization (ISO) on Health Informatics was formed in 1998 following a decade of increasingly international cooperation among health informatics standards organizations [20]. The parent ISO organization is based in Geneva, and has the status of a non-governmental organization, recognized by law in many countries. ISO accepted the United States' offer to hold the Secretariat for TC 215; the Secretariat is managed by HIMSS (Healthcare Information and Management Systems Society) on behalf of ANSI (American National Standards Institute) who is the US member to the ISO community.

The scope of TC 215 includes architecture, frameworks, and models; systems and device interoperability; semantic content; security, safety, and privacy; pharmacy and medicines business; traditional medicine; personalized digital health, artificial intelligence. Example standards that have a high profile within the technical committee include ISO 13606 for Electronic Health Record Communication [21], ISO 13940 System of Concepts for Continuity of Care [22], and ISO 12967 Health Informatics Service Architecture [23].

While de novo standards are created within TC 215 working groups, increasingly the Technical Committee is recognizing, harmonizing, or adopting standards efforts among related standards development organizations. Internationally recognized agreements exist for the European CEN TC251 and HL7 to "fast track" standards balloted in those organizations. A newly established Joint Initiatives Council includes these fast-track organizations in addition to CDISC and SNOMED, to further strengthen international collaboration and synergy among international health information standards organizations.

3.3 SNOMED CT

SNOMED [24] (Systematized Nomenclature of Medicine) International is a not-for-profit organization that owns and maintains SNOMED CT, stated to be the world's most comprehensive clinical terminology. As stated by SNOMED International, "SNOMED International plays an essential role in improving the health of humankind by determining standards for a codified language that represents groups of clinical terms. SNOMED CT enables healthcare information to be exchanged globally for the benefit of patients/citizens, care providers and other stakeholders." [25]

SNOMED was initiated by the College of American Pathologies (CAP) in 1973 and revised into the 1990s, but in 1999, CAP's SNOMED Reference Terminology (SNOMED RT) was merged and expanded with the United Kingdom's National Health Service Read Codes (a coding system predominately used in primary care electronic health record systems) to produce SNOMED Clinical Terminology

(SNOMED CT). The structure of SNOMED CT differs from prior versions, based on sub-type hierarchy, supported by defining relationships based on description logic, versus a hierarchical classification system used before. A main use case of SNOMED CT is as the core terminology for electronic health records, covering clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices, and specimens. The January 2020 release of SNOMED CT includes more than 350,000 concepts and cross maps to other terminologies, such as ICD-9, ICD-10, LOINC, and supports ANSI, DICOM, HL7, and ISO standards. SNOMED CT enables information input in to an EHR system during the course of patient care, while ICD facilitates information retrieval, or output, for secondary data purposes.

SNOMED CT consists of four primary core components: (1) concept codes, which identify clinical terms via numerical codes; (2) descriptions, which are textual descriptions of concept codes; (3) relationships, between concept codes that have a related meaning; and (4) reference sets, used to group concepts or descriptions.

3.4 LOINC

Logical Observation Identifiers Names and Codes (LOINC) is a database and universal standard for identifying medical laboratory observations, initially developed by the Regenstrief Institute, a US not-for-profit medical research organization, in 1994. It has expanded to include nursing diagnosis, nursing interventions, outcomes classification, and patient care datasets beyond the original focus on medical laboratory codes [26].

LOINC's primary use case is to assist in electronic exchange and gathering of clinical results, comprising two parts, (1) laboratory LOINC, and (2) clinical LOINC. In 1999, the HL7 Standards Development Organization recommended LOINC as a preferred code set for laboratory test names in transactions between healthcare facilities, laboratories, laboratory testing devices, and public health authorities.

3.5 The International Classification of Diseases (ICD)

The World Health Organization maintains an international classification of diseases that has been utilized for over a century for the systematic recording, analysis, interpretation, and comparison of mortality and morbidity data collected in different countries or regions and at different times. It has served the epidemiological and public health fields, and governments, to enable insights into disease causation, prevalence, and distribution and therefore informed the design of health systems, awareness of unmet health needs, public health strategies, and prevention programmers.

The latest version of the ICD, ICD-11, was adopted by the 72nd World Health Assembly in 2019 and came into effect on January 1, 2022. It is a significant advance on prior releases by being both a classification, its original purpose, and a terminology system that can provide multilingual vocabularies for clinical and public documentation in registries, electronic health record systems, and prevention information systems [27].

The WHO also maintains the International Classification of Functioning, Disability and Health [28] (ICF) and the International Classification of Health Interventions [29] (ICHI), which are similarly used on a worldwide basis in multiple languages.

3.6 DICOM

Digital Imaging and Communications in Medicine (DICOM) is the standard for communicating and managing medical imaging information and related data. Originally developed by the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) in the early to mid-1980s. It was originally the output of the combined standards committee of both organizations, with the third substantive iteration being known as DICOM 3.0 in 1993, to differentiate it from prior versions, but also to identify it as a fully-fledged standard [30].

Large-scale deployment was initially with the US Army and Air Force, as part of the Medical Diagnostic Imaging Support (MDIS) program, and in the first military Picture Archiving and Communication System (PACS). The main focus to date has been utilization with imaging equipment vendors, and healthcare IT organizations, and utilization of other standards in addition of DICOM are necessary for clinical applications, and for research, such as IHE, HL7, FHIR, or SNOMED CT.

3.7 IHE

Integrating the Healthcare Enterprise (IHE) is a non-profit organization in the United States, established in 1988 by a consortium of radiologists and information technology experts. IHE created and facilitates interoperability improvements for health care IT systems. The IHE group collects use cases, case requirements, identifying available standards, developing technical guidelines which manufacturers can implement, focusing on a clinical information need or clinical workflow scenarios [31].

IHE is recognized by ISO as a Standards Development Organization, although it mainly develops profiles of other standards to be used, often in combination, to achieve interoperability for specific use cases. The profiles are recognized in themselves as standards. For example, IHE promotes the coordinated use of established standards, such as HL7 and DICOM, to optimize clinical care.

There are numerous standards developed for differing aspects of health data, for instance, National Drug Codes (NDC) of the FDA [32], which serves as its identifier of drugs, with a publication of the listing in the NDC Directory and updated daily. The WHO hosts a similar Anatomical Therapeutic Chemical Code (ATC) directory [33], with codes assigned to a medicine according to the organ or system it works on and how it works.

A whole class of procedure codes designed to identify surgical, medical, or diagnostic interventions with a variety of coding systems, such as SNOMED CT, as described above (3.3), but also ICD-9 and ICD-10 procedure coding [34], as referred to above too (3.5), initiated by the US Centers for Medicare and Medicaid Services, in collaboration with 3M Health Information Systems in 1995, with the now current ICD-10-PCS, updated annually since 1998.

Internationally we are seeing a coalescing of standards, with an emphasis on interoperability of standards, versus development of additional, ad hoc, new standards. As we seek interoperability of our data capture, data communication, clinical interpretation, and utilization for research, we need to utilize common standards locally, nationally, and globally. A critical issue to date has been the lack of standards adoption in extremis, with the need for wider implementation and agreement between differing healthcare system stakeholders via common standards use.

Increasingly, specific standards, most covered in this chapter, are being mandated by authorities, regulators, manufacturers, and research organizations or collaborations. The development of research networks is also reinforcing the need and use of specific standards to facilitate interoperability, syntactic and semantic, to enhance efficiencies in research and standardization of the research process, from data harmonization, methods to analytics. This is also a need in the regulatory domain, both for evidence-based decision-making and rapid research requirements, such as pharmaco-surveillance or risk management.

4 The Main Standards Used to Support Clinical Trials

4.1 CDISC

The Clinical Data Interchange Standards Consortium (CDISC) is a standards developing organization working to, “enable information system interoperability to improve medical research and related areas of healthcare.” [35]

Since initiating as a voluntary initiative in 1997, and then through a not-for-profit organization, CDISC has iterated multiple standards, foundational, for data exchange and in specific therapeutic areas. Evolving work on HL7 FHIR to CDISC has produced an *initial* joint mapping implementation guide from the former to the latter, facilitating use of real-world data with, e.g., clinical trial data, but further development is required. Unlike the healthcare standards referred to above, CDISC

standards cover both interchange and data content (storage), which can be used. The standards utilized support a model for planning (Protocol Representation Model, PRM), a model for data collection (CDASH), a Study Data Tabulation Model (SDTM) defining the structure, attributes, and content of study datasets, an Analysis Dataset Model (ADaM), and Operational Data Model (ODM) and a series of vocabulary and content (Therapeutic Area) standards.

CDISC standard and processes are required by the United States' FDA and Japan's PMDA, facilitating efficiencies in the approval times following clinical research from data capture and exchange through to analytics.

4.2 Federated Data Networks

Essentially, a Federated Data Network (FDN, sometimes referred to as distributed data network) is a managed architecture that allows for the sharing of mutual resources for RWD use, for primary or secondary care settings and clinical care decision-making as well as research use, whilst preserving the primacy of the RWD at a local level. Data is not moved from its source hosting, (though hybrid models can exist with local and central data hosting), with the research question or query moving to where the data is originally hosted, with aggregation of the results centrally or delivered to the researcher, so-called data visiting [36].

It is a sociotechnical construct, including the technical architecture and tools to facilitate the network, with governance aspects (socio), based on agreements, codes of conduct, and adherence to legal and privacy requirements (such as the EU General Data Protection Regulation—GDPR [37]) through privacy by design, facilitating the community's use of the data in the network.

The technical architecture in an FDN allows for source data to remain secure behind its sociotechnical firewalls, i.e., technical security through to approvals and ethical oversight. Web-based tools and technologies mean source data can be analyzed where it remains, especially if it is organized in such a way as to facilitate this, e.g., via a common data model (CDM, see later), supported by central portals and management, inclusive of metadata-driven catalogs.

Though different in sociotechnical aspects, FDNs such as the FDA's SENTINEL [38], PCORNET [39], OHDSI [40], IMI ADVANCE (now being sustained and maintained by the vac4EU initiative [41]), IMI ConcePTION [42], IMI EHDEN (European Health Data & Evidence Network) [12], and commercial providers, and the future European Medicines Agency's DARWIN EU (Data Analysis and Real-World Interrogation Network) [13] and proposed legislative EHDS (European Health Data Space) [14] already exist or are being built in open science or commercial communities. Use of such principles as FAIR [43], (findable, accessible, interoperable, and reusable) data, provides the framework for exploiting the benefits of an FDN, enabled by the use of CDMs, metadata, standardized analytical tools, and fit for purpose methodologies, and in particular data discovery (Fig. 3).

The FDN framework may particularly suit the European Union's need across diverse and heterogenous Member States with varying degrees of digital maturity. Ultimately, a hybrid of centralized and federated approach is likely. There may

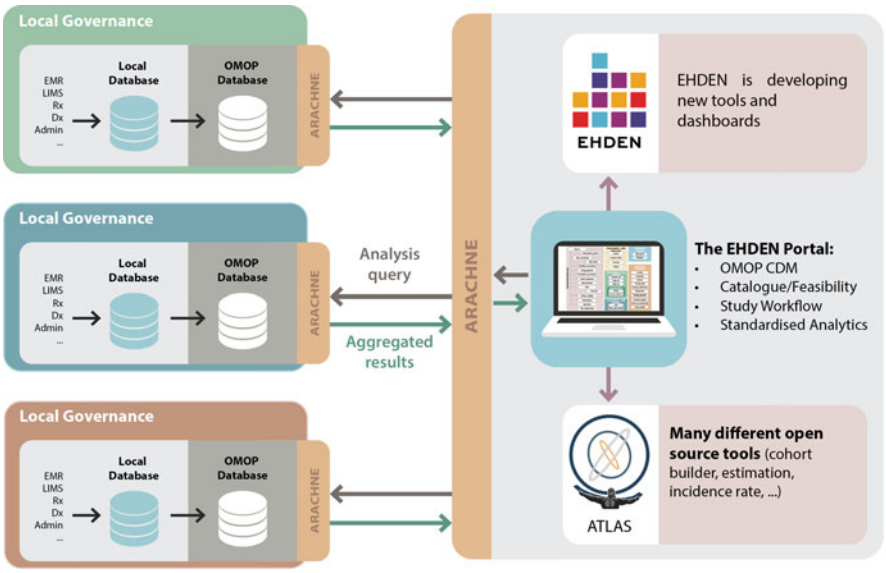

Fig. 3 Schematic of exemplar FDN. (Source: EHDEN [12])

be technical and methodological reasons for using a centralized data hosting architecture, albeit within a federated network, such as central databases or data lakes. For the needs of European healthcare systems, clinical care, and research, a mixed ecology of architectures will most probably support a diversity of needs and use cases.

While centralized data architectures have existed for some time, whether databases, data warehouses, or data lakes, this has been prohibitive in expense and resources, especially at scale, and with increasing scrutiny and legal, governance, and privacy restrictions, more complicated for the data custodian or controller and researcher with regards to data sharing and networking. Certainly, within the European landscape, increasing responsibilities cause additional overheads for central architectures. Moreover, the need for transparency in the use of real-world health data means open science sociotechnical architectures are needed, versus proprietary and/or black box approaches, especially from a regulatory perspective.

This may be related to privacy concerns, but also for instance the need by regulatory authorities to understand the analytical path from source data to evidence. Europe as a consortium of 27 Member States, and as such broad, network research require porous digital borders, as is the case for data portability to support patient mobility, necessitating federated approaches to overcome these difficulties, particularly in allowing remote, secure interrogation, but not movement of data. Moreover, being able to utilize a CDM to harmonize languages is also an advantage in network, multisite studies across borders, albeit common coding at source across the EU would be ideal, but unlikely.

A concern expressed by some is the contemporaneous nature of the data being mapped, i.e., how often is it refreshed following the original mapping to a CDM.

This is highly dependent on the source data custodian's refresh cycle, and this can vary between, e.g., on a 24-hour cycle to weeks or months, but many aspects of the mapping refresh, inclusive of for iterations of the CDM itself, can and are being increasingly automated.

Access to data is more about the terms of access, rather than direct access to RWD. The administrative burden, for instance, for approvals and contracts in conducting real-world, and especially network studies, is significant and well known to those conducting such research. Though clear governance requirements are a necessity, there needs to be mechanisms to address the administrative burden associated with them, and indeed models, such as Data Permit Authorities (DPAs), for instance, FinData [44] in Finland, or the French Health Data Hub [45] may point to a potential construct to do so.

4.3 What Is a Common Data Model, and Why Use One?

A CDM is essentially a construct, a means to an end to help organize RWD into a common structure, formats, and terminologies across diverse, heterogeneous, and multiple source datasets. It addresses a central need to be able to curate data for analysis on a contemporaneous and continuous basis, not on a per study basis, or for large-scale, geographically diverse, network studies of multiple data sources [29].

This inevitably has benefits with regards to reducing the latency and resource requirements overall to conducting research at scale and ensuring quality more rapidly, versus other methods, especially in supporting an FDN (though CDMs can be used for centralized databases too). The mapping process itself inherently incorporates data quality audit of both the source and the CDM-mapped data, with iterative stages per mapping cycle and over time.

A key concept is the need to standardize data which has been collected, stored, and curated differently, whether in an institution, or across data sources, up to an international scale. The CDISC standard, utilized especially for randomized clinical trials (RCT), is a common data model, facilitating regulatory authorities such as the FDA to receive, analyze, and opine on diverse studies across the pharmaceutical industry. The SENTINEL CDM was designed to address the need to do the same for RWD with an emphasis on regulatory pharmacovigilance in the United States, and the Observational Health Data Analytics and Informatics (OHDSI) global collaboration's Observational Medical Outcome Partnership (OMOP) CDM is facilitating a global open science network, amongst others [29].

The FDA created the Sentinel Initiative to meet a mandate by Congress in the FDA Amendments Act of 2007. Through the Sentinel Initiative, FDA aims to develop new ways to assess the safety of approved medical products, including drugs, vaccines, and medical devices [30] (Fig. 4).

The Sentinel System helps to answer the FDA's questions on approved medical products. It does this by creating algorithms that analyze electronic healthcare

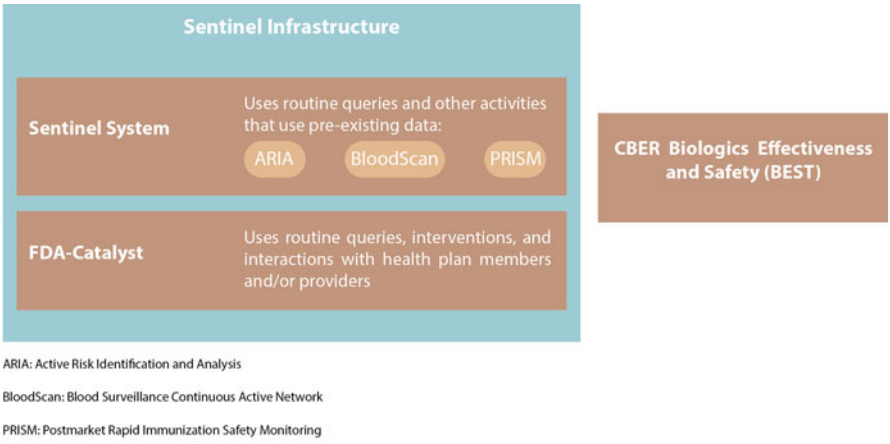

Fig. 4 The Sentinel Initiative infrastructure and governance. (Source: The Sentinel Initiative [38])

data, using statistical methods to study relationships and patterns in medical billing information and electronic health records.

Within the Sentinel System is the Active Postmarket Risk Identification and Analysis System (ARIA). The ARIA system has two main components. The first component is the healthcare data formatted in the Sentinel CDM. The second component includes analytical tools for internal analyses. Congress mandated ARIA in the United States FDA Amendments Act (FDAAA) of 2007 [46]. ARIA is the most widely used portion of the Sentinel System.

FDA-Catalyst supplements the Sentinel System. The data FDA-Catalyst provide come from interactions with patients and/or providers. FDA-Catalyst combines this data with data included in the Sentinel infrastructure.

Standardization can ensure that diverse data is broadly mapped to common schema, ontologies, and vocabularies, for instance, with OMOP, SNOMED. Furthermore, it can support the use also of standardized analytical methods and tools, on top of the CDM mapped data, following extraction, transformation, and loading (ETL), or mapping into the CDM. Exemplars of studies, such as drug utilization, safety, regulatory, and studies for HTA, lend themselves to greater consistency and commonality of methodological approach afforded by standardized analytics on top of a CDM (as for instance SENTINEL ARIA's system). The use of a CDM can underpin the operation of an FDN via facilitation of distributed data querying across multiple data sources, all mapped to the same CDM, from studies through to federated predictive analytics.

Reviews and comparisons of differing CDMs exist, but the EMA's own evaluation of CDMs from a regulatory perspective probably has guiding principles that can be utilized more broadly [49]:

Structure

- The CDM
  - Can be defined as a mechanism by which the raw data are standardized to a common structure, format, and terminology independently from any particular study to allow a combined analysis across several databases/datasets.
  - Should not be considered independently of its ecosystem, which incorporates standardized applications, tools and methods, and a governance structure.
  - The ability to access source data should be retained.
  - Should be the simplest that achieves security, validity, and data sufficiency.
  - Should be intuitive and easy to understand.
  - Should enable rapid answers to urgent questions when required, be efficient and feasible.

Operation/Governance

- The CDM
  - Governance model must respect data privacy obligations across all data partners and regions.
  - The CDM should be built with sustainability as a priority.
  - Development should maximally utilize data partners' expertise. The CDM must be agreed on and accepted by the participating data partners.
  - Must have version control.
  - Should be dynamic, extendable, and learn from experience.
  - Value package should be clear to data partners.

Quality of Evidence Generation

- The CDM
  - Must operationalize reliability and validity by building clear and consistent business rules around transformation of data across multiple databases. Where divergence is unavoidable this should be recorded.
  - Focus should be on data characterization to understand if the data is fit for purpose.
  - Should be transparent on how data is defined, how it is measured and incorporate and document its corresponding validation.
  - Should allow transparency and reproducibility of data, tools, study design to facilitate credible and robust evidence across multiple datasets.

Utility

- The CDM
  - Should provide a common set of baseline concepts which should enable flexibility when required and meets the needs of potential users.
  - All the concepts that are commonly used in safety and effectiveness studies should be mapped to the CDM to maximize regulatory utility.

– Should address recognized use cases for which an established need is present.

Currently only two CDMs cover the majority of these principal requirements at significant scale, SENTINEL's in the United States, and the OMOP CDM internationally. For Europe, there is little utilization of the SENTINEL CDM, but expanding adoption of the OMOP CDM. (The CDISC ODM referred to earlier is a data model used for the submission of clinical trials evidence to medicines regulators, but is not currently widely used as a real-world data analysis representation.)

Via the EHDEN project, the European Union via IMI and 13 pharmaceutical companies are funding € 30 million over the duration of the project to accelerate utilization of the OMOP CDM across the European region, with also more than 20 other IMI projects utilizing this CDM.

In recent years, the OHDSI OMOP CDM has become an international standard for working with RWD in RWE generation, with greater than 2 billion health records mapped to the OMOP CDM globally, equating to approximately 800 million patients, and an accelerating body of literature from international studies, all characteristic for their scale and speed, whilst preserving quality. The FDA, whilst running SENTINEL, is also funding OHDSI through the FDA Centre for Biologics Evaluation and Research (CBER) [47] for biologics and vaccines pharmacovigilance, and both the DARWIN EU® and EHDS programs potentially look to include the OMOP CDM and OHDSI research framework.

The open science approach within OHDSI was demonstrated during the COVID-19 pandemic, through a study-a-thon and continuing research protocols, through its international research studies [48]. Such approaches responded to the need for the right data to be in the right place at the right time, for the right questions, at time of public health emergency, whereas more traditional approaches, via considerable per study curation, would likely still have not reported, especially for large scale studies with multiple data sources for across the European region. Outputs from this international research collaboration were utilized via the FDA and EMA for guidance to clinicians, for instance, on the safety profile of hydroxychloroquine with or without azithromycin in treating COVID-19 early in the pandemic.

Comparisons of common data models exist, as discussed in the EMA report on CDMs in 2018 and shown in Table 2.

The OMOP CDM was designed from the ground up for research purposes initially in North America in 1997, and with an emphasis on epidemiology, utilizing, e.g., US Claims data, but has been expanded over the following years, both in terms of data types incorporated, and study types supported, as well as for geographies. The original founding partners of the OMOP were the US Food and Drug Administration (FDA), Pharmaceutical Research and Manufacturers of America (PhRMA), and the Foundation for the National Institutes of Health (FNIH). OHDSI now develops and iterates the OMOP CDM. More recently, this has included regulatory

Table 2 Comparison of three CDMs

FDA SENTINEL	PCORnet	OMOP
Focused use (pharmacovigilance)	Clinical care emphasis	Broad use cases
US based	US based	Global use
Distributed data network with data queries run locally	Predominately EHR data	Broad, comprehensive model to incorporate claims data, EHRs and surveys
Predominately US claims data, minimum, but expanding EHR data	Principle of minimum mapping	Substantial mapping of content and concepts to standardize multiple different coding systems
Strict version control	Strict version control	Strict version control
Built upon principle of minimal mapping and no derived values	Flexibility for individual data partner to add data/domains to local CDMs	Iterative development by community for data/domains additions in global CDM
Source data retained		Source data retained
Extendable	Based on SENTINEL CDM	Extendable

European Medicines Agency; A Common Data Model for Europe? – Why? Which? How?; London 2018 [49]

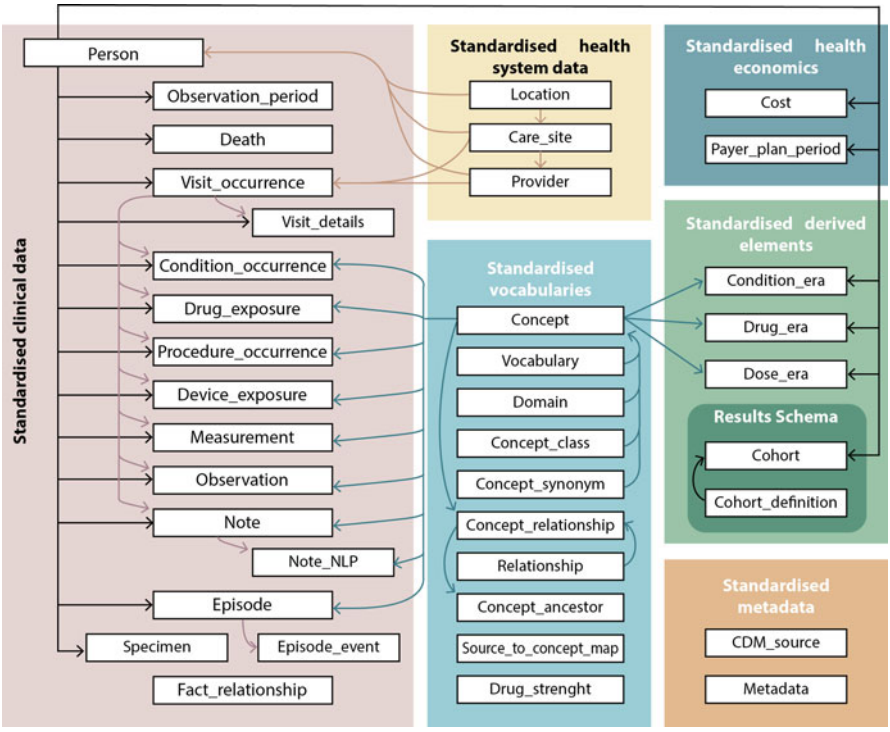

Fig. 5 OMOP common data model schema. (Source: https://ohdsi.github.io/CommonDataModel/)

use cases, and developments to enable health technology assessment (HTA) studies, or precision medicine use cases. Due to the open science emphasis of OHDSI, there is a focus on transparency, replication of results, and development of methodologies for fit-for-purpose RWE generation and observational research (Fig. 5).

The OMOP CDM and OHDSI framework do not support every conceivable use case, and likely a mixed ecology of applications, methods, and tools will be required to do so, which is a reality of working in the real world setting, but further interoperability, e.g., between HL7 FHIR (for facilitating health data exchange) and OMOP CDM (designed for RWD analysis), in particular to support outcomes research are being addressed, and hopefully accelerated (with the recent announcement of a global collaboration). The OMOP CDM utilizes the SNOMED and LOINC standards as its core, standard vocabularies.

On top of the OMOP CDM are the standardized analytical tools to support analysis of OMOP-mapped data, in particular supporting characterization, population-level estimation, and patient-level prediction studies. ATLAS is a free, publicly available, web-based tool developed by the OHDSI community that facilitates the design and execution of analyses on standardized, patient-level, observational data in the CDM format. The ATLAS tool is deployed as a web application in combination with the OHDSI WebAPI and is typically hosted on Apache Tomcat. Performing real-time analyses requires access to the patient-level data in the CDM and is therefore typically installed behind an organization's firewall. However, there is also a public ATLAS, and although this ATLAS instance only has access to a few small simulated datasets, it can still be used for many purposes including testing and training. It is even possible to fully define an effect estimation or prediction study using the public instance of ATLAS, and automatically generate the R code for executing the study. That code can then be run in any environment with an available CDM without needing to install ATLAS and the WebAPI. Other open source tools to facilitate mapping, support data quality evaluation as well as analysis have and are being developed, with more information available from the open access Book of OHDSI [50].

Skilled and knowledgeable epidemiologists with multiyear experience of the OMOP CDM, mapping datasets and analysis using the OHDSI framework is a prerequisite now for some positions. A helpful example of this is a company's ability to make quicker decisions in feasibility as to the efficacy of being able to conduct a substantive study, inclusive of with regulatory authorities, assisted by a transparent, reproducible methodology in being able to debate the company's viewpoint.

Federation and the use of the OMOP CDM is also now supporting therapeutic area-focused initiatives within the company as it proceeds to expand its collaboration with potential Data Partners.

Other projects in the European Innovative Medicines Initiative (IMI) have developed CDMs, such as ADVANCE [41], or latterly ConcePTION [42], in vaccines and pregnancy research, respectively, with the former using a CSV format CDM and Jerboa data processing software and R scripts, and the latter using a syntactic model, but neither have widespread adoption outside of their respective projects.

5 Making Data Fit for Shared Use

5.1 FAIR Principles

Findable, accessible, interoperable, and reusable are principles espoused in 2016 by Wilkinson et al. [43] As described by the authors, there is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—came together to design and jointly endorse a concise and measurable set of principles that they referred to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals.

A European program, GO-FAIR [51], a bottom-up promotion of FAIR principles, and an IMI project, FAIRplus [52], making life science data FAIR have outlined the practical implementations of the FAIR principles, which are outlined below:

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

- F1. (Meta)data are assigned a globally unique and persistent identifier.
- F2. Data are described with rich metadata (defined by R1 below).
- F3. Metadata clearly and explicitly include the identifier of the data they describe.
- F4. (Meta)data are registered or indexed in a searchable resource.

Accessible

Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorization.

- A1. (Meta)data are retrievable by their identifier using a standardized communications protocol.
- A1.1 The protocol is open, free, and universally implementable.
- A1.2 The protocol allows for an authentication and authorization procedure, where necessary.
- A2. Metadata are accessible, even when the data are no longer available.

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles.
- I3. (Meta)data include qualified references to other (meta)data.

Reusable

The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.
- R1.1. (Meta)data are released with a clear and accessible data usage license.
- R1.2. (Meta)data are associated with detailed provenance.
- R1.3. (Meta)data meet domain-relevant community standards.

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource (the infrastructure component).

5.2 Data Quality

Quality of data can be viewed as in the eye of the beholder, with data and analysis being able to answer some questions, but not all (there is perhaps not a universal truth based on any one dataset). Intrinsic to this concept are the relative indicators of quality of the source data, and the various attributes of measuring quality via a growing number of quality initiatives.

Brennen et al. (JAMIA, 2000) stated that data quality in and across diverse data sources (e.g., electronic health records, claims), "[is] the problem of ensuring the validity of the clinical record as a representation of the true state of the patient." [53].

Quality of health data, or real-world data, needs to represent quality of the source data and the curated data used for analysis, inclusive of such attributes as errors, completeness, missingness, biologic implausibility (e.g., finding male pregnancies, or BMI values inconsistent with humans).

Various initiatives and standards incorporate data quality processes, for instance, OHDSI has created data quality dashboards that can evaluate the OMOP-mapped dataset in comparison to the source dataset across a quality criterion, running a script across the OMOP CDM [54].

The European Institute for Innovation through Health Data provides a data quality assessment service, again criterion-driven, across nine dimensions (Fig. 6):

Name	Definition
Completeness	Data values are present
Consistency	Data satisfy constriants (format, allowable ranges and values, domain rules, relations)
Correctness	Values are true and unbiased with respect to their real-world state
Uniqueness	Records representing a single patient are not replicated
Timeliness	Data is up-to-date to their real world state for the task at end
Stability	Data inherent concepts and statistics are comparable among sources (hospital, professional, etc) and over time
Relevance	Data are useful for their task
Contexualization	Data are annotated with the acquisition context, their meaning and semantics
Trustworthiness	Data can be trusted based on the reputation of the stakeholders involved in their acquisition

Fig. 6 Nine data quality dimensions, suitable for health data, assessed by the European Institute for Innovation through Health Data [55]

Arguably, data quality is an emerging sub-specialism, but a critically important one in addressing confidence in being able to assess the quality of data as part of an overall assessment to engender confidence in analytical outputs and evidence generated. From a regulatory domain perspective, this will be a standard component of assessing research and studies carried out and ensuring validity in the proposed guidance from regulatory authorities using real-world data.

5.3 Research Infrastructures and Platforms

Europe is driving the momentum for big health data research through three transnational initiatives, EHDEN, DARWIN EU®, and the EHDS, which have been mentioned throughout this chapter. There are additionally disease-specific networks in vaccination and pregnancy, also mentioned earlier, and a growing number of national research infrastructures in countries such as Germany, France, Switzerland, and the United Kingdom. The three major Europe-wide initiatives are briefly summarized below.

5.3.1 EHDEN

The European Health Data and Evidence network (EHDEN, 2018–2024) is an infrastructure start-up within the Europe's IMI, an overarching public private partnership fostering innovation in healthcare and earlier access to such innovation for EU citizens. EHDEN was created to address the common bottlenecks encountered when harmonizing datasets to the OMOP common data model, at an industrial scale across the European region. Ultimately, EHDEN is building a region-wide federated data network, supporting FAIR data use, with a centralized architecture to enable a digital study workflow, data visiting/remote analysis which is privacy preserving via standardized analytical tool pipeline within the OHDSI research framework. As of time of writing, EHDEN is working with 187 Data Partners in 29 European countries across the region, and is continuing to expand [12].

Successful applicant Data Partners receive financial subgrants, technical support for mapping their data to the common data model (via EHDEN-certified small-to-medium enterprises (SMEs) in a unique marketplace of trained technical businesses), and can join the Open Science community in terms of evidence generation in multisite, network, and rapid studies; upskilling and training are also provided on tools, skills, and methods to support Data Partners, SMEs, and researchers via an EHDEN Academy (https://academy.ehden.eu). At the time of writing, EHDEN is working with 64 SMEs in more than 20 countries. Sustainability via a not-for-profit legal entity beyond the IMI phase will continue, expand, and develop the EHDEN open science community and network, as well as research programs, use cases, methodological innovation, and training.

5.3.2 DARWIN EU®

DARWIN EU® will deliver real-world evidence from across Europe on diseases, populations, and the uses and performance of medicines. This will enable EMA and national competent authorities in the European medicines regulatory network to use these data whenever needed throughout the lifecycle of a medicinal product [13].

DARWIN EU® will support regulatory decision-making by

- Establishing and expanding a catalog of observational data sources for use in medicines regulation.
- Providing a source of high-quality, validated real-world data on the uses, safety, and efficacy of medicines.
- Addressing specific questions by carrying out high-quality, non-interventional studies, including developing scientific protocols, interrogating relevant data sources, and interpreting and reporting study results.

The range of approved healthcare databases enabling distributed data access via DARWIN EU® will evolve and expand over time. The former HMA/EMA Big Data Task Force originally recommended developing DARWIN EU®. The creation of

DARWIN EU® features in the EMA-HMA Big Data Steering Group workplan and the European medicines agencies network strategy to 2025.

EMA will be a principal user of DARWIN EU®, by requesting studies to support its scientific evaluations and regulatory decision-making. A service provider will act as the DARWIN EU® Coordination Centre and be responsible for setting up the network and managing its day-to-day operations.

EMA will also play a central role in developing, launching, and maintaining DARWIN EU®, by

- Providing strategic direction and setting standards
- Overseeing the coordination center and monitoring its performance
- Ensuring close links to European Commission policy initiatives, particularly the EDHS, and delivering pilots
- Reporting to EMA's Management Board, the HMA and European Commission

The advent of DARWIN EU® will be a paradigm shift for regulatory science and decision-making in Europe, perhaps mirroring the FDA SENTINEL program, but incorporating the OMOP CDM and OHDSI research framework at its core, in a federated network. Ultimately, DARWIN EU® will be an accelerator for evidence-based decision-making using real-world data to complement clinical trial data and other data sources in providing insights into real-world outcomes, whether positive or negative.

5.3.3 European Health Data Space (EHDS)

In order to unleash the full potential of health data, the European Commission is presenting a regulation to set up the European Health Data Space, one of a number of data spaces across multiple industries and domains. Draft legislation was published in May 2022 for review and approval by the European Council and European Parliament [14].

The proposal

- Supports individuals to take control of their own health data
- Supports the use of health data for better healthcare delivery, better research, innovation, and policy-making
- Enables the EU to make full use of the potential offered by a safe and secure exchange, use, and reuse of health data

The European Health Data Space is a health-specific ecosystem comprising rules, common standards and practices, infrastructures, and a governance framework that aims at

- Empowering individuals through increased digital access to and control of their electronic personal health data, at national level and EU-wide, and support to their free movement, as well as fostering a genuine single market for electronic health record systems, relevant medical devices, and high-risk AI systems (primary use of data)

- Providing a consistent, trustworthy, and efficient setup for the use of health data for research, innovation, policy-making, and regulatory activities (secondary use of data)

As such, the European Health Data Space is a key pillar of the strong European Health Union, and it is the first common EU data space in a specific area to emerge from the European strategy for data. DARWIN EU® is designated as an EHDS pathfinder project, paving the way in its own development for the EHDS.

The legislative path will take some time, and implementation of any eventual, agreed legislation will not impact until the second half of the 2020s. An EC Joint Action, Towards the European Health Data Space (TEHDAS) commenced in February 2021 to develop European principles for the implementation of the EHDS, supported by 25 EU member states and a myriad of NGOs, SMEs, academic and commercial entities. TEHDAS is focused on [56].

- Solutions for the trustworthy secondary use of health and health care data with a view to promoting the digital transformation of European health systems
- Guidance on ensuring data quality such as anonymization of data and handling of data disparity

Work package 6 of TEHDAS is focused on excellence in data quality and has written a number or recommendation reports in 2022 on data interoperability, and data quality, providing frameworks and working concepts in these domains. The former lists a number of interoperability standards on data discoverability (at data source and variable levels) and on standards for the development of common data models, and describes some basic features: typology of interest, utility, and domain/s. This list is the basis for the work to come on aiming the description of their actual use, challenges in their implementation, issues on maintenance and sustainability.

The latter, on data quality, explores and synthesizes the existing knowledge and experiences on data quality frameworks (DQFs) in the context of cross-border sharing of federated secondary use health data with the aim to identify good practice within this area and make recommendations. The report builds on the work regarding data quality already undertaken the TEHDAS Joint Action and will be further updated with chapters on interoperability standards. This first part of the final report contains recommendations on the European Health Data Space (EHDS) data quality framework.

6 Conclusion

Numerous countries across the world are advancing their work on developing learning health systems, interoperable, federated networks, with FAIR data using agreed, aligned standards, data models and standardized analytical tools, and are at various stages. Most notably, the United States, the United Kingdom, certain EU

member states, such as Germany, France, Spain, Italy, Greece, and Finland, as well as South Korea and Singapore, China, Japan, and India are just some representing government or bottom-up initiatives.

Health data, whether used for care delivery, continuity of care, quality improvement, decision-making or research, is a critical success factor. The need to learn from health data at scale has never been more compelling, and its interoperability and quality are the key enablers of this scale. This chapter has explained the role and diversity of interoperability standards that are needed across the care and research spectrum. However, despite their individual maturity and capability of delivering interoperability, there remains a grand challenge of standards adoption. Too many health ICT products and networks either fail to take a standards-based approach to the health data they process or adopt only some standards in a patchy and highly customized way and so are not really interoperable.

The business drivers for the health ICT sector are recognized to be weak, and procurements insufficiently precise and stringent, so that the market push for standards adoption is too slow, as discussed in a recent multistakeholder round table report [57]. The report includes 14 recommendations and calls to action related to the greater uptake and promotion of interoperability, the first six of which are reproduced here as they relate to accelerating the adoption of standards and especially target actions the EC and Member States can make as part of implementing the European Health Data Space.

The report lists additional calls to action on interoperability relating to enforcing the adoption of interoperability standards by health ICT developers, the strategic governance of interoperability and ensuring wider awareness and engagement.

What has been evident for some time is that we have been attempting to meet increasing complex healthcare needs and evidence generation still using at best twentieth-century methods, and with an increasing emphasis on the value and potential of real-world data, advances have to be made. In particular in the regulatory domain, qualifications and approvals can no longer be reliant on clinical trial data alone, albeit remaining pivotal. Complementary developments for this type of data, such as the adoption of the CDISC model and family of standards, is being replicated with similar initiatives for real-world data, as described in this chapter, pointing to a radically different environment with mandated and aligned standards and models being implemented at scale across the global learning health system.

It is important for decision-makers, funders, ICT companies, and initiatives that seek to establish real-world evidence generation platforms and networks to ensure that they adopt and promote the wider adoption of interoperability standards, the FAIR principles and data quality as described in this chapter, and thereby contribute to a global momentum to scale up the usability of real-world data for trustworthy evidence generation.

References

1. Electronic Health Records Market Share, Size, Trends, Industry Analysis Report. Polaris Market Research, 2021. Available from https://www.polarismarketresearch.com/industry-analysis/electronic-health-records-ehr-market. Last accessed July 2022.
2. Kalra D. Scaling up the big health data ecosystem: engaging all stakeholders. J Int Soc Telemed eHealth 2020;8:e16
3. Brown S, Hudson C, Hamid A, Berman G, Echefu G, Lee K, Lamberg M, Olson J. The pursuit of health equity in digital transformation, health informatics, and the cardiovascular learning healthcare system. American Heart Journal Plus: Cardiology Research and Practice, 2022 (e-publication). https://doi.org/10.1016/j.ahjo.2022.100160
4. Horgan D, Hackett J, Westphalen C, Kalra D, Richer E, Romao M, Andreu A, Lal J, Bernini C, Tumiene B, Boccia S, Monserrat A. Digitalisation and COVID-19: The perfect storm. Biomed Hub 2020;5:511232. https://doi.org/10.1159/000511232
5. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. Nature. 2015 Oct 15;526(7573):336–42. https://doi.org/10.1038/nature15816
6. Kalra D. The importance of real-world data to precision medicine. Personalized Medicine 2018:16 (2);1–4. https://doi.org/10.2217/pme-2018-0120
7. Wise J, Möller A, Christie D, Kalra D, Brodsky E, Georgieva E, Jones G, Smith I, Greiffenberg L, McCarthy M, Arend M, Luttringer O, Kloss S, Arlington S. The positive impacts of Real-World Data on the challenges facing the evolution of biopharma. Drug Discovery Today 2018 Volume 23, Issue 4, April 2018, Pages 788–80. https://doi.org/10.1016/j.drudis.2018.01.034.
8. Horgan D, Romao M, Morré SA, Kalra D. Artificial Intelligence: Power for Civilisation – and for Better Healthcare. Public Health Genomics 2019;22:145–161. https://doi.org/10.1159/000504785
9. Kalra D, Strübin M. Editorial: Personal Health Systems. Frontiers in Medicine 2020; 7:694–695. https://doi.org/10.3389/fmed.2020.591070
10. Kölker S, Gleich F, Mütze U, Opladen T. Rare Disease Registries Are Key to Evidence-Based Personalized Medicine: Highlighting the European Experience. Front Endocrinol (Lausanne). 2022 Mar 4;13:832063. https://doi.org/10.3389/fendo.2022.832063.
11. Cuggia M, Combes S. The French Health Data Hub and the German Medical Informatics Initiatives Two National Projects to Promote Data Sharing in Healthcare. IMIA Yearbook of Medical Informatics, Schattauer, 2019, 28 (1), pp.195–202. https://doi.org/10.1055/s-0039-1677917
12. European Health Data & Evidence Network. Please see https://www.ehden.eu. Last accessed July 2022.
13. Data Analysis and Real World Interrogation Network. Please see https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu. Last accessed July 2022.
14. Please see https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en. Last accessed July 2022.
15. Singh G, Schulthess D, Hughes N, Vannieuwenhuyse B, Kalra D. Real world big data for clinical research and drug development. Drug Discov Today. 2017 Dec 30. pii: S1359-6446(17)30595-0. https://doi.org/10.1016/j.drudis.2017.12.002. PMID: 29294362.
16. The European Institute for Innovation through Health Data. How is health data being used to benefit society? Available from https://www.i-hd.eu/knowledge-center/how-is-health-data-being-used-to-benefit-society. Last accessed July 2022.
17. Suchard M, Schuemie M, Krumholz H, You S, Chen R, Pratt N et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. The Lancet 2019, 394: 1816–1826. https://doi.org/10.1016/S0140-6736(19)32317-7

18. Li X, Raventós B, Roel E, Pistillo A, Martinez-Hernandez E, Delmestri A et al. Association between covid-19 vaccination, SARS-CoV-2 infection, and risk of immune mediated neurological events: population based cohort and self-controlled case series analysis BMJ 2022; 376 :e068373 doi:https://doi.org/10.1136/bmj-2021-068373
19. Please see http://www.hl7.org/. Last accessed July 2022.
20. Please see https://www.iso.org/committee/54960.html. Last accessed July 2022.
21. ISO 13606 Electronic Health Record Communication. Please see https://www.iso.org/standard/67868.html. Last accessed August 2022.
22. ISO 13940 System of Concepts for Continuity of Care. Please see https://www.iso.org/standard/58102.html. Last accessed August 2022.
23. ISO 12967 Health Informatics Service Architecture. Please see https://www.iso.org/standard/58102.html. Last accessed August 2022.
24. Please see https://www.snomed.org/snomed-ct/why-snomed-ct. Last accessed July 2022.
25. Please see https://www.snomed.org/snomed-international/who-we-are. Last accessed July 2022.
26. Please see https://loinc.org. Last accessed July 2022.
27. Please see https://www.who.int/standards/classifications/classification-of-diseases. Last accessed July 2022.
28. Please see https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health. Last accessed July 2022.
29. Please see https://www.who.int/standards/classifications/international-classification-of-health-interventions. Last accessed July 2022.
30. Please see https://www.dicomstandard.org/. Last accessed July 2022.
31. Please see https://www.ihe.net/. Last accessed July 2022.
32. Please see https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory . Last accessed July 2022.
33. https://www.whocc.no/atc_ddd_index/. Last accessed July 2022.
34. Please see https://www.cms.gov/Medicare/Coding/ICD10/downloads/pcs_final_report2010.pdf. Last accessed July 2022.
35. Please see https://www.cdisc.org/. Last accessed July 2022.
36. Weeks, J and Pardee, R. 2019 Learning to Share Health Care Data: A Brief Timeline of Influential Common Data Models and Distributed Health Data Networks in U.S. Health Care Research. eGEMs (Generating Evidence & Methods to improve patient outcomes), 7(1): 4, pp. 1–7 https://doi.org/10.5334/egems.279
37. The EU General Data Protection Regulation. Please see https://gdpr-info.eu. Last accessed July 2022.
38. US Food and Drug Administration (FDA) Sentinel Initiative. Please see https://www.fda.gov/safety/fdas-sentinel-initiative. Last accessed July 2022.
39. The [US] National Patient-Centred Clinical Research Network (PCORNET). Please see https://pcornet.org. Last accessed July 2022.
40. Observational Health Data Science and Informatics. Please see https://www.ohdsi.org. Last accessed July 2022.
41. Vac4EU. Please see https://vac4eu.org. Last accessed July 2022.
42. The IMI Conception project. Please see https://www.imi-conception.eu. Last accessed July 2022.
43. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
44. Finnish Social and Health Data Permit Authority (Findata). Please see https://findata.fi/en/. Last accessed July 2022.
45. French Health Data Hub. Please see https://www.health-data-hub.fr (in French). Last accessed July 2022.

46. Food and Drug Administration Amendments Act. Please see https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/food-and-drug-administration-amendments-act-fdaaa-2007. Last accessed July 2022.
47. FDA Center for Biologics Evaluation and Research (CBER). Please see https://www.fda.gov/about-fda/fda-organization/center-biologics-evaluation-and-research-cber. Last accessed July 2022.
48. COVID-19 Study-a-thon 2020. Please see https://www.ehden.eu/covid19-study-a-thon/. Last accessed July 2022.
49. European Medicines Agency. Workshop report: A Common Data Model for Europe? – Why? Which? How?. London 2018. Available from https://www.ema.europa.eu/en/documents/report/common-data-model-europe-why-which-how-workshop-report_en.pdf. Last accessed July 2022.
50. The Book of OHDSI. Please see https://ohdsi.github.io/TheBookOfOhdsi/OhdsiAnalyticsTools.html. Last accessed August 2022.
51. GO FAIR. Please see https://www.go-fair.org. Last accessed July 2022.
52. FAIRplus. Please see https://fairplus-project.eu. Last accessed July 2022.
53. Patricia Flatley Brennan, William W. Stead, Assessing Data Quality: From Concordance, through Correctness and Completeness, to Valid Manipulatable Representations, Journal of the American Medical Informatics Association, Volume 7, Issue 1, January 2000, Pages 106–107, https://doi.org/10.1136/jamia.2000.0070106
54. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. Journal of the American Medical Informatics Association (JAMIA) 2021 Sep; 28(10): 2251–2257. https://doi.org/10.1093/jamia/ocab132.
55. The European institute for Innovation through Health Data. Data quality dimensions. Please see https://www.i-hd.eu/how-to-assess-data-quality-to-trust-what-you-learn/. Last accessed July 2022.
56. Towards the European Health Data Space. Please see https://tehdas.eu. Last accessed July 2022.
57. Scaling up the availability and reusability of big health data: 2021 Recommendations based on calls to action for health data ecosystems. Available from https://www.i-hd.eu/wp-content/uploads/2022/04/EHDS-Round-Table-3.pdf. Last accessed July 2022.

Privacy-Preserving Record Linkage for Real-World Data

Tianyu Zhan, Yixin Fang, and Weili He

1 Introduction and Motivation

Real-world data (RWD) are data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources [1]. These sources include electronic health records (EHRs), claims or billing activities, medical product or disease registries, patient-generated data including home-use settings, and health data gathered from other sources including mobile devices [2]. In addition to data from completed clinical trials, RWD is an emerging source of healthcare data that has become more readily available by the day [1, 2]. It is of great interest to aggregate several RWDs to provide new insights on health care outcomes.

Combining clinical trial data with RWD offers a more comprehensive longitudinal evaluation of health status beyond the maximum follow-up time of a clinical trial [3]. For example, a randomized controlled clinical trial (RCT) of evaluating pravastatin in preventing coronary heart disease was linked to routinely collected administrative health records [3, 4]. By increasing the follow-up time from 5 years to 15 years, this record linkage study was able to evaluate several long-term outcomes, including cardiovascular measures, quality-adjusted life years, and hospital administration status [4]. New research can be continuously conducted by performing data linkage with more recent RWD to have a better and full understanding of the initial intervention taken in the original RCT.

Another branch is to combine multiple RWD databases. Linking administrative claims data to EHR allows the researchers to leverage the complementary advantages of each data source to enhance study validity, as claim databases

T. Zhan (✉) · Y. Fang · W. He
Data and Statistical Sciences, AbbVie Inc., North Chicago, IL, USA
e-mail: tianyu.zhan.stats@gmail.com

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_7

usually contain extensive data on diagnoses, medications, healthcare utilization, and expenditure often lacking clinical details while EHR provides clinical details often absent in other datasets [3, 5]. EHR only captures patients' information within the specific healthcare network, but claim databases record all healthcare encounters. This allows for a more accurate or complete definition and evaluation of an exposure or outcome with proper and adequate confounding adjustment [3, 5]. Registry data can also be linked with EHR. For example, Alberta Cancer Registry data containing demographic and treatment information were linked to EHR, hospital discharge data, and census data [6]. Treatment patterns, adherence to treatment guidelines, and disparities in the receipt of treatment of colorectal cancer were investigated in this data linkage study [3, 6]. Moreover, the same type of RWD can be combined from different institutions. Hospital records from childhood and adulthood of patients with type 1 diabetes were linked to determine the relationship between glycaemic control trajectory and the long-term risk of severe complications [3, 7].

Record linkage or data linkage is a process of associating records from two or multiple datasets with the aim of identifying connections that belong to the same entity, for example, the same person [8]. Linking data from different sources plays an important role in improving data quality, enriching data for further analysis, and generating new insights [9]. This is a general method to enrich data with applications in many areas, such as health care [3, 10–12], finance [13–15], and business [16–18]. Generally speaking, record linkage of datasets within the same organization does not involve privacy and confidentiality concerns [9]. For example, a pharmaceutical company may link data from an RCT with its corresponding long-term extension study based on a unique subject identifier to comprehensively evaluate the maintenance of treatment effect. Similarly, the same database owner can link its claims and EHR before deidentification.

For RWD that are usually collected from a variety of sources or institutions [1], the process of data linkage should not disclose subject level identifying information per laws or regulations [9], for example, the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulations (GDPR) in Europe. Privacy-Preserving Record Linkage (PPRL) techniques are appealing in practice with the aim of identifying matching records that refer to the same entities in different databases without compromising privacy and confidentiality of these entities [9, 19, 20]. A cohort of quasi-identifiers is encoded to mask confidential information and then utilized to link records [9]. To ensure patients' privacy, some variables, for example, status of a rare disease, will not be allowed in linkage when there is risk of identifying specific patients.

In this chapter, we provide a high-level review of PPRL to motivate its applications to RWD. We review several methods for data preparation in Sect. 2 and methods for linkage in Sect. 3. Some performance evaluation approaches are discussed in Sect. 4. An illustration of performance probabilistic record linkage on real datasets is presented in Sect. 5. Concluding remarks are provided in Sect. 6.

2 Data Preparation Methods

2.1 Data Preprocessing Methods

Data preprocessing refers to the task of converting raw data to a standard format for accurate and efficient matching [21, 22]. There are generally three steps in this process. First, data cleaning is conducted to remove unrelated or unwanted information for matching, delete duplicated records, and convert inputs to a consistent form. Either hard-coded rules or look-up tables are used in this step [22]. The second step is utilizing look-up tables to standardize tokens, which are usually referred to as the values that are separated by whitespace characters in attributes, with the goal of correcting typographical errors or variations, or standardizing abbreviations [22]. For example, "bevely park" or "bevelly park" is standardized as "beverley park" based on the look-up table in the FEBRL system [22, 23]. The third step is the segmentation of the tokenized attribute values into single pieces of information that are suitable for downstream data matching [22]. The challenge is to identify the most likely and meaningful assignment because there are often several possible assignments of tokens [22].

Since data from multiple sources are to be harmonized, the Observational Medical Outcomes Partnership (OMOP) Common Data Model [24] can be adopted to allow for the systematic analysis of disparate observational databases. Data are transformed to a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes) and then are utilized to perform systematic analyses using a library of standard analytic routines that have been written based on the common format.

In PPRL, data masking or encoding is an additional step to transform original data to masked data [25]. Data elements, such as names or social security numbers, need to be de-identified to protect privacy. Moreover, several relatively nonsensitive attributes need to be masked as well because their combination may reveal identifying information. For example, nearly 90% of the U.S. population had a unique combination of zip code, gender, and date of birth [22, 26]. The level of such deidentification is important because a mild layer can still disclose private and sensitive data, while a heavy one may lose discriminating power to distinguish between matches and non-matches. Several specific techniques are reviewed in the next section.

2.2 Privacy Protection Methods

If there is no privacy concern, one can directly link records from different datasets with personal identifying information if available. However, in applications to RWD, especially RWD from different institutions, such data linkage should

be conducted without disclosing privacy and confidentiality information. In this section, we review some techniques for protecting privacy in record linkage.

2.2.1 Separation Principle

The data can usually be separated into personally identifying data which contain sensitive information and content data with clinical information for research [25, 27]. A data custodian sends personally identifying information to a linkage unit or a third party, which performs data linkage to determine which records belong to the same person. Such linkage information is sent back to the data custodian. Then, researchers receive content data along with linkage information from the data custodian to perform further analysis.

This separation principle is classified under the so-called three-party protocols that utilize a third party for performing the linkage [25]. As compared with its counterpart "two-party protocols" with no third-party involvement, three-party protocols require fewer resources in communication and computation to compare records but are also considered less secure due to the existence of a third party [25].

2.2.2 Secure Hash Encoding

This technique uses one-way irreversible hash encoding functions to convert sensitive information to hash code [28–30]. Having access to a hash code makes it nearly impossible with current computing technology to learn its original string value [30]. However, dictionary attack is possible with masking functions, where an adversary masks a large list of known values using various existing masking functions until a matching masked value is identified [30]. A possible mitigation is the Hashed Message Authentication Code (HMAC) as a keyed masking approach [31]. With HMAC, dataset owners exchange and add a secret code to data before masking [9]. A major limitation of secure hash encoding is that it can only adopt deterministic linkage methods to identify exact matches, but not probabilistic linkage, because even a single character difference in a string will lead to a completely different hash code [30]. As discussed in Sect. 3.2, probabilistic linkage has advantages of accommodating data entry error when performing matching between records.

2.2.3 Phonetic Encoding

Phonetic encoding techniques convert string to code based on pronunciation [32]. For example, Soundex is the best known phonetic encoding algorithm [33]. It keeps the first letter and converts the rest into numbers according to an encoding table [32]. Phonetic encoding inherently provides privacy and is a blocking technique of reducing the number of comparisons in linkage to increase scalability [9]. It also supports probabilistic linkage to tolerate typographical variations [30, 34]. Two

drawbacks of phonetic encodings are that they are language dependent and are vulnerable to frequency attacks, where the frequency distribution of a set of masked values is matched with the distribution of known unmasked values in order to infer the original values of the masked values [9, 35].

2.2.4 Bloom Filters

Bloom filters technique was proposed by Schnell et al. [20] to calculate the similarity between two encrypted strings for use in probabilistic record linkage procedures. It first converts a string to a set of consecutive letters (q-grams) or a set of tokens [36] and then computes similarity between two strings by the Dice coefficient [20]. Bloom filter demonstrates high quality in the evaluation of privacy-preserving string comparison [25, 37]. Filtering techniques can also be applied based on Bloom filters to increase scalability to large datasets, for example, excluding unnecessary comparisons based on q-grams [9].

3 Linkage Methods

After strings are encrypted to mask personal identifying information, the next step is to merge datasets by finding matching records. Deterministic linkage and probabilistic linkage are two common methods [16, 38–40].

3.1 Deterministic Linkage

In the deterministic linkage, only record pairs that matched exactly are accepted as links [40]. This can be based on a single attribute or several attributes. This method is easy to implement in practice and can be applied to most methods of masking personal identifying information including Secure Hash Encoding discussed in Sect. 2.2.2. This method is typically computationally more efficient as compared with the probabilistic linkage as discussed next. A major limitation of this method is that even a single character difference of data entry error between a pair of original values results in a matched classification [9].

3.2 Probabilistic Linkage

Probabilistic linkage methodology addresses record linkage problems under conditions of uncertainty [41] and allows imperfect matches due to partially inaccurate or missing data [40]. The Fellegi–Sunter method [39] is a popular and well-known

algorithm for probabilistic record linkage [42]. For each record pair, a weight is computed based on the probability that the field agrees given a record pair matches (called the m probability) and the probability of chance agreement given an unmatched pair (called the u probability) [39, 41]. A composite weight for each record can also be calculated for multiple linkage variables adjusting agreement or disagreement status for each variable, with zero weight assigned for missing values [40]. A threshold on the weight is chosen to classify records as matches or non-matches [40], or a consider zone for clerical review with another cutoff value [41]. The specific setting of cutoff value is critical and difficult in probabilistic linkage [43] and can be selected to optimize f-measure, introduced in the next section [40].

Typical probabilistic linkage methods classify individual record pairs independently from other pairs and therefore aim at a many-to-many matching scenario [22]. Additional restrictions can be applied to accommodate one-to-one and one-to-many matching scenarios. In one-to-one matching, a simple approach is to sort the matched pairs based on similarity values and then assign pairs to confirmed matches in a greedy fashion [22]. However, this method may yield a sub-optimal solution because it does not consider all records simultaneously and it is possible that not all records can be paired. Several more advanced methods have been developed to solve this constrained optimization problem [44], for example, treating a class of algorithms as an auction problem [45].

There are also quite a few Bayesian record linkage techniques proposed to accommodate uncertainty. A fully Bayesian approach to record linkage was developed to compute posterior probability of matching [46]. In a unified Bayesian framework, matching uncertainty is naturally accounted for in estimating population size by using samples of multivariate categorical variables [47]. A Bayesian graphical approach is proposed to simultaneously detect duplicate records within files and link records across files [48]. Partial Bayes estimates were derived for bipartite matching to quantify uncertainty in matching decisions while leaving uncertain parts undeclared [49].

3.3 Unsupervised Classification Methods

With the objective of protecting privacy, the classification labels of either matched or unmatched records are not available. Supervised classification methods cannot be directly applied. Alternatively, unsupervised classification or clustering methods can be adopted to PPRL. For example, K-means algorithm is a popular iterative clustering method based on similarity measure [50, 51]. The Damerau–Levenstein distance [52, 53] or the Jaro–Winkler distance [54, 55] can be used to measure similarity for text-based variables [42]. Other unsupervised methods include agglomerative clustering (bottom-up) and divisive clustering (top-down) as two paradigms in hierarchical clustering, self-organizing maps, etc. [56]

4 Performance Evaluation

We first review some measures and then some methods to assess performance of PPRL.

4.1 Measures

Since record linkage is a classification problem, there are two types of errors that can be generated: false negative (FN) and false positive (FP) [40]. A higher number of FNs contribute to a lower sensitivity, which is defined as

$$\text{Sensitivity} = \frac{\text{Number of True Positive}}{\text{Number of True Positive} + \text{FN}}, \tag{1}$$

while FP is related to positive predictive value (PPV),

$$\text{PPV} = \frac{\text{Number of True Positive}}{\text{Number of True Positive} + \text{FP}}. \tag{2}$$

Ideally, both FN and FP need to be minimized, but there is a trade-off between these two types of error in practice [40]. F-measure,

$$\text{f-measure} = 2 \times \frac{\text{Sensitivity} \times \text{PPV}}{\text{Sensitivity} + \text{PPV}}, \tag{3}$$

is a harmonic mean of sensitivity and PPV [43]. The f-measure reaches a high value only if both sensitivity and PPV are high [40] and is more appealing in practice than single metrics [43]. Other measures, such as area under the receiver operating characteristic (ROC) curve and Youden's index, can also be used to evaluate performance.

Note that some common evaluation metrics for classification problems may not be proper for record linkage, for example accuracy, which is defined as the total number of true positives and true negatives divided by the total number of pairs. The reason is that the majority of record pairs correspond to non-matched pairs (true negatives), and the number of true negatives dominates the calculation of accuracy [22].

4.2 Assessment Method

In practice, it is challenging to evaluate linkage performance based on the above measures because the underlying true labels of either matched or unmatched

pairs are not available. Manual assessment of all individual records would reveal sensitive information, which is in contradiction to the objective of PPRL [9]. This comprehensive manual review is also not feasible given the relatively large number of comparisons with even moderate datasets. Moreover, even with personal identifying information, there may not exist a gold standard because several RWDs do not share the same unique identifier [57].

A clerical review can be implemented by a third party to routinely scan and manually review a small proportion of links [25, 58]. However, this selective review can also be time-consuming and may not be feasible for large datasets [25]. There is an increased privacy risk because personal sensitive information is regularly manually examined [25]. Under the framework of interactive PPRL, parts of sensitive data are revealed for manual assessment [9, 59]. However, there are still some open questions in real applications, for example, how to ensure the revealed information is limited to a certain level of detail and is also sufficient for manual assessment [9].

An alternative approach to obtain benchmark datasets is to generate synthetic data based on the characteristics of real data, for example, distributions of variables and proportion of missing data [9, 57, 60, 61]. Given synthetic data with known classification labels, one can perform cross-validation to fine-tune parameters in data linkage, for example the cutoff values in probabilistic linkage [62]. Multiple replicates of synthetic data can be simulated to report the average of certain evaluation measure [57].

5 Demonstration with the R Package `RecordLinkage` on Dataset NHANES

In this section, we illustrate how to perform probabilistic record linkage with the R package `RecordLinkage` [63] on the real dataset NHANES (National Health and Nutrition Examination Survey) [64] from CDC (Centers for Disease Control and Prevention). NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States with interviews including demographic, socioeconomic, dietary, and health-related questions [64]. This example is for illustrative purposes. Variables being included are not intended to identify unique patients, as NHANES might enroll different participants across years.

For demonstration, we consider the first ten records in two demographics files of survey cycle 2015–2016 and 2017–2018 with variables RIAGENDR (Gender), RIDRETH1 (Race), and DMDBORN4 (Country of Birth). RIAGENDR is utilized as the blocking variable. String comparison is based on RIDRETH1 and DMDBORN4 with the Jaro–Winkler method [55] to compute the similarity between strings. The threshold values are set as -4 and 0, such that records with matching weights less than -4 are classified as non-links, records with weights greater than

Table 1 Linked pairs detected by probabilistic linkage

Row index of dataset 1	Row index of dataset 2
1	2
3	2
8	2
4	8
5	3
5	7

or equal to 0 are classified as links, and remaining records are possible links for clerical review.

There are 50 record pairs for evaluation with the blocking of RIAGENDR. Based on the classification of probabilistic linkage, 19 pairs are categorized as non-links, 25 as possible links, and 6 as links, which are shown in Table 1. Record No. 4 from dataset 1 is uniquely linked to record No. 8 from dataset 2. Since each pair is categorized independently, a record from one dataset can be linked to multiple records in the other dataset. For example, records No. 1, 3, and 8 from dataset 1 are linked to the same record No. 2 in dataset 2, while records No. 3 and 7 from dataset 2 are linked to record No. 5 in dataset 1. Clerical review can be further performed to determine potential exact one-to-one mapping for record No. 5 in dataset 1 and record No. 2 in dataset 2.

```
## load R packages
library(RecordLinkage); library(foreign)

## code to import file: https://wwwn.cdc.gov/nchs/data/tutorials/
     file_download_import_R.R
## Download NHANES 2015-2016 to temporary file
download.file("https://wwwn.cdc.gov/nchs/nhanes/2015-2016/DEMO_I.
     XPT", tf1 <- tempfile(), mode="wb")
## Create Data Frame From Temporary File
DEMO_I3 <- foreign::read.xport(tf1)

## Download NHANES 2017-2018 to temporary file
download.file("https://wwwn.cdc.gov/nchs/nhanes/2017-2018/DEMO_J.
     XPT", tf2 <- tempfile(), mode="wb")
## Create Data Frame From Temporary File
DEMO_J3 <- foreign::read.xport(tf2)

## Create data with the first 10 records and three variables:
## RIAGENDR: Gender, RIDRETH1: Race, DMDBORN4: Country of birth
DEMO_I3_OUT = DEMO_I3[1:10, intersect(colnames(DEMO_I3), colnames
     (DEMO_J3))[c(4, 7, 13)]]
DEMO_J3_OUT = DEMO_J3[1:10, intersect(colnames(DEMO_I3), colnames
     (DEMO_J3))[c(4, 7, 13)]]

## Convert RIAGENDR and RIDRETH1 to character variables
DEMO_I3_OUT[, 2] = as.character(DEMO_I3_OUT[, 2])
DEMO_J3_OUT[, 2] = as.character(DEMO_J3_OUT[, 2])
```

```
DEMO_I3_OUT[, 3] = as.character(DEMO_I3_OUT[, 3])
DEMO_J3_OUT[, 3] = as.character(DEMO_J3_OUT[, 3])

## Perform record linkage with RIAGENDR as a block variable, and
     a similarity function based on Levenshtein distance of
     variables RIAGENDR and RIDRETH1
rpairs=compare.linkage(DEMO_I3_OUT,
DEMO_J3_OUT,
blockfld=c(1),
strcmp =c(2, 3),
strcmpfun = jarowinkler
)

# calculate weights based on default m and u probabilities
rpairs.w <- fsWeights(rpairs, m = 0.95, u=rpairs$frequencies)

# classify records with thresholds -4 and 0
rpairs.fit = fsClassify(rpairs.w, threshold.upper = 0, threshold.
     lower = -4)

# show results
print(summary(rpairs.fit))

# show linked records
rpairs.fit$pairs$is_match = rpairs.fit$prediction
print(rpairs.fit$pairs[rpairs.fit$pairs$is_match=="L", ])
```

6 Discussion

There are several additional points to consider when performing PPRL on real-world data. First of all, missing data or missing values are common in real-world data. A simple method is to remove records or attributes with missing values, but this leads to information loss [22]. Rule-based imputation methods are more proper to take account of distributions of attributes and correlations between attributes [65, 66]. Alternatively, the probabilistic linkage method discussed in Sect. 3.2 can intrinsically handle this by assigning zero weights for missing attributes when calculating the composite weight.

Another challenge is scalability because the number of potential pairs for evaluation is the product of the number of records in two datasets leading to quadratic complexity. Blocking or indexing is a common technique to eliminate non-matched records. For example, standard blocking uses the values of so-called blocking key values (BKVs) to partition all records into disjoint blocks [25]. To accommodate incorrect or missing BKVs, one can conduct blocking in an iterative fashion. The non-matched pairs filtered by the first BKV are further sent to the second BKV for partition, and so on and so forth to the last BKV [25]. This can be viewed as a hybrid framework to combine deterministic linkage in Sect. 3.1

for blocking and probabilistic linkage in Sect. 3.2 for downstream matching. Other filtering techniques are also available to reduce the search space based on similarity measures and the length of tokens [25].

The results from data linkage on RWD may also be used to support regulatory decision-making for study drugs. Based on a recent FDA guidance on RWD [67], the protocol should clearly describe data sources, the information that will be obtained, linkage methods, and the accuracy and completeness of data linkages over time. Sensitivity analysis should also be performed to evaluate the robustness of results based on probabilistic linkage methods [67].

PPRL is a relatively new area in record linkage. To apply PPRL on RWD, there are several challenges and future research topics. Additional work is needed to guide statistical inference of estimates from integrated datasets under potential mismatch errors. This problem is even more challenging to evaluate empirically because true classification labels are not available due to privacy concerns. Generating realistic synthetic data is in itself a formidable challenge [22]. The missing data issue adds another layer of challenge because the assumption of missing not at random can be more common in RWD.

References

1. FDA. Framework for FDA's Real-World Evidence Program. 2018. https://www.fda.gov/media/120060/download.
2. Jie Chen, Martin Ho, Kwan Lee, Yang Song, Yixin Fang, Benjamin A Goldstein, Weili He, Telba Irony, Qi Jiang, Mark van der Laan, et al. The current landscape in biostatistics of real-world data and evidence: clinical study design and analysis. *Statistics in Biopharmaceutical Research*, pages 1–14, 2021.
3. Donna R Rivera, Mugdha N Gokhale, Matthew W Reynolds, Elizabeth B Andrews, Danielle Chun, Kevin Haynes, Michele L Jonsson-Funk, Kristine E Lynch, Jennifer L Lund, Helen Strongman, et al. Linking electronic health data in pharmacoepidemiology: appropriateness and feasibility. *Pharmacoepidemiology and Drug Safety*, 29(1):18–29, 2020.
4. Alex McConnachie, Andrew Walker, Michele Robertson, Laura Marchbank, Julie Peacock, Christopher J Packard, Stuart M Cobbe, and Ian Ford. Long-term impact on healthcare resource utilization of statin treatment, and its cost effectiveness in the primary prevention of cardiovascular disease: a record linkage study. *European Heart Journal*, 35(5):290–298, 2014.
5. Kueiyu Joshua Lin and Sebastian Schneeweiss. Considerations for the analysis of longitudinal electronic health records linked to claims data to study the effectiveness and safety of drugs. *Clinical Pharmacology & Therapeutics*, 100(2):147–159, 2016.
6. N Sharaf Eldin, Y Yasui, A Scarfe, and M Winget. Adherence to treatment guidelines in stage II/III rectal cancer in Alberta, Canada. *Clinical Oncology*, 24(1):e9–e17, 2012.
7. Mary White, Matthew A Sabin, Costan G Magnussen, Michele A O'Connell, Peter G Colman, and Fergus Cameron. Long term risk of severe retinopathy in childhood-onset type 1 diabetes: a data linkage study. *Medical Journal of Australia*, 206(9):398–401, 2017.
8. M Winglee, R Valliant, and F Scheuren. A case study in record linkage. *Surv Methodol*, 31(1):3–11, 2005.
9. Dinusha Vatsalan, Ziad Sehili, Peter Christen, and Erhard Rahm. Privacy-preserving record linkage for big data: Current approaches and research challenges. In *Handbook of Big Data Technologies*, pages 851–895. Springer, 2017.

10. Douglas Iain Ross Boyle and Naomi Rafael. Biogrid Australia and GRHANITE$^{\mathring{U}}$: Privacy-protecting subject matching. In *Health Informatics: The Transformative Power of Innovation*, pages 24–34. IOS Press, 2011.
11. James H Boyd, Anna M Ferrante, Christine M O'Keefe, Alfred J Bass, Sean M Randall, and James B Semmens. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. *BMC Health Services Research*, 12(1):1–8, 2012.
12. Nicole L Pratt, Christina D Mack, Anne Marie Meyer, Kourtney J Davis, Bradley G Hammill, Christian Hampp, Soko Setoguchi, Sudha R Raman, Danielle S Chun, Til Stürmer, et al. Data linkage in pharmacoepidemiology: A call for rigorous evaluation and reporting. *Pharmacoepidemiology and Drug Safety*, 29(1):9–17, 2020.
13. Kunho Kim and C Lee Giles. Financial entity record linkage with random forests. In *Proceedings of the Second International Workshop on Data Science for Macro-Modeling*, pages 1–2, 2016.
14. Ian Kloo, Matthew F Dabkowski, and Samuel H Huddleston. Improving record linkage for counter-threat finance intelligence with dynamic Jaro-Winkler thresholds. In *2019 Winter Simulation Conference (WSC)*, pages 2467–2478. IEEE, 2019.
15. Antonio Maratea, Angelo Ciaramella, and Giuseppe Pio Cianci. Record linkage of banks and municipalities through multiple criteria and neural networks. *PeerJ Computer Science*, 6:e258, 2020.
16. Howard B Newcombe. *Handbook of record linkage: methods for health and statistical studies, administration, and business*. Oxford University Press, Inc., 1988.
17. William E Winkler. Matching and record linkage. *Business Survey Methods*, 1:355–384, 1995.
18. Andrew T Schneider, Arjun Mukherjee, and Eduard C Dragut. Leveraging social media signals for record linkage. In *Proceedings of the 2018 World Wide Web Conference*, pages 1195–1204, 2018.
19. Chris Clifton, Murat Kantarcioğlu, AnHai Doan, Gunther Schadow, Jaideep Vaidya, Ahmed Elmagarmid, and Dan Suciu. Privacy-preserving data integration and sharing. In *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 19–26, 2004.
20. Rainer Schnell, Tobias Bachteler, and Jörg Reiher. Privacy-preserving record linkage using bloom filters. *BMC medical informatics and decision making*, 9(1):1–11, 2009.
21. Tim Churches, Peter Christen, Kim Lim, and Justin Xi Zhu. Preparation of name and address data for record linkage using hidden Markov models. *BMC Medical Informatics and Decision Making*, 2(1):1–16, 2002.
22. Peter Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, 2012.
23. Peter Christen. Development and user experiences of an open source data cleaning, deduplication and record linkage system. *ACM SIGKDD Explorations Newsletter*, 11(1):39–48, 2009.
24. OHDSI. Observational Health Data Sciences and Informatics. https://ohdsi.org/.
25. Sean M Randall, Anna M Ferrante, James H Boyd, Jacqueline K Bauer, and James B Semmens. Privacy-preserving record linkage on large real world datasets. *Journal of Biomedical Informatics*, 50:205–212, 2014.
26. Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.
27. Christopher W Kelman, A John Bass, and Cashel DJ Holman. Research use of linked health data—a best practice protocol. *Australian and New Zealand Journal of Public Health*, 26(3):251–255, 2002.
28. L Dusserre, C Quantin, and H Bouzelat. A one way public key cryptosystem for the linkage of nominal files in epidemiological studies. *Medinfo. MEDINFO*, 8:644–647, 1995.
29. Catherine Quantin, Hocine Bouzelat, FAA Allaert, Anne-Marie Benhamiche, Jean Faivre, and Liliane Dusserre. How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure. *International Journal of Medical Informatics*, 49(1):117–122, 1998.

30. Dinusha Vatsalan, Peter Christen, and Vassilios S Verykios. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969, 2013.
31. Hugo Krawczyk, Mihir Bellare, and Ran Canetti. HMAC: Keyed-hashing for message authentication, 1997.
32. Peter Christen. A comparison of personal name matching: Techniques and practical issues. In *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)*, pages 290–294. IEEE, 2006.
33. David Holmes and M Catherine McCabe. Improving precision and recall for Soundex retrieval. In *Proceedings. International Conference on Information Technology: Coding and Computing*, pages 22–26. IEEE, 2002.
34. Alexandros Karakasidis, Vassilios S Verykios, and Peter Christen. Fake injection strategies for private phonetic matching. In *Data Privacy Management and Autonomous Spontaneous Security*, pages 9–24. Springer, 2011.
35. Hongbo Liu, Hui Wang, and Yingying Chen. Ensuring data storage security against frequency-based attacks in wireless networks. In *International Conference on Distributed Computing in Sensor Systems*, pages 201–215. Springer, 2010.
36. Elizabeth A Durham, Murat Kantarcioglu, Yuan Xue, Csaba Toth, Mehmet Kuzu, and Bradley Malin. Composite bloom filters for secure record linkage. *IEEE transactions on knowledge and data engineering*, 26(12):2956–2968, 2013.
37. Elizabeth Durham, Yuan Xue, Murat Kantarcioglu, and Bradley Malin. Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. *Information Fusion*, 13(4):245–259, 2012.
38. Howard B Newcombe, James M Kennedy, SJ Axford, and Allison P James. Automatic linkage of vital records. *Science*, 130(3381):954–959, 1959.
39. Ivan P Fellegi and Alan B Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.
40. Ying Zhu, Yutaka Matsuyama, Yasuo Ohashi, and Soko Setoguchi. When to conduct probabilistic linkage vs. deterministic linkage? a simulation study. *Journal of biomedical informatics*, 56:80–86, 2015.
41. Matthew A Jaro. Probabilistic linkage of large public health data files. *Statistics in Medicine*, 14(5–7):491–498, 1995.
42. Jana Asher, Dean Resnick, Jennifer Brite, Robert Brackbill, and James Cone. An introduction to probabilistic record linkage with a focus on linkage processing for WTC registries. *International Journal of Environmental Research and Public Health*, 17(18):6937, 2020.
43. Peter Christen and Karl Goiser. Quality and complexity measures for data linkage and deduplication. In *Quality measures in data mining*, pages 127–151. Springer, 2007.
44. Richard J Trudeau. *Introduction to graph theory*. Courier Corporation, 2013.
45. Dimitri P Bertsekas. Auction algorithms for network flow problems: A tutorial introduction. *Computational Optimization and Applications*, 1(1):7–66, 1992.
46. Marco Fortini, Brunero Liseo, Alessandra Nuccitelli, and Mauro Scanu. On Bayesian record linkage. *Research in Official Statistics*, 4(1):185–198, 2001.
47. Andrea Tancredi and Brunero Liseo. A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, 5(2B):1553–1585, 2011.
48. Rebecca C Steorts, Rob Hall, and Stephen E Fienberg. A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516):1660–1672, 2016.
49. Mauricio Sadinle. Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112(518):600–612, 2017.
50. Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.
51. Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.
52. Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

53. Fred J Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.
54. Matthew A Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.
55. William E Winkler. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, page 354–369, 1990.
56. Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
57. Katie Harron, Ruth Gilbert, David Cromwell, and Jan van der Meulen. Linking data for mothers and babies in de-identified electronic health data. *PloS One*, 11(10):e0164667, 2016.
58. Glenda Lawrence, Isa Dinh, and Lee Taylor. The centre for health record linkage: a new resource for health services research and evaluation. *Health Information Management Journal*, 37(2):60–62, 2008.
59. Hye-Chung Kum, Ashok Krishnamurthy, Ashwin Machanavajjhala, Michael K Reiter, and Stanley Ahalt. Privacy preserving interactive record linkage (PPIRL). *Journal of the American Medical Informatics Association*, 21(2):212–220, 2014.
60. Peter Christen and Dinusha Vatsalan. Flexible and extensible generation and corruption of personal data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1165–1168, 2013.
61. Khoi-Nguyen Tran, Dinusha Vatsalan, and Peter Christen. GeCo: an online personal data generator and corruptor. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2473–2476, 2013.
62. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.
63. Murat Sariyar and Andreas Borg. *RecordLinkage: Record Linkage Functions for Linking and Deduplicating Data Sets*, 2022. R package version 0.4-12.3.
64. National Health and Nutrition Examination Survey, howpublished = https://www.cdc.gov/nchs/nhanes/index.htm, note = Accessed: 2022-01-23.
65. Ivan P Fellegi and David Holt. A systematic approach to automatic edit and imputation. *Journal of the American Statistical association*, 71(353):17–35, 1976.
66. Thomas N Herzog, Fritz J Scheuren, and William E Winkler. *Data quality and record linkage techniques*. Springer Science & Business Media, 2007.
67. FDA. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products. 2021. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory.

Part III
Causal Inference Framework and Methodologies in RWE Research

Causal Inference with Targeted Learning for Producing and Evaluating Real-World Evidence

Susan Gruber, Hana Lee, Rachael Phillips, and Mark van der Laan

1 Introduction

Targeted Learning (TL) is a statistical framework for efficient learning from data [1]. TL's systematic approach addresses many of the critical barriers in analyses of studies incorporating real-world data (RWD), including any non-randomization of the exposure, treatment non-compliance, time-varying confounding, and incomplete capture of the outcome [2]. Core principles for valid causal inference include (1) specifying a causal model and realistic statistical model consistent with expert knowledge and characteristics of the data generating process; (2) specifying a target of estimation (i.e., estimand) consistent with the goals, design, and conduct of the study; (3) analyzing the data using targeted minimum loss-based estimation (TMLE), a generalization of targeted maximum likelihood estimation, coupled with super learning (SL); and (4) assessing robustness of study findings via diagnostics and sensitivity analyses. The TL estimation roadmap codifies these principles and the use of TMLE + SL for optimally estimating causal effects and association

S. Gruber (✉)
Putnam Data Sciences, LLC, Cambridge, MA, USA
e-mail: sgruber@putnamds.com

H. Lee
Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA

R. Phillips · M. van der Laan
Division of Biostatistics, University of California at Berkeley, Berkeley, CA, USA

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_8

measures from data [1–3]. Despite legitimate concerns, RWD can provide valuable insights in areas increasingly important for regulatory, policy, and clinical decision-making.

A 2018 draft framework issued by the US Food and Drug Administration (FDA) highlights opportunities and challenges in producing real-world evidence (RWE) in support of regulatory decision-making [4]. RWD sources, such as electronic health records (EHR), medical claims, product and disease registries, and personal wearable devices, produce copious amounts of data. However, generating reliable RWE about the use, risks, and benefits of medical products necessitates careful study design, conduct, data analysis, and interpretation. Key considerations include whether the RWD are fit for purpose, whether the study provides adequate scientific evidence, and whether study conduct meets regulatory requirements [4–8].

Studies ranging from randomized controlled trials (RCT) in clinical settings, to non-randomized interventional single arm trials with external controls, to observational studies (OS) rely in varying degrees on RWD [9]. However, RWD sometimes suffers from incomplete or mis-specified measures of subject characteristics, exposures, and outcomes. Intercurrent events can disrupt the measure or interpretation of the outcome. Whether or not treatment is randomized, these aspects of RWD increase the difficulty of drawing accurate, interpretable insights into safety and efficacy in broad populations under real-world conditions.

This chapter describes the TL approach to causal inference that addresses these challenges. The TL roadmap provides a step-by-step guide to producing and evaluating RWE [1–3, 10, 11]. It accounts for all components of the ICH E9(R1) Guideline definition of an estimand: *population, treatment, outcome variable*, s*ummary measure*, and *intercurrent events* [12]. In alignment with the guidelines, the roadmap defines the target causal estimand as a parameter of the probability distribution of the data. Initial steps in the roadmap characterize the data-generating process prior to data collection. This promotes transparent definitions of the causal estimand in terms of a causal model and the statistical estimand in terms of a statistical model. The choice of estimator and the scope of sensitivity analyses are also pre-specified.

TMLE+SL provide efficient, consistent estimation of statistical parameters, and inference. The final step in the roadmap offers a transparent process for assessing the validity of a causal interpretation. These concepts are illustrated through an analysis of time-to-event data from a single arm study with a synthetic external control arm. The population of interest is defined by the population included in the single arm study. TMLE+SL are used to evaluate the marginal cumulative incidence ratio of treatment versus comparator among the treated (ATT). The chapter concludes with a summary of other ways to utilize RWD throughout the pharmaceutical pipeline with the help of TL.

2 Targeted Learning Estimation Roadmap

2.1 Step 0

Step 0 of the roadmap (Fig. 1) concerns the clinical question of interest and the plan for acquiring study data that is suitable for addressing the substantive question of interest to domain experts. This question is initially expressed in terms of a gap in scientific knowledge, and will ultimately be formulated as a statistical question that can be answered from data. Each of the five ICH E9(R1) elements of an estimand are clarified in this step: (1) Inclusion/exclusion criteria define the *study population*; (2) precise definitions of the *treatment,* including time of initiation, background therapies, and comparators to define each study arm; (3) clear criteria for identifying or measuring the *outcome* on or before a specific follow-up period; (4) a meaningful *summary measure*, such as a risk difference, hazard ratio (HR), or dose-response curve, is decided upon [12]; (5) identify likely *intercurrent events*, such as treatment non-adherence, loss-to-follow-up (LTFU), and competing risks. These post-randomization events can potentially disrupt the treatment-outcome associations in the data and/or have an impact on defining the (identifiable and estimable) treatment effect. Therefore, considering intercurrent events this early in the process allows the study team to identify an appropriate strategy for ameliorating their impact on the eventual study finding, collecting relevant data, and/or to define the realistic estimand that respects the underlying data-generation. For example, experts can consider whether incorporating a competing risk into a composite outcome is the appropriate scientific question to investigate, e.g., "stroke or myocardial infarction" vs. "myocardial infarction" alone. With these elements in mind, Step 0 culminates in characterizing the process that gives rise to the data over time.

For a simple example, consider a hypothetical retrospective cohort study to compare the impact of utilizing etomidate/midazolam as a sedative during routine screening colonoscopy versus a comparator drug propofol/midazolam on systolic blood pressure at the end of sedation. The summary measure will be the marginal additive treatment effect (ATE). The population of interest consists of non-pregnant

Fig. 1 The targeted learning estimation roadmap

Step 0. Formulate the substantive question(s), and describe the experiment giving rise to the data
Step 1. Define a realistic statistical model for the data
Step 2. Define a causal model, and causal parameter of interest
Step 3. Specify statistical parameter, and identifying assumptions
Step 4. Estimation and Inference using TMLE + SL
Step 5. Interpretation and substantive conclusion, supported by sensitivity analyses

adults between the ages of 20 and 85 who had a screening colonoscopy at the outpatient center of an urban hospital between January 1, 2019 and December 31, 2019. The dataset will consist of independent and identically distributed (i.i.d.) observations $O = (W, A, Y)$, where $W$is a vector of baseline covariates (*age, sex, pulse*, systolic blood pressure (*SBP*), irritable bowel syndrome (*IBS*, y/n), A is a binary indicator of treatment with etomidate ($A = 1$) or propofol ($A = 0$), and Y is the post-sedation SBP. Assume that subject matter experts have ensured that W contains all potential confounders of the treatment–outcome association Although patient status during the procedure can modify the administration of the sedative over time, this is a downstream effect of treatment choice at baseline, so is not viewed as an intercurrent event that confounds the treatment–outcome relationship.

2.2 Step 1

Step 1 is the specification of a realistic statistical model, i.e., a set of possible joint distributions of the data. Domain knowledge can be used to restrict the model by ruling out distributions that are incompatible with known truth. The key is to avoid making unrealistic restrictive assumptions that preclude the true data distribution. For example, restricting the conditional distribution of a point treatment to the set of main terms logistic regression models is appropriate when treatment is randomized, but overly restrictive when treatment decisions were made by clinicians and patients, potentially involving complex interactions among baseline patient characteristics. In our running example, the likelihood of the data can be factorized as $\mathcal{L}(O) = p_Y(Y|A, W)\ p_A(A \mid W)\ p_W(W)$.

2.3 Step 2

Step 2 approaches the problem from a causal perspective. A causal model specifies known conditional independencies in the data. Directed acyclic graphs (DAG) provide a visual representation of a causal model [13]. Nodes in the graph represent endogenous variables: covariates, treatment, and outcomes in the causal model. Exogenous variables unaffected by others in the graph are denoted by U. An arrow between two nodes depicts a possible causal relationship. The node at the origin of the arrow is termed the *parent node*. The absence of an arrow encodes knowledge of true statistical independencies. Some independencies are inherent in the time ordering of the data, while others stem from domain knowledge. The DAG in Fig. 2. indicates that *age, sex, pulse, SBP,* and *IBS* potentially impact the outcome. The absence of an arrow between *IBS* and *A* indicates that treatment choice is known to be independent of IBS status. The DAG indicates that the only potential confounders of the treatment–outcome association are *age, sex, pulse, and SBP.*

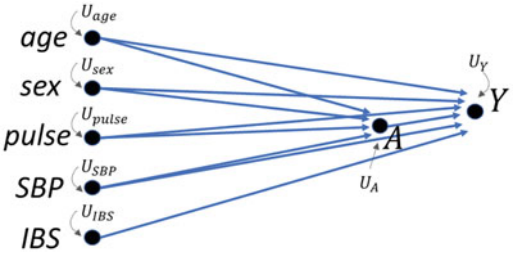

Fig. 2 Directed acyclic graph (DAG) depicting a causal model consistent with the time ordering and with expert knowledge that IBS status does not influence the treatment decision

The relationships depicted in the DAG can also be expressed as a collection of functions in a structural causal model (SCM) [13]. In an SCM, each variable is defined as a function of its parents and exogenous variables, $age = f_{age}(U_{age})$, $sex = f_{sex}(U_{sex})$, $pulse = f_{pulse}(U_{pulse})$, $SBP = f_{SBP}(U_{SBP})$, $IBS = f_{IBS}(U_{IBS})$, $A = f_A(age, sex, pulse, SBP, U_A)$, $Y = f_Y(age, sex, pulse, SBP, IBS, A, U_Y)$.

We next define a causal quantity in terms of the causal model in the full data, where potential counterfactual outcomes arising under each treatment of interest are available. The counterfactual full data consists of observations $O^{Full} = (W, Y^0, Y^1)$,where Y^a denotes a counterfactual outcome observed under exposure to treatment a. The estimand of interest is typically a causal contrast between counterfactual distributions of the outcome. An individual level causal contrast comparing two treatments, a_0 and a_1, is expressed as a function of Y^{a_0} and Y^{a_1}, e.g., $\psi^{causal}_{ATE} = EY^{a_1} - EY^{a_0}$ is a causal additive effect of treatment (ATE), $\psi^{causal}_{RR} = EY^{a_1}/EY^{a_0}$ is a causal relative risk (RR), etc. The causal quantity of interest in our running example is the difference in post-sedation SBP, ψ^{causal}_{ATE}.

2.4 Step 3

Step 3 of the roadmap defines the statistical parameter that can be estimated from data that can be observed in the real world, where it is only possible to capture the outcome a subject experienced under the received treatment. This statistical estimand, ψ^{obs}, must be defined in terms of the features of the distribution of the observable data, $O = (W, A, Y_A)$, rather than features of the underlying full data defined in the causal model.

In our running example, the statistical ATE parameter is given by $\psi^{obs} = E[E(Y|A = 1, W) - E(Y|A = 0, W)]$. Identifying assumptions link this statistical estimand with ψ^{causal}_{ATE} [14]. The first of these is the *consistency assumption* stating that for a subject, i, who experiences a treatment or exposure at level a, the observed outcome, Y_i, is equivalent to the counterfactual outcome, Y^a_i.

The *positivity assumption* states that within strata defined by confounders in W, there must be a positivity probability of receiving treatment at all levels under consideration, $0 < P(A = a \mid W) < 1$, $a = 1$ or 0. In many applications, the outcome may be subject to missingness (missingness indicator $\Delta = 1$

when the outcome is observed, $\Delta = 0$ when the outcome is missing). For these situations, the SCM would contain an additional function describing the missingness mechanism, $\Delta = f_\Delta(W, A, U_\Delta)$. The observed data would be given by $O = (W, A, \Delta, \Delta Y_A)$, where $\Delta Y_A = Y_A$ when $\Delta = 1$ and is missing when $\Delta = 0$. The definition of the causal parameter remains unchanged, while the statistical estimand must explicitly account for missingness, e.g., for the ATE $\psi^{obs} = E[E(Y|\,\Delta = 1, A = 1, W) - E(Y|\,\Delta = 1, A = 0, W)]$. The positivity assumption must hold with respect to missingness as well. Within strata defined by A and W, there must be a positive probability that the outcome will be observed, $0 < P(\Delta = 1 \mid A, W) \leq 1$.

The *randomization assumption* of no unmeasured confounding states that treatment and outcome missingness are independent of the counterfactual outcome given the past, $Y^a \perp A, \Delta \mid W$. This corresponds with assuming that Δ is independent of Y, given (W, A), and A is independent of Y^a given $\Delta = 1, W$.

When these identifying assumptions hold, ψ^{causal} is identifiable from the data. Whether they hold or not, ψ^{obs} is a statistical estimand clearly defined as a function of the true data distribution.

Working through steps 0–3 is sometimes an iterative process. It is important to note that these steps occur without examining study data. When designing a statistical analysis plan for the purpose of regulatory approval, the process occurs prior to data collection. Post-market safety surveillance may involve secondary analysis of existing data that includes some real-world elements, but the actual data doesn't play a role in steps 0–3.

2.5 Step 4

Step 4 focuses on the statistical estimation problem. A critical tenet of TL is that an estimator is a pre-specified mapping from data to a scalar. Traditional parametric modeling approaches typically regress Y on A, optionally adjusting for additional covariates. Coefficients in the model are estimated using maximum likelihood, and the coefficient in front of A is interpreted as the conditional treatment effect, e.g., a log hazard ratio in a Cox model, a log odds ratio in a logistic model, or an additive effect in a linear model. This approach is limited in that if the model is not correct then the effect estimate will be biased (unless treatment is randomized [15]). This model assumes the treatment effect is homogeneous and that there are no effect modifiers, such as drug–drug interactions. It also assumes monotonicity and linearity in the dose-response relationship between treatment and the outcome. If any of these assumptions are unwarranted, effect estimates will be biased. Furthermore, in high dimensional settings, it is impossible to a priori specify a correct parametric model.

If we are interested in learning from RWD about treatment effects in diverse populations, then adopting a more flexible methodology is a better alternative. Desirable properties of an estimator are that it is consistent, regular, asymptotically

linear (RAL) and thereby asymptotically normal, efficient in the sense that the normal limit distribution has minimal variance. Asymptotically linear means that the estimator minus estimand equals the empirical mean of a function of O_i called the influence curve of the estimator, plus an asymptotically negligible remainder. Root-n convergence of such an estimator implies that $\sqrt{n}\left(\psi_0 - \psi_n\right) \xrightarrow{d} N\left(0, \sigma^2\right)$, where n is sample size, ψ_0 is the true parameter value, ψ_n is the estimated value, and σ^2 is the variance. An efficient estimator is one in which σ^2 is the variance of what is known as the efficient influence curve (EIC) [16, 17]. The EIC is a mathematical object that can be computed for any statistical model and bounded pathwise differentiable target parameter. It is derived as the canonical gradient of the derivative of the target parameter viewed as a function of the data density [18]. Theory teaches us that an estimator is asymptotically efficient if and only if it is asymptotically linear, with influence curve the canonical gradient. TMLE possesses each of these properties, and promotes consistency by incorporating non-parametric estimation of the key functionals of the data distribution [1].

TMLE+SL couples an efficient estimator with machine learning to flexibly model outcome regressions, propensity scores, and missingness mechanisms. The combination provides estimates consistent with the process that gave rise to the data. Unlike machine learning alone, TMLE+SL is tailored towards providing efficient unbiased estimation of the target parameter and valid inference. Its practical utility includes the ability to account for baseline and time-varying confounding, intercurrent events, and missing outcomes, in estimating any pathwise-differentiable parameter of interest, in point treatment problems, longitudinal analyses, and analyses of time-to-event data [1, 19].

TMLE is a two-step procedure. In a point treatment problem, the first step uses SL to obtain initial estimates of the outcome regression (Q), propensity scores and missingness mechanisms (collectively denoted by $G = (g_A(A, W), g_\Delta(\Delta, A, W))$, while it estimates the expectation over W with the empirical mean. The second, so-called *targeting* step, involves fluctuating the initial outcome model to improve the bias variance trade-off for ψ^{obs} by ensuring the EIC has empirical mean 0. Statistical theory shows that in this estimation, problem estimators having this property are double robust (DR), i.e., consistent if either Q or G are correctly specified [16, 20, 21]. When both Q and G are correctly specified, these estimators are efficient.

The variance of ψ_n, $\sigma^2_{\psi_n} = \sigma^2/n$, can be used to evaluate p-values and to construct confidence intervals. When positivity is an issue, IC-based confidence intervals might provide less than the nominal coverage [21]. One alternative recognizes that as a result of using TMLE, we have a targeted estimate of the data density, p_n^*.This allows us to bootstrap by sampling from p_n^*, then carrying out the targeting step in the bootstrapped sample and evaluating the parameter estimate to create a finite sampling distribution. This *targeted bootstrap* picks up the behavior of the second order remainder term that is asymptotically negligible, but can be relatively large when positivity is an issue [22]. Quantile-based confidence intervals Wald-type confidence intervals can be constructed based on the bootstrapped estimate of the

variance, or based on the quantiles of the bootstrapped distribution. This approach avoids re-estimation of Q and G, thus, is quite computationally feasible.

A second alternative is to recognize that $\sigma^2_{\psi_n}$ is itself a pathwise-differentiable parameter that can be estimated from data in a robust, targeted manner using TMLE. We can fit the data density with a TMLE, obtaining an estimate p_n^* targeted towards $\sigma^2(p)$, then estimate the variance with $\sigma^2\left(p_n^*\right)$. This plug-in estimate of the variance offers improved inference in sparse data settings.

Several features distinguish TMLE from other DR estimators. As a substitution estimator, TMLE is guaranteed to remain within the bounds of the statistical model (e.g., outcome regression estimates remain within the possible range) [23]. A substitution estimator is an estimator of type $\Psi\left(p_n^*\right)$, with p_n^* an estimator of the true density, p_0, and being an element of the statistical model $\mathcal{M}$. This plug-in property of the TMLE improves finite sample bias and variance relative to non-plug-in estimators, which can even produce negative estimates of a probability in sparse data situations. Several advantages arise from targeting an initial density estimator. TMLE can be utilized for parameters where no estimating equation approach exists or where the estimating equations have multiple or no solutions. It also allows TMLE to incorporate machine learning while remaining RAL [21]. For this reason, we refer to TMLE as the bridge from machine learning to statistical inference. Another finite sample advantage is that estimation of the G components of the likelihood can be tailored based on residual bias in the parameter estimate evaluated with respect to the initial estimate of Q. This approach, known as collaborative TMLE (C-TMLE), can improve the bias/variance trade-off by conditioning on only a subset of confounders, while remaining DR [24, 25]. C-TMLE is particularly useful when there are near-violations of the positivity assumption and in high dimensional settings. TMLE can also naturally incorporate additional targeting for the purpose of additional statistical robustness properties, or simultaneously target many parameters, including an entire survival curve [26].

Consistency and efficiency of the TMLE rest on successfully modeling the Q and G components of the likelihood. It is impossible to know in advance which parametric or machine learning algorithm is optimal. This challenge motivates the use of SL to simultaneously consider multiple approaches, relying on cross-validation to select the best algorithm (discrete SL) or the best combination of algorithms (ensemble SL) from a user-specified collection known as the library [27, 28]. Aside from the library specification, SL performance depends on the complexity of the underlying prediction or regression function, the cross-validation scheme, and choice of loss function [29, 30]. Practical advice and a flowchart for specifying a super learner tailored to the task at hand and characteristics of the data are available in the literature [30].

Given the theoretical properties of SL, it is natural to wonder why we do not simply obtain SL predictions for each observation under each counterfactual value for A of interest and evaluate the plug-in estimator. Consider the ATE and let $\overline{Q}_n^{SL}$ (a, W) be the predicted value from the SL fit of the outcome regression when $A = a$. Why not directly evaluate $\psi_n^{SL} = \frac{1}{n}\sum_{i=1}^{n} \overline{Q}_n^{SL}(1, W) - \overline{Q}_n^{SL}(0, W)$?

Table 1 Simulation study: Bias, variance (var) and mean squared error (MSE) for the unadjusted, IPTW, and TMLE+SL estimators of the ATE

Estimator	Bias	Var	MSE
unadj	1.774	1.024	4.170
IPTW	0.210	0.058	0.102
TMLE+SL	0.053	0.020	0.023

The answer stems from the fact that SL and other machine learning algorithms approximate an unknown prediction function by optimizing a global loss. While this is ideal for the purpose of prediction, it offers a less than ideal bias/variance tradeoff for estimating ψ^{obs}. There are no guarantees on rates of convergence with respect to the target parameter and no guarantees of asymptotic linearity. Thus valid inference is precluded, except when using special sieve maximum likelihood estimators (MLE), such as plug-in highly adaptive lasso-MLE [31].

2.5.1 Simulation Study

A simulation study demonstrates how following the roadmap guidelines impact study estimates. We compare the unadjusted estimate of the ATE with an estimate obtained using inverse probability of treatment weighting (IPTW) [32] and TMLE+SL. Stabilized weights for the IPTW estimator were based on propensity scores estimates from a main terms logistic regression model. This common practice is a slight misspecification of the true PS model. TMLE+SL estimates were obtained using the *tmle* R package, with the default settings [33]. One thousand datasets of size $n = 500$ were generated (Appendix A.1). Bias, variance, mean squared error (MSE) are reported in Table 1. Although IPTW greatly reduced bias and variance compared with the unadjusted estimator, results illustrate that TMLE+SL was 75% less biased than IPTW, with 66% smaller variance. These gains stem from using machine learning to minimize model misspecification bias, and from TMLE's efficiency property.

2.6 Step 5

Step 5 is to assess the interpretation and robustness of the study finding resulting from step 4. *Diaz and van der Laan (2013)* define the *causal gap* as the difference between the statistical estimand (ψ^{obs}) and the causal estimand (ψ^{causal}) [34]. That paper proposes a sensitivity analysis to explore how different values of the hypothesized gap would impact the effect estimate and confidence interval bounds. If a small causal gap would reverse the substantive conclusion, then the study findings are not robust. This might imply that the study does not provide substantial evidence for regulatory or other decision-making. If, on the other hand, even a

large causal gap would not change the substantive conclusion, then the evidence produced by the study could be acted upon with confidence. This non-parametric sensitivity analysis can complement other sensitivity analyses deemed appropriate by regulators.

Any causal gap would be due to violations of the underlying identifying assumptions. The consistency and randomization assumptions are not testable from data. Their plausibility rests on subject matter experts with knowledge of the underlying data generating process and data capture. The impact of practical violations of the positivity assumption could also be evaluated as part of this sensitivity analysis. In addition, diagnostics examining baseline differences between treatment and control groups and assessing their overlap can provide important insight. In studies where parametric methods are used in TMLE rather than an advanced SL, additional sensitivity analyses addressing those restrictive statistical assumptions would also be required. However, this approach is not recommended. A better alternative is to set up outcome blind simulations before specifying the full TMLE and SL to evaluate if the SL is sufficiently data adaptive and make adjustments accordingly. The primary goal of the sensitivity analysis is to address non-testable assumptions.

3 Case Study: Single-Arm Trial with External Controls

This section illustrates how to follow the TL roadmap to foster the development of transparent, interpretable, and reliable RWE.

RWE plays an important role in single-armed trials, where outcomes in the treated arm can be contrasted with outcomes in external comparators. External data sources include historical or concurrent trials with similar inclusion/exclusion criteria or RWD. A key challenge is identifying a comparator group where the observed causal contrast can be attributed to the effect of treatment, rather than other differences in the populations, background therapies, monitoring schedules, etc. [35].

We illustrate TL in this context through an analysis of time-to-event data from a real-world single-arm study combined with a synthetic external control arm. Data were downloaded from Project Data Sphere, a repository of oncology data from biopharmaceutical companies, academic medical centers, and government organizations (www.projectdatasphere.org). Our dataset consists of observations on $n = 371$ subjects in the comparator arm of a Phase III RCT sponsored by Eli Lilly and Company (IMCL CP12-0606/TRIO-012) comparing progression-free survival (PFS) in previously untreated patients with HER2-negative, unresectable, locally recurrent, or metastatic breast cancer [36]. The real-world comparators who received a placebo plus docetaxel are viewed as the treated group in our case study. We simulated data on 1000 subjects in an external comparator arm from a similar population (Appendix A.2) and carried out a retrospective cohort study on the combined dataset ($n = 1371$). The next subsections step through the TL roadmap to generate and evaluate RWE concerning our simulated treatment effect.

3.1 Apply the TL Estimation Roadmap

3.1.1 Step 0

The goal of our study is to understand the impact of treatment with docetaxel + background therapy vs. placebo + background therapy on disease progression. The target population is reflected by the real-world study's inclusion/exclusion criteria.

- Female patients at least 18 years of age with histologically or cytologically confirmed, HER2-negative breast adenocarcinoma
- At study entry, the disease must be metastatic or locally recurrent and inoperable with curative intent
- Patients may not have received chemotherapy or biologic therapy for metastatic or locally recurrent, inoperable breast cancer

The treatment under consideration in our case study ($A = 1$) is docetaxel (75 mg/m^2). The (simulated) comparator ($A = 0$) is a placebo consisting of only the histidine-buffered formulation vehicle. Both treatments were administered as an approximately one-hour intravenous infusion on Day 1 ($\pm$ 3 days) of each 21-day cycle. We will contrast the time-to-disease progression in each study arm. Our primary interest is the intention-to-treat (ITT) effect, which is not affected by non-compliance or discontinuation of treatment. The ICH guidelines refer to this as the *treatment-policy strategy* for dealing with intercurrent events that are considered irrelevant in defining the treatment effect of interest [12]. However, mortality is a competing risk that would preclude observing the time to progression. To address the clinically relevant question regarding progression-free survival (PFS), we define a composite outcome, *disease progression or death*. The summary measure is the cumulative incidence of disease progression or death by $t = 60$ months.

The data consists of n i.i.d observations $O = \left(W, A, \Delta, \widetilde{T}\right)$, where W is a vector of baseline covariates (*age,* body surface area (*bsa*), Eastern Cooperative Oncology Group performance status (*ecog*), left ventricular ejection fraction (*lvef)*, measurable lesion (Y/N) (*lesion*), menopausal status (*meno*), and triple negative status (*tripleNeg)*), A is a binary treatment indicator, Δ is an indicator of the event type (0: censoring, 1: progression or death), $\widetilde{T}$, the last time point at which a subject was monitored, is the minimum of the censoring and outcome event times (C and T, respectively). All real-world participants were followed up for longer than 60 months, thus the only censoring event is administrative censoring at $t = 60$.

Subjects in the comparator group are younger on average than subjects in the treatment group, and are 15% more likely to be pre-menopausal (Table 2). Among the 91% of subjects in the comparator group who experienced an outcome event within 60 months, the crude mean PFS was 37 months. Among the 63% of subjects in the treatment group who experienced an outcome event within 60 months, the crude mean PFS was 28 months.

Table 2 Mean and standard deviation (SD) of baseline characteristics of patients by trial arm, and standardized mean difference (SMD)

Covariate	Mean (SD)		
	Comparator arm	Treatment arm	SMD
age	50.25 (10.80)	54.18 (10.02)	−0.38
bsa	1.86 (0.18)	1.76 (0.19)	0.55
ecog	0.36 (0.02)	0.38 (0.02)	−0.05
lesion	0.17 (0.01)	0.19 (0.02)	−0.04
meno	0.41 (0.02)	0.27 (0.02)	0.30
tripleNeg	0.24 (0.01)	0.22 (0.02)	0.05

3.1.2 Step 1

The statistical model is most naturally expressed in terms of the intensities of the failure and censoring time processes, $N(t) = I\left(\widetilde{T} \leq t, \Delta = 1\right)$,and $A_C(t) = I\left(\widetilde{T} \leq t, \Delta = 0\right)$, with respect to the observed history. For the sake of generality we will assume discrete (T, C) on a discrete time scale that is sufficiently fine so that all formulas have their analogs for continuous (T, C). We can recode the observation $O = \left(W, A, \Delta, \widetilde{T}\right)$ in discrete time as the time-ordered longitudinal data structure $O = (W, A, N(0), A_C(0), \ldots, N(\tau), A_C(\tau), N(\tau + 1))$, where τ is a maximal follow up time so that each $\widetilde{T} \leq \tau + 1$. This can be more succinctly expressed by suppressing the time ordering, $O = \left(W, A, \overline{N}(\tau + 1), \overline{A}_C(\tau)\right)$, where the overbar denotes the entire history.

The likelihood of O can be factorized according to the time ordering: $p(O) = q_W(W) g_A(A|W) \prod_{t=0}^{\tau} g_{A_C(t)}\left(A_C(t)|\overline{N}(t), \overline{A}_C(t-1), A, W\right) \prod_{t=0}^{\tau+1} q_{N(t)}\left(N(t)|\overline{N}(t-1), \overline{A}_C(t-1), A, W\right)$, where each conditional density is conditioning on the variables realized prior to the variable in question.

Furthermore, $g_{Ac(t)}\left(1|\overline{N}(t), \overline{A}_C(t-1), A, W\right) = I\left(A_C(t-1) = 0, N(t) = 0\right) \lambda_C(t \mid A, W)$, and $q_{N(t)}(1 \mid \overline{N}(t-1), \overline{A}_C(t-1), A, W) = I\left(N(t-1) = 0, A_C(t-1) = 0\right) \lambda_T(t \mid A, W)$, where $\lambda_C(t|A, W) = P(C = t \mid C > t-1, N(t) = 0, A, W)$ and $\lambda_T(t|A, W) = P(T = t \mid T > t-1, A_C(t-1) = 0, A, W)$. Under the coarsening at random assumption on C stating that censoring and event times are conditionally independent given T, A, W, the conditional hazard functions reduce to $\lambda_C(t|A, W) = P(C = t \mid C > t-1, A, W)$ and $\lambda_T(t|A, W) = P(T = t \mid T > t-1, A, W)$. In other words, under this assumption these intensities of $N(t)$ and $A_C(t)$ reduce to indicators of being at risk of changing values multiplied by the conditional hazards of C and T, respectively. Thus, the density of O can be parameterized as $p = p_{q_w, \lambda_C, \lambda_T, g}$. A statistical model $\mathcal{M}$ for the density p of O is determined by assumptions on λ_C and g, with q_w and λ_T remaining nonparametric.

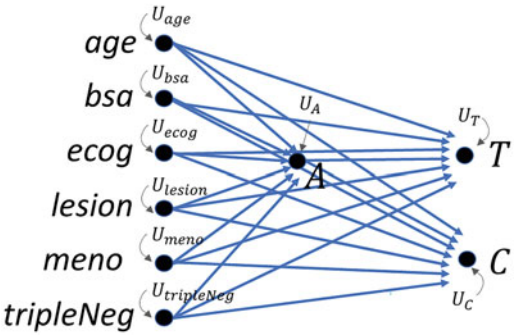

Fig. 3 Directed acyclic graph (DAG) representation of the causal model for the case study

3.1.3 Steps 2 and 3

Our causal model is represented by the DAG in Fig. 3. All baseline covariates are potential confounders of the association between treatment and outcome event time, T.

We are interested in contrasting the 60-month cumulative incidence of disease progression or mortality under exposure to the study treatment vs. the comparator. Subjects in our treatment group are representative of our target population, while subjects in the comparator group are relatively younger and 1.5 times more likely to be pre-menopausal. Thus, we define our causal parameter as the cumulative incidence ratio (CIR) of disease progression or mortality by 60 months among the treated, an average treatment effect among the treated (ATT).

Consider an intervention-specific causal parameter among the treated in terms of the full data, $\psi_a^{causal} = P\left(T^a \geq 60 \mid A = 1\right)$ .Then $P\left(T^a \geq 60 \mid A = 1\right) = E\left(P\left(T^a \geq 60 \mid A = 1, W\right) | A = 1\right) = E_{W|A=1}\left(P\left(T \geq 60 \mid A = 1, W\right)\right) = E_{W|A=1} \prod_{s \leq 60}\left(1 - \lambda_T\left(s \mid A = 1, W\right)\right)$. We denote the latter expression by $\psi_a(P)$, a parameter defined in terms of the observed data distribution, establishing the identification of $\psi_a^{causal}\left(P_X\right)$ under exposure a.

Note that $P\left(T^a \leq 60 \mid A = 1\right) = 1 - \psi_a^{causal}$. Next we define the conditional survival function of T given A, W as $S\left(t|A, W\right) = \prod_{s \leq t} 1 - \lambda_T\left(s|A, W\right)$ so that $\psi_1^{obs} = E_{W|A=1} S\left(60|A = 1, W\right)$ and $\psi_0^{obs} = E_{W|A=1} S\left(60|A = 0, W\right)$. The CIR among the treated is a function of these two statistical estimands, $\psi_{CIR-ATT}^{obs} = \left[1 - \psi_1^{obs}\right] / \left[1 - \psi_0^{obs}\right]$.

3.1.4 Step 4

A customized version of the *survtmle* R package was used to estimate the 60-month CIR of disease progression or mortality among the treated (https://github.com/benkeser/survtmle/tree/att) [37]. SL was used to estimate the propensity score and the failure time process [38]. The number of cross-validation folds was set to 20, and the SL library contained logistic regression, lasso regression, and general

additive models [39–41]. Specifying "ATT = TRUE" when invoking the survtmle function returned estimates of the cumulative event incidence among the treated in each study arm ($\hat{\mu}_0 : comparator, \hat{\mu}_1 : treatment$), and the covariance matrix, $\Sigma = \begin{bmatrix} \hat{\sigma}_0^2 & \hat{\sigma}_{0,1} \\ \hat{\sigma}_{0,1} & \hat{\sigma}_1^2 \end{bmatrix}$. From these results, we evaluate $\psi_n^{CIR-ATT} = \hat{\mu}_1/\hat{\mu}_0$. By the delta method, the variance on the log scale is given by $\sigma_{logCIR}^2 = \sigma_1^2/\mu_1^2 + \sigma_0^2/\mu_0^2 - 2\sigma_{0,1}/(\mu_1\mu_0)$ .We found a CIR among the treated of 0.77 (95% CI: 0.71, 0.83), indicating that treatment reduced the 60-month cumulative incidence.

3.1.5 Step 5

To understand whether the study finding provides sufficiently reliable RWE to support an actionable conclusion or regulatory decision, we consider the direction and magnitude of the causal gap, the difference between the statistical and causal parameters defined earlier. Fig. 4 examines how the point estimate, $\psi_n^{CIR-ATT}$, and 95% CI bounds change as a hypothesized causal gap grows larger, towards and away from the null value of 1. The study's point estimate and 95% CI are at 0 on the x-axis, representing an unbiased estimate of $\psi_{CIR-ATT}^{causal}$ under an assumption that there is no causal gap, $\delta = 0$. Hypothetical gaps are shown on the x-axis in absolute units, δ, and on an alternate x-axis in terms of "adjustment units," the difference between the adjusted and crude estimate (0.77–0.70=0.07).

The plot illustrates that any positive causal gap produces an even larger protective effect of treatment on PFS. A causal gap in the negative direction would have to be at least −0.24, or approximately 0.24/0.07=3.64 times (taking into account

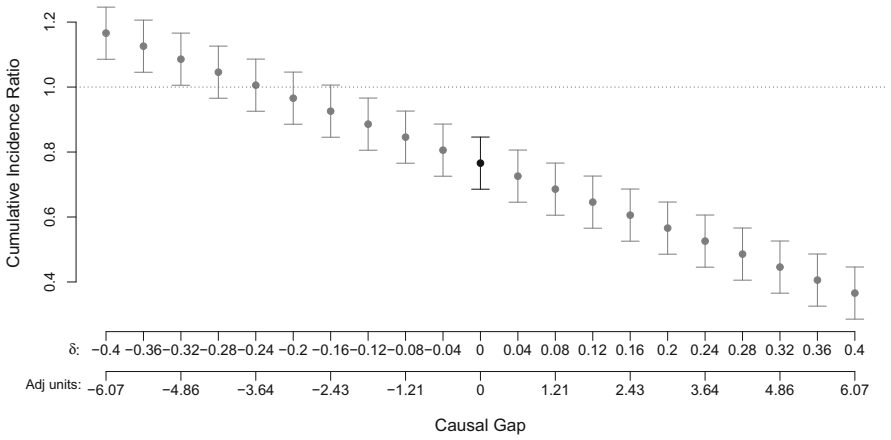

Fig. 4 Sensitivity plot showing point estimates and confidence intervals for the cumulative incidence ratio among the treated under presumed causal gap, δ, between −0.4 and 0.4, or approximately six times the magnitude of the adjustment due to measured confounders (Adj units)

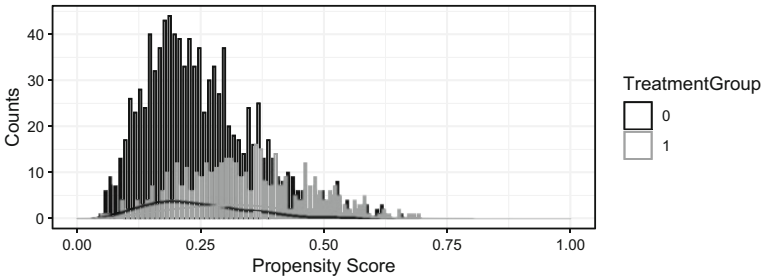

Fig. 5 Distribution of propensity scores by treatment group

rounding error) larger than the adjustment due to measured confounders, for the point estimate to be above 1, and approximately -0.32, or nearly five times larger than the adjustment due to measured covariates for the 95% CI to exclude the null.

Next, we would confer with experts to determine plausible limits on the size and direction of the causal gap, with the understanding that the existence of a causal gap must stem from one or more violations the identifying assumptions. We examine each of these in turn.

In our case study, the consistency assumption is very likely met in the treatment arm due to the careful monitoring of the PFS by the study team throughout follow-up. It is met in the comparator arm through simulation. The positivity assumption only needs to hold with respect to the distribution of data in the treatment group, since we are evaluating an ATT parameter, $0 < P(A = 0 \mid W) \leq 1$. In our data, the estimated propensity score in the treated group is between 0.05 and 0.69, suggesting the positivity assumption is met. A diagnostic showing the propensity score distribution among treated and comparator groups shows good overlap (Fig. 5).

The randomization assumption is generally untestable, though it holds by design in randomized studies with no right censoring. In single-arm trials that utilize external comparator arm data the randomization assumption needs to account for all factors, S, predictive of the outcome that determine trial arm membership, such that $Y^a \perp S \mid A, \Delta, W$. For example, differences over time, region, inclusion criteria, and study conduct need to be listed and evaluated for their potential to induce positivity violations (e.g., all comparators received a now-discontinued background therapy), and to confound the treatment–outcome associations. Differences in unmeasured confounders can also increase the causal gap.

In our case study, all three causal assumptions appear to be met, therefore the plausible causal gap in our study should be close to $\delta = 0$. The point estimates change very little when δ is small, and the CIs largely overlap. Thus, the sensitivity analysis strongly supports interpreting the study finding as a reliable estimate of the true causal effect of treatment, and for concluding that treatment is protective.

4 Conclusion

Essential components of a study incorporating RWD include careful study design, conduct, and complete, accurate data capture. Producing reliable, interpretable RWE also requires learning from data using a rigorous methodology that is transparent and flexible. The TL roadmap provides a systematic approach to generating and evaluating RWE. The potential outcomes framework provides a unified, systematic approach to defining causal estimands regardless of randomization. This is in alignment with FDA's definition of RWE [42]. Clearly articulating the underlying identification assumptions can contribute to evaluating whether the data are suitably fit for purpose. TMLE+SL can appropriately adjust for bias due to baseline and time-dependent confounders, intercurrent events, and missing outcomes. Utilizing data-adaptive machine learning avoids imposing additional statistical assumptions beyond those required for identification. An analysis using TMLE+SL can be completely pre-specified to satisfy regulatory requirements, while remaining data adaptive and providing valid inference [43].

TL is a general approach that can be tailored towards parameters of interest beyond those traditionally evaluated, for example, mediation analysis with time-varying mediators and exposures [44]. The approach is to first define the desired causal quantity in a causal model, then specify the corresponding statistical parameter identified through the G-computation formula [45], derive the efficient influence curve, and finally develop a targeted estimator for the target parameter.

Beyond estimating causal effects, in the pre-clinical phase, TL can be used to rank drug candidates by their potential for Phase I success, or to identify differentially expressed genes meriting further investigation [46, 47]. TL can be used to design sequential adaptive randomized trials to optimize trial design, or to optimize individualized treatment rules for precision medicine [48, 49]. SL-based outcome phenotyping is useful for cohort identification, and for identifying health outcomes of interest in safety and efficacy studies [50]. The TL paradigm can be utilized throughout the pharmaceutical pipeline for optimal learning from data.

Acknowledgement This publication is based on research using information obtained from www.projectdatasphere.org, which is maintained by *Project Data Sphere*. Neither *Project Data Sphere* nor the owner(s) of any information from the web site have contributed to, approved, or are in any way responsible for the contents of this publication. The opinion and information provided in this publication are those of the authors and do not necessarily reflect the official views or policies of the US Food and Drug Administration.

A Appendix

A.1 Simulation Study Data Generation Process

One thousand datasets of size $n = 500$ were generated as follows: age~$U(20, 85)$, sex ~ Bernoulli(0.4), pulse ~ $N(70, 5^2)$, SBP ~ $N(130, 10^2)$, IBS ~ Bernoulli(p_{IBS}) with $p_{IBS} = 0.08 + 0.07sex + 0.1\text{I}(age > 50)$, A ~ Bernoulli(p_A) with

$p_A = expit(-1.8 - 0.01age + 0.3sex + I(pulse < 65) + 0.01SBP)$, $Y \sim N(\mu, 1)$, with $\mu = SBP - 10 + 8A + 0.05age + 8sex + 5A\, I(pulse < 65) + 2A\, I(SBP < 120) - 2IBS$.

A.2 Case Study Data Generation Process

Observations in the synthetic comparator arm were generated as follows. First, $n = 1000$ values for *age* were sampled with replacement from the real-world data with probability inversely proportional to age, then shifted by a random amount, $\varepsilon_{age} \sim N(-2, 9)$. The remaining covariates were generated sequentially by fitting covariate-specific main terms regression models to the real-world data, then adding random noise to predictions from the model based on the previously generated covariates: $bsa = \hat{E}(bsa|age) + \varepsilon_{bsa} \sim N(0.1, 0.034)$; $ecog \sim Bernoulli(p_{ecog})$, where $p_{ecog} = \hat{E}(ecog \mid bsa, age)$; $lesion \sim Bernoulli(p_{lesion})$, where $p_{lesion} = \hat{E}(lesion \mid ecog, bsa, age)$; $meno \sim Bernoulli(p_{preMeno})$, $p_{preMeno} = \hat{E}(preMeno|\, lesion, ecog, bsa, age)$; $tripleNeg \sim Bernoulli(p_{tripleNeg})$, where $p_{tripleNeg} = \hat{E}(pretripleNeg \mid meno, lesion, ecog, bsa, age)$. The outcome event time was generated from an exponential model fit on the real-world data that included *age, bsa, ecog, lesion, meno, tripleNeg* as main terms, an interaction of *tripleNeg* with *bsa*, and an indicator of $age > 75$, plus a random shift in the negative direction, $\varepsilon_{age} \sim U(-10, -6)$, that injected a protective treatment effect into the data. Administrative censoring was imposed at $t = 60$.

Disclaimer This chapter reflects the views of the authors and should not be construed to represent FDA's views or policies.

References

1. van der Laan, M.J., Rose, S.: Targeted Learning: Causal Inference for Observational and Experimental Data, Springer (2011)
2. Gruber, S., Phillips, R.V., Lee, H., Ho, M., Concato, J., van der Laan, M.J.: Targeted learning: Towards a future informed by real-world evidence. *Statistics in Biopharmaceutical Research*, 2023 [in press]
3. Petersen M.L., van der Laan M.J.: Causal Models and Learning from Data: Integrating Causal Modeling and Statistical Estimation. *Epidemiology*, 25, 418 (2014) https://doi.org/10.1097/EDE.0000000000000078
4. FDA Guidance Document (2018): Framework for FDA's Real-world Evidence Program, https://www.fda.gov/media/120060/download
5. FDA Guidance Document (2021): Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory.
6. Levenson, M.: Regulatory-Grade Clinical Trial Design Using Real-World Data. *Clinical Trials*. 17(4), 377–382. doi:https://doi.org/10.1177/1740774520905576 (2020)

7. Corrigan-Curay, J., Sacks, L., and Woodcock, J.: Real-world evidence and real-world data for evaluating drug safety and effectiveness. *Journal of the American Medical Association*, 320, 867–868 (2018)
8. Simon, G.E., Bindman, A.B., Dreyer, N.A., Platt, R., Watanabe, J.H., Horberg, M., Hernandez, A., Califf, R.M.: When Can We Trust Real-World Data To Evaluate New Medical Treatments? *Clinical Pharmacology and Therapeutics,* (1):24–29 (2022) doi: https://doi.org/10.1002/cpt.2252.
9. Concato, J., Stein, P., Dal Pan, G.J., Ball, R, Corrigan-Curay, J.: Randomized, observational, interventional, and real-world—What's in a name? *Pharmacoepidemiology and Drug Safety.* 29,1514– 1517 (2020). https://doi.org/10.1002/pds.5123
10. Gruber, S., Phillips, R. V., Lee, H., Concato, J., & van der Laan, M. (2022). Evaluating and improving real-world evidence with Targeted Learning. *arXiv preprint arXiv:2208.07283*.
11. Ho, M., van der Laan, M., Lee, H., Chen, J., Lee, K., Fang, Y., He, W., Irony, T., Jiang, Q., Lin, X., Meng, Z.: The current landscape in biostatistics of real-world data and evidence: causal inference frameworks for study design and analysis. *Statistics in Biopharmaceutical Research*, pp.1–14 (2021)
12. ICH (2020). ICH E9(R1) Addendum to Statistical Principles for Clinical Trials on Choosing Appropriate Estimands and Defining Sensitivity Analyses in Clinical Trials, https://www.ich.org/page/efficacy-guidelines
13. Pearl, J.: Causality. Cambridge University Press. (2009)
14. Stone, R.: The assumptions on which causal inferences rest. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(2),455–466 (1993)
15. Rosenblum, M., van der Laan, M.J.: Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics*, *65*(3), 937–945 (2009)
16. van der Laan, M.J., Robins, J.M.: Unified methods for censored longitudinal data and causality. Springer (2003)
17. Tsiatis, A.A.: Semiparametric theory and missing data. Springer (2006)
18. van der Vaart, A.: Asymptotic Statistics. Vol. Chapter 25. Cambridge University Press (2000)
19. van der Laan, M.J., Rose, S.: Targeted learning in data science. Springer (2018)
20. Robins, J.M., Rotnitzky, A.: Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90 (429), 122–129 (1995)
21. van der Laan, M.J., Rubin, D.: Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1) (2006)
22. Coyle, J., van der Laan, M.J.: Targeted bootstrap. In Targeted learning in data science (523–539). Springer (2018)
23. Gruber, S., van der Laan, M.J.: A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics,* 6(1) (2010)
24. van der Laan, M.J., Gruber, S.: Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics,* 6(1) (2010)
25. Ju, C., Gruber, S., Lendle, S.D., Chambaz, A., Franklin, J.M., Wyss, R., Schneeweiss, S., van der Laan, M.J.: Scalable collaborative targeted learning for high-dimensional data. *Statistical Methods in Medical Research.* 28(2), 532–54 (2017)
26. van der Laan, M., Wang, Z., van der Laan, L.: Higher order targeted maximum likelihood estimation. arXiv preprint arXiv:2101.06290. (2021)
27. van der Laan, M.J., Polley, E.C., Hubbard, A.E.: Super learner. *Statistical Applications in Genetics and Molecular Biology,* 6(1) (2007)
28. Polley E.C., van der Laan, M.J.: Super learner in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series,* working paper 266 (2010)
29. LeDell, E.: Scalable super learning. In Handbook of Big Data. Chapman and Hall (2016).
30. Phillips, R.V., van der Laan, M.J., Lee, H., Gruber, S.: Practical considerations for specifying a super learner. *International Journal of Epidemiology, 2023 [in press]*
31. van der Laan, M.J., Rose, S.: Why Machine Learning Cannot Ignore Maximum Likelihood Estimation. arXiv preprint arXiv:2110.12112. 2021 Oct 23.

32. Hernán, M.A., Brumback, B., Robins, J.M.: Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 11(5), 561–70 (2000)
33. Gruber, S., van der Laan, M.J.: tmle: An R Package for Targeted Maximum Likelihood Estimation (v. 1.5.0.2). *Journal of Statistical Software*, 51(13), 1–35 (2012)
34. Díaz I., van der Laan, M. J.: Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *The International Journal of Biostatistics*, 9(2), 149–160 (2013)
35. Seeger, J.D., Davis, K.J., Iannacone, M.R., Zhou, W., Dreyer, N., Winterstein, A.G., Santanello, N., Gertz, B., Berlin, J.A.: Methods for external control groups for single arm trials or long-term uncontrolled extensions to randomized clinical trials. *Pharmacoepidemiology and Drug Safety,* 29(11),1382–1392 (2020)
36. Phase III Study of Docetaxel + Ramucirumab or Placebo in Breast Cancer. ClinicalTrials.gov identifier NCT00703326. Updated December 6, 2021. Accessed August 11, 2022. https://clinicaltrials.gov/ct2/show/NCT00703326
37. Benkeser, D.C., Carone, M., Gilbert, P.B.: Improved estimation of the cumulative incidence of rare outcomes. *Statistics in Medicine*, 37(2),280–293, (2017) doi:https://doi.org/10.1002/sim.7337
38. Polley, E., LeDell, E., Kennedy, C., van der Laan, M.: SuperLearner: Super Learner Prediction. R package version 2.0-26 (2019) https://CRAN.R-project.org/package=SuperLearner
39. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (v 3.6.3) (2020) http://www.R-project.org/.
40. Friedman, J., Hastie, T., Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22 (2010)
41. Hastie T.: gam: Generalized Additive Models. R package version 1.16.1. (2019) https://CRAN.R-project.org/package=gam.
42. FDA Guidance Document (2021): Considerations for the Use of Real-World Data and Real-World Evidence To Support Regulatory Decision-Making for Drug and Biological Products, https://www.fda.gov/media/154714/download
43. Gruber, S., Lee, H., Phillips, R., Ho, M., & van der Laan, M.: Developing a Targeted Learning-Based Statistical Analysis Plan. *Statistics in Biopharmaceutical Research*, 1–8 (2022)
44. Zheng, W., van der Laan, M.J. Mediation analysis with time-varying mediators and exposures. Chapter 17 in Targeted Learning in Data Science, Springer (2018)
45. Robins, J.M.: A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease*, 40(2), 139s–161s (1987)
46. Wang, G., Schnitzer, M.E., Menzies, D., Viiklepp, P., Holtz, T.H., Benedetti, A.: Estimating treatment importance in multidrug-resistant tuberculosis using Targeted Learning: An observational individual patient data network meta-analysis. *Biometrics*, 76, 1007– 1016, (2020) https://doi.org/10.1111/biom.13210
47. Wang, L., Sun, X., Jin, C., Fan, Y., Xue, F.: Identification of Tumor Microenvironment-Related Prognostic Biomarkers for Ovarian Serous Cancer 3-Year Mortality Using Targeted Maximum Likelihood Estimation: A TCGA Data Mining Study. *Frontiers of Genetics.* 12:625145 (2021) doi: https://doi.org/10.3389/fgene.2021.625145
48. Chambaz, A., van der Laan, M.J.: TMLE in adaptive group sequential covariate-adjusted RCTs. In Targeted Learning, Springer (2011)
49. van der Laan, M.J., Petersen, M.L.: Statistical learning of origin-specific statically optimal individualized treatment rules. *The International Journal of Biostatistics*, 3(1) (2007)
50. Carrell, D.S., Gruber, S., Floyd, J.S., Bann, M., Cushing-Haugen, K., Johnson, R, Graham, V, Cronkite, D, Hazlehurst, B, Felcher, A.H., Bejin, C.A.: Improving methods of identifying anaphylaxis for medical product safety surveillance using natural language processing and machine learning. *Pharmacoepidemiology and Drug Safety*, 30, 16–17 (2021)

Estimand in Real-World Evidence Study: From Frameworks to Application

Ying Wu, Hongwei Wang, Jie Chen, and Hana Lee

1 Introduction

Estimand is the target of estimation to address the scientific question of interest [1]. Precise definition of estimand helps elucidate what is to be estimated and thus clarifies what question can (or cannot) be answered using observed data. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9(R1) [1] addendum on estimands and sensitivity analysis (henceforth abbreviated as the ICH E9(R1) or the addendum) provides a structured framework for constructing estimands in clinical trials by focusing on five attributes: treatments, population, variable (endpoint), intercurrent events (ICE) along with strategies to handle these events, and population-level summary.

Although the addendum states that the framework is also applicable to single-arm trials and observational studies, constructing estimands for real-world evidence (RWE) studies is not as straightforward and often requires additional considerations. The *Real-World Evidence Scientific Working Group (SWG) of the American Statistical Association Biopharmaceutical Section* published Chen et al.'s [2] paper, in which they discussed the challenges in constructing estimands for RWE studies in great detail. More specifically, Chen et al. [2] elucidated (1) similarities and

Y. Wu
Department of Biostatistics, Southern Medical University, Guangzhou, China

H. Wang
Medical Affairs & Health Technology Assessment Statistics, AbbVie, North Chicago, IL, USA

J. Chen
Taimei Technology & Overland Pharma, Warrington, PA, USA

H. Lee (✉)
Food and Drug Administration, Silver Spring, MD, USA
e-mail: Hana.Lee@fda.hhs.gov

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_9

differences in estimand attributes between traditional clinical trials and RWE studies; (2) presented points-to-consider when defining real-world estimands; and (3) provided a roadmap for constructing real-world estimands. We expand the discussion in Chen et al. [2] and provide additional considerations with respect to the construction of estimands in RWE studies.

We will begin with an overview of existing frameworks that might be useful to define real-world estimands, as in Chen et al. [2] However, we propose to consider an additional framework which was not a part of Chen et al.—a targeted learning framework [3]. We elucidate how each framework can be used to define target estimand as well as to identify sources of bias and underlying assumptions associated with the selected estimand, which are specific to RWE studies. We also elaborate on how the use of potential outcome notation can provide a basis for precise definition of different types of ICE and corresponding strategies handling them, which can ultimately provide a well-defined, transparent definition of a target estimand and thus inform appropriate study design and analysis. These are also illustrated using various case examples.

The rest of this chapter is organized as follows: Section 2 presents an overview of existing estimand-related frameworks. Section 3 delineates how one can define real-world estimands based on each framework using various case examples. Section 4 provides a summary and discusses additional considerations for constructing real-world estimands.

2 Frameworks Relevant to Real-World Estimands

This section reviews four frameworks that can provide guidance on how to define estimands for RWE studies: the ICH E9(R1) [1], target trial framework [6], causal inference framework [4, 5] (a.k.a., *Neyman-Rubin causal inference framework* in some other literature) and targeted learning framework [3]. Of note, all of these frameworks, except the ICH E9(R1), consider more than estimand encompassing study question, estimand, design, analysis, and/or interpretation of findings. In addition, the causal inference framework is the basis for all the other frameworks. We also highlight that all of these frameworks are based on the same notion of causality; however, each framework has its own distinct perspectives. We illustrate how some elements in each of these frameworks can be used as guiding principles for constructing estimands, identify sources of bias and underlying assumptions in RWE studies.

2.1 The Estimand Framework in ICH E9(R1)

The ICH E9(R1) presents a structured framework for constructing estimands in clinical trials by describing five attributes of an estimand. As mentioned before,

application of the ICH E9(R1) framework to RWE studies might not be straightforward. This is mainly because specification of the five estimand attributes depends not only on the research question, but also on real-world data (RWD) sources and complexity in real-world clinical practice. Here, we review the ICH E9(R1) and discuss challenges in applying the framework to real-world settings. See Chen et al. [2] for more detailed discussion regarding similarities and differences with respect to the five estimand attributes between traditional randomized controlled trials (RCT) and RWE studies.

1. *Treatments*. The treatment condition of interest, and, as appropriate, the alternative treatment condition to which comparison will be made. It is important to clearly articulate the treatment regime of interest, which could be individual interventions, combinations of interventions administered concurrently, or a complex sequence of interventions. In RWE studies, various treatment use patterns (e.g., treatment non-adherence, dosage adjustment, treatment switching, concomitant use of multiple medications, or initiation of some dynamic treatment regime that adjusts for treatment strategy based on accumulated patient information) are often observed in routine clinical practice [7–11]. Therefore, articulating the treatment (regime) of interest is one of key considerations in defining estimands for RWE studies. In addition, having clarity on the start of follow-up (i.e., time zero) is crucial in RWE studies. For example, Hernán et al. [6] illustrated how bias may be introduced when cohort entry time, follow up time, and initiation of a treatment are not synchronized. Unlike RCT, where the follow up starts at the time of treatment assignment, subjects in RWE studies may have a span of time during which outcome could not occur before treatment initiation, which will introduce immortal time bias [12]. See also Sect. 2.2 and a single-arm trial example in Sect. 3.1.
2. *Population*. The population of patients targeted by the clinical question, which can be the entire study population and/or a subgroup/stratum of patients defined by particular characteristics such as demographic and clinical characteristics, or ICE (non-)occurrence. The target population for RWE studies, typically defined with a set of less restrictive inclusion and exclusion criteria than those of traditional clinical trials, may include patients with more diverse demographics, clinical characteristics (e.g., multiple comorbidities), geographic areas and study sites, all of which can lead to heterogeneity in target population. Also, types and patterns of ICE occurrence might be much more complicated in RWE studies compared to those in RCT which require additional considerations on selecting appropriate principal stratum.
3. *Variable (endpoint).* The endpoint to be obtained from each patient that is required to address the clinical question. In RWE studies, blinding to the endpoint data, which may already exist in selected RWD, should be enforced during the conduct of RWE studies to avoid investigator/analyst bias. An independent endpoint adjudication committee can be set up to ensure validity and reliability of the endpoint when necessary. Unlike traditional clinical trials, surrogate endpoints are less likely to be used as primary endpoint in RWE studies. Instead, single-time measured clinical endpoints such as death or hospitalization are often used [13].

4. *Intercurrent events and their handling strategies*. ICE are events occurring after treatment initiation that affect either the interpretation or existence of the endpoints associated with the clinical question of interest. The ICH E9(R1) discusses five strategies for handling ICE: *treatment policy*, *hypothetical*, *composite variable*, *principal stratum strategy*, and *while-on-treatment* strategies. See the ICH E9(R1) [1] for details with regards to each strategy. In RCT, most ICE are induced by treatment efficacy or safety profile (e.g., intolerability or lack of efficacy) and terminal events (e.g., death) [2, 14]. ICE in RWE studies are more complicated and likely to be induced by patient behaviors and routine care practice. Chen et al. [2] classifies ICE in RWE studies into five categories: (1) events due to safety concerns; (2) events due to lack of efficacy; (3) events related to behavioral factors (e.g., preference for certain treatment, convenience use of a treatment, doctor–patient relationship, etc.); (4) events related to non-behavioral factors (such as change of medical insurance policy affecting the use of current treatments, improvement of health condition, etc.); and (5) terminal events.
5. *Population-level summary*. Population-level summary of variables/endpoints that serves as a basis for comparison between different treatments (or treatment strategies), such as difference in mean/median survival time, response rate, etc. Unlike RCT, where some standard statistical methods (e.g., regression models) are used to estimate causal treatment effects, specific causal inference methods are often required in RWE studies to ensure comparability of study groups in terms of measured covariates. In addition, ascertainment bias due to baseline window/period [15] and selection bias due to missing information may appear in RWE studies. These biases are often hard to address via analytic methods because reasons for missingness are typically unknown or not well-captured in RWD. In this perspective, defining an estimand in RWE studies can be an iterative process dependent upon RWD quality. Not only that, but all causal methods require some form of untestable assumptions such as no unmeasured confounding [16]. Therefore, it is essential to understand sources of bias and identify underlying assumptions associated with the selected population-level summary (as well as to other attributes such as endpoint). To support interpretation and evaluate robustness of study findings, it is important to consider sensitivity analyses under various, clinically plausible departures from the underlying assumptions. These include, but are not limited to, different mechanisms of missing data, different definitions of analysis set, different causal inference methods, different combinations of covariates in analysis models, and assumptions on unknown or unmeasured confounding variables.

2.2 Target Trial Framework

Target trial framework is a useful tool to identify and prevent some common methodological pitfalls that may introduce biases in observational studies by thinking through an ideal, hypothetical randomized trial called *target trial* and by

attempting to emulate the trial using large observational databases [6]. Therefore, this framework allows to explicitly delineate potential sources of bias in RWE studies and enables to evaluate RWD fit-for-purpose. In this framework, a causal question and corresponding study design are articulated by specifying the following seven attributes referred to as *target trial protocol components*: (1) eligibility criteria, (2) treatment strategies, (3) treatment assignment, (4) start and end of follow-up, (5) outcomes, (6) causal contrasts, and (7) a data-analysis plan. These attributes can be mapped into the ICH E9(R1) estimand attributes and can be used to construct real-world estimands. For example, the eligibility criteria in (1) correspond to the population attribute of the ICH E9(R1), treatment strategies and assignment in (2) and (3) correspond to the treatment and ICE attributes of the ICH E9(R1), etc. Hampson et al. and Umemura et al. [33–35] illustrated the utility of the target trial framework as a tool to define estimands for RWE studies. The target trial framework might be particularly useful to facilitate communications between statisticians and other domain experts because: (a) it is directly connected with a notion of RCT and (b) all estimand and design components are illustrated in non-technical language. This framework has been widely used in various scientific domains including pharmaco-epidemiology [31, 32].

As mentioned earlier, feasibility of the target trial emulation can serve as a basis to evaluate whether (1) a selected estimand is an addressable quantity based on available data and/or (2) the data is fit-for-purpose to address the selected casual question of interest. Therefore, the selection of estimand and evaluation of data fit-for-purpose may require iterative operations in the study development process. In other words, the key scientific question may be determined by the availability in RWD sources, rather than a pre-specified question determining the rest of the study development process. Such iteration should be minimized, and the design should be chosen to match the study objectives and estimands.

Although the framework highlights the importance of articulating the treatment strategy, the lack of specification of ICE and strategies to handle the ICE might be concerning due to complexity and high frequency of ICE in RWE studies, as well as limitations in RWD sources. Combined with the ICH E9(R1) estimand attributes, the target trial protocol components may help articulate the treatment of interest.

2.3 Causal Inference Framework

Causal inference framework that utilizes the potential outcome (or counterfactual) language and corresponding mathematical notation can provide a basis to define estimands in a precise and transparent manner, even under highly complicated scenarios such as multiple time-varying ICE or confounders, informative missingness, etc. An appealing feature of this framework is that it provides a quantitative form of estimands that assists to understand the assumptions needed to estimate them from the available data. Ho et al. [17] and Lipkovich et al. [18] demonstrate how the use of the causal inference framework and potential outcome language/notation can

help define causal estimands for both randomized and non-randomized studies. We illustrate how the potential outcome notation can be applied to define estimands by focusing on different strategies to address ICE [18–21]. We adapt the notation used in Chen et al. [2]

Consider a study in which we are interested in comparing two different treatment strategies. Let $Y(a;t)$ be the potential outcome under treatment strategy a and receipt of treatment status t. Note that a and t may differ, for example, a person prescribed treatment 1 might not take the drug as directed, or switched to another drug, say drug 0. Let $T(a)$ be the receipt of treatment under initiation of treatment a, and $Y(a) = Y(a;T(a))$ be the potential outcome under initiation of treatment a. This notation is useful to define subgroups of interest. For example, $T(1) = 1$ represent patients who initiated treatment 1 and continued to take the treatment 1. Similarly, $T(0) = 0$ represent patients initiated treatment 0 and continued to take the treatment 0. In other words, $T(1) = 1$ and $T(0) = 0$ are so called "compliers." In addition, this notation is useful to set a hypothetical scenario and to re-define potential outcomes under such scenario. For example, we can force $T(a)$ to be at a specific level, say $T(a) = t$ regardless of patients' actual treatment receipt status, and examine a treatment effect where everybody in a population is forced to initiate and stay on treatment t.

Based on this notation, we can now define various estimands of interest:

1. *Treatment policy estimand*. An average treatment effect (ATE) measured in mean difference under the treatment policy strategy (i.e., regardless of ICE) can be defined as $E[Y(1)] - E[Y(0)]$. To paraphrase in words, this estimand corresponds to the difference in mean of potential outcomes in a world in which everyone had initiated the treatment strategy $a = 1$ versus the same person had initiated a reference treatment strategy $a = 0$, regardless of any ICE experience. Similarly, an average treatment effect among the treated group (ATT) under the same population-level summary and the same ICE-handling strategy can be defined as $E[Y(1)|\,T(1)] - E[Y(0)|\,T(1)]$. If one concerns a treatment policy strategy intended to apply to all qualifying patients, the target population should be the whole (indicated) patient population and estimand should be the ATE. If the question concerns a policy of withholding a treatment among those currently receiving (or not receiving it), the estimand should be ATT (or average treatment effect among the untreated, ATU).
2. *Hypothetical estimand*. Now suppose that we are interested in the treatment effect under no ICE occurrence that are plausible in practice. For example, suppose that additional medications other than study treatment 0 or 1 should be available in an RWE study for ethical reasons. However, our interest lies on a treatment effect in the absence of the additional medications (or when they are not available). Define a new set of potential outcomes, $Y\,(0;\,t) = Y\,(1;\,t) = Z(t)$. Then an ATE under the hypothetical strategy can be defined as $E[Z(1)] - E[Z(0)]$, which represents the difference in mean of potential outcomes in a world in which everyone was forced to take treatment $a = 1$ versus a world in which everyone was forced to take treatment $a = 0$. It is important to ensure that a hypothetical scenario of

interest is precisely defined and clinically relevant, as well as that selected RWD are sufficient quality to support corresponding analysis. For example, suppose that a study that considers an additional medication use as an intercurrent event and the hypothetical strategy to handle the event. Corresponding analysis should account for potential non-random selection of the additional medication use which requires sufficient and accurate (covariate) information on reasons for the additional medication use. If some of the information are not collected in RWD, the target hypothetical estimand should not be selected as no reliable estimator exists. More often than not, hypothetical estimands require additional, untestable assumptions than the other estimands and thus anticipate more comprehensive set of sensitivity analyses. See chapter "The Need for Real World Data/Evidence in Clinical Development and Life Cycle Management, and Future Directions" of this book for more details.

3. *Composite variable estimand*. With the composite variable strategy, an estimand incorporates ICE as a part of outcome definition. Of note, there could be many possible outcomes after incorporating ICE and thus there is a need to pre-define a set of clinically relevant outcomes of interest. For example, if a binary outcome such as heart failure (yes/no) is the primary interest, but receipt of a rescue medication, which is an intercurrent event, is considered to define an outcome, there are four different combinations—heart failure with and without receiving the rescue medication, no heart failure with and without receiving the rescue medication. If the occurrence of an intercurrent event is considered a treatment failure, some of these combinations can be merged and the primary outcome may only consider two levels—no heart failure and no rescue medication as success versus others. In this case, an ATE measured in difference in means under the composite variable strategy can be defined as $E[Y^{'(1)}] - E[Y^{'(0)}]$, with $Y^{'}(a) = 1$ indicating (potential) occurrence of heart failure and/or receipt of the rescue medication under treatment a. For continuous outcomes, defining a composite variable estimand is more complex. Approaches, including dichotomization of original continuous scale, assignment of specific values for patients with ICE, using the worst value and modified summary measures such as quality-adjusted survival or trimmed means, can be employed [22–27]. It is also possible to rank patients with ICE according to timing or severity of the ICE, incorporating more granular level information about the ICE [28, 29]. In general, treatment discontinuations due to lack of efficacy or tolerability is regarded as treatment failure, and it is reasonable to assign an unfavorable value (e.g., worst possible score).
4. *Principle stratum estimand*. With the principal stratum strategy, we are interested in a sub-population defined by occurrence of a specific intercurrent event. For example, we may be interested in a treatment effect in a (principle) stratum of patients who can tolerate study treatments (including both the treatment and control). Let S be an indicator of the intercurrent event which corresponds to the treatment tolerability ($S = 1$ if tolerated and 0 otherwise). Then the principal stratum consists of patients that would tolerate under both treatment and control, i.e., $\{S(1) = 1\} \cap \{S(0) = 1\}$. Therefore, an ATE in the principal stratum can be

expressed as $E\{Y(1) \mid S(1) = 1, S(0) = 1\} - E\{Y(0) \mid S(1) = 1, S(0) = 1\}$. Note that the principal stratum estimand is a local average treatment effect. To draw inference about this estimand, the principal stratum of patients' needs to be identified and thus some additional (identification) assumptions are often required [30]. For more detailed examples of this estimand, see Bornkamp et al. [30]

5. *While-on estimand*. With this strategy, we are interested in a treatment effect under treatment adherence. Therefore, the ATE can be defined as $E[Y(1)|T(1) = 1] - E[Y(0)|T(0) = 0]$, i.e., the effect among the compliers. If per-protocol is considered, we are interested in the treatment effect among those who complied to the study protocol.

Although we do not present here, the benefit of considering potential outcome notation is even greater for longitudinal studies with potential time-varying confounding and informative censoring. See Gruber et al. [51] for more details on how potential outcome notation enables to define complex RWE study estimands for observational longitudinal studies in a precise and transparent way. As the definition of estimand further informs the choice of analytic methods, estimation, and inference, the clarity in estimand is also important from a design as well as from an analytic perspective.

As Chen et al. [2] pointed out, a limitation of this framework is that stakeholders other than statisticians might not be familiar with the notation and may need extensive training to understand the concept. Therefore, this framework might be useful to facilitate communication on estimands particularly among statisticians (e.g., when used in statistical analysis plan), but not involving the other stakeholders.

2.4 Targeted Learning Framework

Targeted learning (TL) by van der Laan and colleagues [3] is a statistical framework that provides a systematic roadmap on defining, generating, and evaluating evidence from data, while utilizing an efficient estimation approach [36–38]. Also see chapter "Causal Inference with Targeted Learning for Producing and Evaluating Real-World Evidence" of this book for more details. The TL roadmap [39] consists of a prerequisite step 0 and 5 subsequent steps:

Step 0. Formulate a well-defined question. Describe the study objective reflecting underlying data generating mechanism. The key estimand attributes in the ICH E9(R1), including ICE, are naturally integrated in this step.

Step 1. Define a realistic statistical model for the data. Here, "realistic" implies *NOT* imposing any unknown/unnecessary modeling assumptions such as parametric modeling assumptions. It also includes exploiting knowledge to reduce the size of the statistical model, e.g., known bounds on the outcome, knowledge of the treatment assignment mechanism.

Step 2. Define a causal model and causal estimand (i.e., target causal parameter) in terms of potential outcomes. The causal estimand should be consistent with the pre-specified ICE and ICE strategies considered in the pre-requisite Step 0.

Step 3. Specify a statistical parameter, i.e., a parameter in terms of observed data that is aligned with or best approximates the target causal parameter. Note that there are two different parameters—causal parameter (the ultimate target parameter, but defined based on potential outcomes) and statistical parameter (now defined based on observed data). In this step, one needs to identify and specify assumptions needed to link the causal parameter to the statistical parameter.

Step 4. Conduct statistical estimation and draw causal inference. The TL framework utilizes targeted maximum likelihood estimation (or targeted minimum loss-based estimation; TMLE) coupled with super learning (which is an ensemble of various machine learners) as an efficient estimation tool [3]. Other types of estimators can be used to estimate the same causal estimand. See chapter "Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods" on recent statistical development for comparative effectiveness research beyond propensity-score methods.

Step 5. Sensitivity analyses to assess findings under different hypothetical magnitudes of the causal gap and interpretation of results.

Steps 0–3 in the TL framework are relevant to construct estimands as well as to evaluate RWD fit-for-purpose. Strengths of this framework include, but are not limited to, the fact that it (1) provides a unified, systematic way to construct estimands based on the potential outcome language/notation; (2) provides a guidance on designing a study consistent with the selected estimands and thus helps evaluate data-fit-for-use; (3) enforces to think through and clearly state underlying assumptions associated with the selected estimand which helps planning on sensitivity analyses; and (4) avoids unnecessary probabilistic and modeling assumptions (e.g., linearity, normality, etc.), which enables to use an efficient estimation approach utilizing the state-of-the-art machine learning techniques. A potential limitation of this framework is that it does not explicitly state estimand or design attributes as in the ICH E9(R1) or the target trial frameworks. Therefore, the TL framework might be best utilized when used together with the ICH E9(R1) and/or the target trial frameworks, for the purpose of defining estimands.

3 Examples of Estimands in Real-World Evidence Studies

This section presents three RWE study examples and delineates how estimands can be defined using the four frameworks. Within each example, estimands are described based on two frameworks. We use the ICH E9(R1) for all three examples,

then consider one additional framework to illustrate potential utility of the other framework to further enhance clarity on estimand definition. See also chapter "Examples of Applying Causal-Inference Roadmap to Real-World Studies" of this book for more illustrative examples using the TL framework.

3.1 Single-Arm Trial with External Control

For some disease areas where RCT are infeasible (e.g., a disease with extremely low incidence rate) or unethical (e.g., a life-threatening disease for which no efficacious treatments are available), single-arm trials might be considered to demonstrate efficacy and safety of a medical product. Single-arm trials often use RWD to construct external controls (historical or concurrent). See the FDA draft Guidance on Rare Diseases: Common Issues in Drug Development [40] and the ICH E10 on the choice of control group in clinical trials [41] for situations where external controls can be used.

Gökbuget et al. [42] provide an example of a single-arm trial using RWD to form an external control. They compared outcomes from a phase 2 single-arm study [43] of safety and activity of blinatumomab among 189 adult patients with B-precursor Ph-negative relapsed or refractory acute lymphoblastic leukemia (R/R ALL). For the external control, the authors used a historical data from European national study groups as well as large historical sites data from Europe and the United States. Chen et al. [2] provided the five, ICH E9(R1) estimand attributes of the primary estimand for this study in great detail. Here we revisit this example using both the ICH E9(R1) and the target trial frameworks to provide specification of the target estimand (Table 1). Note that Gökbuget et al. [42] did not consider the target trial emulation and therefore there is no benchmark information. Here, we present *what we consider to be a target trial* for the Gökbuget et al. [42] study, and demonstrate an estimand and potential sources of bias on the basis of the assumed target trial.

The primary study objective for the Gökbuget et al. [42] study might be expressed as:

To evaluate the effect of blinatumomab among adult patients with B-precursor Ph-negative relapsed/refractory acute lymphoblastic leukemia. Now this can be much elaborated by using the ICH E9(R1). Here, we excerpted the primary estimand attributes from Chen et al. [2]:

- Population: adult patients ($\geq$18 years) with B-precursor Ph-negative relapsed/refractory acute lymphoblastic leukemia (R/R ALL)
- Treatment: blinatumomab (9 ug/day for the first 7 days and 28 ug/day thereafter) by continuous intravenous infusion over 4 weeks every 6 weeks (up to five cycles) (experimental arm), or salvage therapy (possibly multiple lines) (historical control arm)

Table 1 Specification of a target trial protocol and the target trial emulation using a single-arm trial with external control in Gökbuget et al. [42]

Protocol component	Target trial	Emulation in Gökbuget et al. [42]
Eligibility criteria	Same as the population attribute in the ICH E9(R1)	The single arm: Same as the target trial. The external control: The historical data did not capture all eligibility criteria applied in the original single-arm trial due to the limited availability in some variable information
Treatment strategies	Same as the treatment attribute in the ICH E9(R1)	The single arm: Same as the target trial The external control: Among patients in historical data with information on several lines of salvage therapy, only the endpoints for the last available salvage therapy were selected. This was to mimic the likely period when a patient would enter the single-arm trial, as the time period of the historical data was from 1990 to 2013, and the patients in the single-arm trial were enrolled over the period 2010–2014
Treatment assign-ment	Randomly assign eligible patients to each treatment strategy—blinatumomab or salvage therapy	Randomization was emulated through weighting outcomes using propensity score-based inverse probability of treatment methods to balance predetermined prognostic baseline factors between patients in the single-arm trial and patients in the historical data set
Outcomes	Same as the endpoint attribute in the ICH E9(R1)	CR was defined differently between the single-arm trial and historical data
Follow-up	Patients were followed from the random treatment assignment until the CS occurrence, or until maximum 24 months after the randomization	The single arm: Same as the target trial. The external control: Start of last salvage therapy in the historical data. Patients in the historical data were not subject to a maximum length of follow-up and could be followed until death
ICE and strategies	Same as the ICE and ICE strategy attribute in the ICH E9(R1)	The single arm: Same as the target trial The external control: Patients with missing CR information in historical data were excluded. We are unable to quantify how accurately the ICE information was identified and ascertained from the RWD sources
Statistical analysis	Direct between-group comparison of CR rates measured in OR scale	Same as the population-level summary attribute in the ICH E9(R1)

- Endpoint: complete remission (CR) within the first two treatment cycles in all blinatumomab-treated patients (experimental arm) or after salvage therapy (historical control arm)
- ICE: death before the first response assessment or adverse events leading to treatment discontinuation before the first response assessment. Treatment policy strategy was considered for primary objective.

- Population-level summary: comparison of rates of CR between the two groups measured in odds ratio (OR) scale, after inverse probability of treatment weighting using propensity score

Now Table 1 provides how considering target trial components can help identify feasibility of considering the ICH E9(R1)-based estimand, or the estimand attributes, for the Gökbuget et al. [42] study. Deviations from target trial components may inform potential sources of bias and limitation of the historical data.

As illustrated in this example, the use of target trial components can further increase the clarity of estimands (or estimand attributes) and help identify potential sources of bias.

3.2 Longitudinal Study with a Static Treatment Regime

A multinational RWE study called CVD REAL aimed to examine whether the benefits of sodium-glucose cotransporter-2 inhibitor (SGLT-2i) empagliflozin in lowering hospitalization for heart failure (HHF) rate among patients with type 2 diabetes mellitus (T2DM), that were observed from a previous randomized trial [44], can be also seen in real-world practice. RWD sources include data collected from health insurance claims, electronic health records of primary care and hospitals, and national registries from six countries. The ICH E9(R1) attributes of the primary estimand for the CVD REAL can be summarized as follows:

- Population: Adult T2DM patients who initiated either SGLT-2i or other glucose-lowering drugs (oGLD), who had at least 1 year data history in the databases
- Treatment: SGLT-2i or oGLD. Note that this is a static treatment regime which does not vary over time based on patients' response to a (sequence of) previous treatment uptake
- Endpoint: HHF, death, and combination of both
- ICE: discontinuation of initiated treatment, change in background glucose-lowering medication, loss to follow-up. While-on-strategy was considered.
- Population-level summary: hazard ratio for time to first endpoint event, estimated from a Cox proportional hazard model after propensity score matching.

Some of these estimand attributes are linked to certain assumptions which are not apparent when verbalized. Focusing on HHF as a sole endpoint outcome for simplicity, now we see how the estimand can be expressed in terms of potential outcome notation and requires to specify some inherent assumptions. In this case, the use of propensity score matching implies that we are interested in the causal hazard ratio, assuming the matching could fully address systematic differences between the two treatment groups. Also, the specification of ICE, particularly the loss to follow-up, assumes that the time of HHF is subject to right censoring. Let $T_{\overline{a}}$ be the potential time of HHF under a specific treatment history $\overline{a}$ which could be different from the actual, observed treatment history denoted by $\overline{A}(t) =$

$\{A(u); 0 \leq u < t\}$, where $A(t)$ represents the actual, observed treatment status at t. At this moment, assume there is no censoring for the sake of illustration. Let $\lambda_{\overline{a}}(t)$ be the potential hazard of HHF at time t when all patients in this study followed a treatment history $\overline{a}$ through time t. Then the causal estimand, which is measured in hazard ratio scale, can be expressed as a causal parameter $\exp(\beta_{\text{causal}})$ in the following marginal structural Cox model [46]:

$$\lambda_{\overline{a}}(t) = \lambda_0(t) \exp\{\beta_{\text{causal}} * a(t)\},$$

where $\lambda_0(t)$ is an unspecified baseline hazard. After propensity score matching, we assume that the parameter β_{causal} is equivalent to β in the following Cox model $\lambda_{\text{T}}\left(t|\overline{A}(t)\right) = \lambda_0(t) \exp\{\beta * A(t)\}$, where T represents a patient's actual, observed time of HHF. Therefore, the use of potential outcome notation makes us differentiate what the target causal parameter is and what we estimate using observed data. From the target causal parameter β_{causal} to the statistical parameter β, we make various assumptions regarding data and model, such as "*the matching can fully address systematic differences between the two treatment groups.*" This requires three causal assumptions associated with the use of propensity score specified in the causal inference framework—consistency, no unmeasured confounding, and positivity assumptions. Consistency means that we assume a patient's response under each study treatment regime is well defined (although generally not observable) and $T_{\overline{A}} = T$ for a patient whose actual treatment history $\overline{A}$ equals to $\overline{a}$. No unmeasured confounding means that information used in propensity score estimation is sufficient to explain the treatment selection mechanism. Positivity assumption means that probability of receiving either treatment is strictly greater than zero over all combinations of different levels of covariates. In other words, all patients should have some probability of receiving both treatments. In addition to these so called "causal assumptions," we also rely on other modeling assumptions too, such as no model misspecifications (for both propensity score and the Cox), proportional hazard, and non-informative censoring. Note that all of these assumptions, except for the proportional hazard and non-informative censoring, are generally not a concern for traditional clinical trials. Therefore, the use of RWD requires additional assumptions and considerations within, which may not be apparent in estimand-defining stage. As the estimand consequently defines design and analysis attributes, having clarity on underlying assumptions with regards to each estimand attribute and its impact on design and methodologic choice is strongly recommended in RWE studies. In the next example, we illustrate how the TL framework provides a systematic roadmap to delineate all of these assumptions while utilizing the potential outcome notation.

Of note, the most widely used ICE handling strategies is treatment policy strategy or while-on strategy. These may be suboptimal for quantifying effectiveness of medical interventions, particularly for chronic disease when hazard ratio is a population-level summary measure, because validity and interpretation of the summary measure estimate heavily depend on assumptions about censoring mechanism. For example, there could be a systematic difference between those who stay on

initial treatment versus those who discontinue or switch. Therefore, methodologies that can account for time-varying nature of treatment assignment/receipt [3, 17, 45–47], as well as potential informative loss/drop-out [48, 49]. might be more appropriate for RWE studies. See Sect. 3.3.

3.3 Longitudinal Study with a Dynamic Treatment Regime

A dynamic treatment regime, also known as an adaptive treatment strategy, is a sequence of treatment decisions that are determined based on patients' response to the treatment. Hernán et al. [50] presented an example of a prospective study of human immunodeficiency virus (HIV)-infected patients using observational data to compare the acquired immunodeficiency syndrome (AIDS)-free survival under the following two dynamic, highly active antiretroviral therapy (HAART) regimes:

- Regime 1: Start HAART when CD4 cell count first drops under 500 cells/μL then always treat.
- Regime 2: Start HAART when CD4 cell count first drops under 200 cells/μL then always treat.

They considered a cohort comprising 2344 HIV-infected individuals included in the French Hospital Database on HIV who had their first CD4 cell count measurement below 500 cells/μL during the study period and had never received antiretroviral therapy before the first measurement. Individuals in the cohort were followed from the first CD4 measurement until a diagnosis of AIDS, death, or end of study period, whichever occurred earlier. Although the primary estimand was not explicitly stated in the original study, the five attributes for the primary estimand might be summarized as follows.

- Population: HIV infected patients with CD4 cell count never below 500 cells/μL and had never received antiretroviral therapy before or at study entry.
- Treatment: Regime 1 and 2 shown above.
- Endpoint: Diagnosis of AIDS, death, or the end of follow-up, whichever comes first.
- ICE: Deviation from one of the two study treatment regimes, e.g., did not start HAART use within 1 month of the first CD4 cell count measurement below 500 cells/μL but started HAART before the CD4 cell count dropped below 200 cells/μL. While-on treatment strategy was considered to censor those who deviated from the two study regimes of interest.
- Population-level summary: Hazard ratio to compare endpoint rates (including mortality rate) estimated from a Cox proportional hazard model after accounting for time-varying nature of treatment and (potential) informative censoring via propensity score weighting.

We now describe how to elaborate this estimand, express it based on potential outcome notation, and identify underlying assumptions following the TL frame-

work. Throughout, we assume that the above estimand attributes are final and focus on the death endpoint for the sake of illustration. As mentioned earlier in Sect. 3.2, the selection of the Cox proportional hazard model as an analytic approach automatically imposes a parametric modeling assumption. Later, in Step 1 of the TL roadmap, we demonstrate how this assumption limits a set of possible data distributions and collection of statistical models. We continue to use the notation introduced in Sect. 3.2.

Step 0: A well-defined question in Step 0 should be able to address the five estimand attributes in ICH E9(R1), which are stated above [51]. In addition, this step of the TL roadmap emphasizes a precise description of the experiment generating the data. This requires specification of data structure including treatment, covariate, and ICE sequence for the longitudinal study. We did not find information on measurement times and frequency from Hernán et al. [50]. However, in general, the data structure considering time-to-event outcome can be expressed as $O = (L, A, \tilde{T}=\min(T,C), \Delta = I(\tilde{T} = \mathrm{T}))$, where L is a vector of time-varying covariates and C is the time of treatment deviation (i.e., the time of ICE occurrence; henceforth censoring time). Assuming the time scale is discretized and 3 time points, data structure may be depicted in Fig. 1:

Step 1: If we were followed the TL roadmap, we would have defined a realistic statistical model, say M, respecting the time ordering of the data generating process O to be consistent with study inclusion/exclusion criteria and would have not imposed any unknown assumptions. Therefore, unlike Hernán et al. [50], we might have not imposed a parametric modeling assumption on the mortality rate (i.e., the Cox model), as well as on the conditional probability of being censored (i.e., a parametric propensity score modeling for the censoring). It is also worth mentioning that the hazard ratio measure is not a quantity that admits a causal interpretation, even in some RCT settings (Aalen et al. [52], Hernán [53]). Alternatively, difference in mean survival time [54] or restricted mean survival time analysis might be considered. Acknowledging these limitations, we will continue to describe how to follow the TL roadmap assuming the setting where Hernán et al. [50] is valid.

Step 2: The causal model under the treatment-confounder feedback as well as informative censoring is represented by the following DAG in Fig. 2:

The causal hazard ratio parameter associated with the while-on treatment strategy can be expressed as the exp (β_{causal}) in the same marginal structural model shown in Sect. 3.2, but now with the treatment being a dynamic regime.

Fig. 1 Longitudinal process of giving rise to the data over time. L and A are collected until the time of event or censoring $\tilde{T}=\min(T, C)$

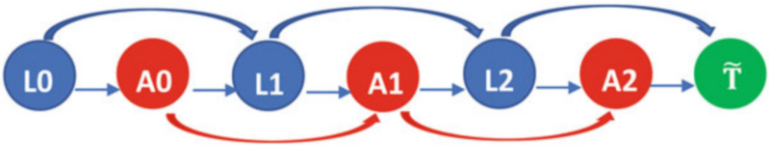

Fig. 2 A causal diagram for three time points with treatment-confounder feedback

Step 3: A statistical parameter in terms of observed data under the assumptions imposed in Hernán et al. [50] and given the above estimand attributes, consists of the following attributes:

1. Is based on inverse probability of informative censoring weight accounting for censoring patients when they stop following one of the study treatment regimes.
2. Compares the survival of the uncensored individuals under each study regime in a weighted analysis adjusting for the treatment-confounder feedback via inverse probability of (time-varying) treatment weights.

Hernán et al. [50] state that g-estimation of nested structural models could be an alternative approach. Regardless, the following assumptions are imposed to link the statistical parameter with the causal parameter of the effect of the dynamic treatment regimes: consistency, no unmeasured confounding, positivity assumptions, no model misspecifications, and no unmeasured reasons for censoring. In particular, the assumptions on no unmeasured confounding and no unmeasured reasons for censoring dictate that the RWD contains sufficient information on all joint risk factors for treatment initiation/discontinuation and mortality (i.e., data is fit-for-purpose).

This example demonstrates that the TL framework is more specific to identify limitations associated with selected estimand attributes or modeling approach than the causal inference framework. Considerations on underlying data generating mechanism, causal model, causal gaps provide a guidance on data fit-for-use evaluation and sensitivity analysis, which could ultimately inform the interpretation of study findings and support decision-making.

4 Summary and Discussion

Table 2 provides a summary of the four frameworks. Broadly speaking, both the ICH E9(R1) and target trial frameworks use non-technical language while the causal inference and TL frameworks utilize potential outcome notation. Therefore, the ICH E9(R1) and target trial frameworks could facilitate communication between various disciplines involved in the formulation of RWE study objectives, while the causal inference and TL frameworks could provide additional clarity on estimand and corresponding choice of statistical methods among statisticians and quantitative

Table 2 A summary table of four frameworks relevant to real-world estimands

	ICH E9(R1)	Target trial framework	Causal inference framework	Targeted learning framework
Overview	Providing a structured framework for constructing estimands by describing five attributes of an estimand	Considering an ideal, hypothetical RCT (target trial) and thinking through how to emulate the RCT using RWD using seven protocol components	Basis for all of the other frameworks encompassing study question to sensitivity analysis and beyond. Potential outcome notation can be used to define estimands in a quantitative manner	A statistical framework that provides a roadmap on defining, generating, and evaluating evidence from data utilizing an efficient estimation approach
Relevance to real-world estimands	Explicitly stating five estimand attributes. Best fit for trial settings. Principles remain the same, but application to RWE studies might not be straightforward	Some or all of the seven protocol components can be used to construct estimands. Although it highlights articulating a treatment strategy, the lack of ICE specification might be concerning	Explicitly defining estimand using mathematical (potential outcome) notation, enhancing clarity on estimands and methodological pitfalls	Some steps in the roadmap are specific to constructing estimands and identification of underlying assumptions
On communication on estimands	Providing a foundation for dialogue between various disciplines involved in the formulation of study objectives	Same as the framework in ICH E9(R1). Using a concept of RCT as the basis for emulation is helpful	Useful to facilitate communication among statisticians and quantitative scientists who understand the notation	Same as the potential outcome notation. Helps to understand difference between causal estimand and statistical estimand, which is less concerning in traditional RCT settings
On study protocol	Help enhance alignment from study planning, design, conduct, analysis, and interpretation	All design components are explicitly illustrated	Help enhance alignment between estimand and method of estimation	Help enhance alignment between estimand, data, and method of estimation
On validity of RWE	Help articulate the strategies for handling ICE, though provide no detailed guidance on analysis method	Help explicitly investigate and identify various sources of bias that might be hidden in observational studies, though provide no detailed guidance on analysis method	Precisely define estimand, though provide no detailed guidance on analysis method.	Help identify and state underlying assumptions. Highlights importance of avoiding unnecessary probabilistic and modeling assumptions, utilizes an efficient estimate-on approach

scientists. Compared to the ICH E9(R1), the target trial framework considers more direct and comprehensive components on RWE study design which allows to explicitly delineate potential sources of bias and enables to evaluate RWD fit-for-purpose. The causal inference framework provides a basis to formally define estimands. Rooted in the causal inference framework, the TL framework further provides a systematic roadmap, from the start to the end of an RWE study, that is more specific in terms of identifying assumptions associated with selected estimand attributes and embedded modeling approach.

Constructing RWE study estimands is complex. It involves increasing level of heterogeneity and complexity in defining attributes of an estimand which could, in part, be driven by RWD fit-for-purpose, patient behaviors, and routine clinical practice. In addition, it is important to understand stakeholders and their research questions for the construction of estimands for RWE studies [2].

Lastly, it is crucial to understand the inherent assumptions connecting the target causal estimand with the corresponding statistical estimand. Unlike traditional clinical trials in which the randomization (and some other factors such as rigorous patient follow-up) could approximately warrant the validity of those assumptions, some of them are not even empirically testable in RWE studies. As different strategies for handling ICE and the choice of estimators require different sets of assumptions, interpretability of study findings will heavily rely on the validity of the underlying assumptions. Rigorous sensitivity analysis should always be accompanied to ensure robustness of study findings.

Disclaimer This chapter reflects the views of the authors and should not be construed to represent FDA's views or policies.

References

1. ICH E9(R1) (2021): Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials, https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf
2. Chen, J., Scharfstein, D., Wang, H., Yu, B., Song, Y., He, W., Scott, J., Lin, X., Lee, H.: Estimands in Real-World Evidence Studies. (Submitted 2022).
3. van der Laan, M., Rose, S.: Targeted Learning: Causal Inference for Observational and Experimental Data. New York: Springer. (2011).
4. Neyman, J.: On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9. *Translated in Statistical Science*. **5**, 465–480 (1923).
5. Rubin, D.: Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. **56**, 688–701 (1974).
6. Hernán, M., Robins, J.: Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology.***183**(**8**), 758–764 (2016).
7. Hernán, M., Scharfstein, D.: Cautions as regulators move to end exclusive reliance on intention to treat. *Annals of Internal Medicine.***168**(**7**), 515–516 (2018).
8. Scharfstein, D.: A constructive critique of the draft ICH E9 Addendum. *Clinical Trials.***16**(**4**), 375–380 (2019).

9. Li, Z., Chen, J., Laber, E., Liu, F., Baumgartner, R. Optimal treatment regimes: An empirical comparison of methods and applications. (Submitted 2021).
10. Ogundipe, O., Mazidi, M., Chin, K., Gor, D., McGovern, A., Sahle, B., Jermendy, G., Korhonen, M., Appiah, B., Ademi, Z.: Real-world adherence, persistence, and in-class switching during use of dipeptidyl peptidase-4 inhibitors: a systematic review and meta-analysis involving 594,138 patients with type 2 diabetes. *Acta Diabetologica.***58**(**1**), 39–46 (2021).
11. Nicholas, J., Edwards, N., Edwards, R., Dellarole, A., Grosso, M., Phillips, A.: Real-world adherence to, and persistence with, once- and twice-daily oral disease-modifying drugs in patients with multiple sclerosis: a systematic review and meta-analysis. *BMC neurology.***20**(**1**), 1–15 (2020).
12. Suissa, S.: Immortal time bias in pharmaco-epidemiology. *American Journal of Epidemiology*. **167**, 492–499 (2008).
13. Mercon, K., Mahendraratnam, B., Eckert, J., Silcox, C., Romine, M., Lallinger, K., Kroetsch, A., Fazili, H., Wosióska, M., McClellan, M.: A Roadmap for Developing Study Endpoints in Real-World Settings (2020). Center for Health Policy at Duke University. https://healthpolicy.duke.edu/sites/default/files/2020-08/Real-World%20Endpoints.pdf
14. Qu, Y., Shurzinske, L., Sethuraman, S.: Defining estimands using a mix of strategies to handle intercurrent events in clinical trials. *Pharmaceutical Statistics.***20**(**2**), 314–323 (2021).
15. ENCePP (2022): Guide on methodological standards in pharmacoepidemiology (Revision 10). https://www.encepp.eu/standards_and_guidances/documents/01.ENCePPMethodsGuideRev.10_Final.pdf
16. Rosenbaum, P., Rubin, D.: The central role of the propensity score in observational studies for causal effects. *Biometrika.***70**(**1**), 41–55 (1983).
17. Ho, M., van der Laan, M., Lee, H., Chen, J., Lee, K., Fang, Y., He, W., Irony, T., Jiang, Q., Lin, X.: The Current Landscape in Biostatistics of Real-World Data and Evidence: Causal Inference Frameworks for Study Design and Analysis. *Statistics in Biopharmaceutical Research*, 1–14 (2021).
18. Lipkovich, I., Ratitch, B., Mallinckrodt, C.: Causal inference and estimands in clinical trials. *Statistics in Biopharmaceutical Research.***12**(**1**), 54–67 (2020).
19. Boeden, J., Bornkamp, B., Glimm, E., Bretz, F.: Connecting Instrumental Variable Methods for Causal Inference to the Estimand Framework. *Statistics in Medicine*. **40**(**25**), 5605–5627 (2021).
20. Ocampo, A., Bather, J.: Single-World Intervention Graphs for Defining, Identifying, and Communicating Estimands in Clinical Trials. *arXiv preprint arXiv:2206.01249v1* (2022).
21. Qu, Y., Luo, J., Ruberg, S.: Implementation of Tripartite Estimands Using Adherence Causal Estimators Under the Causal Inference Framework. *Pharmaceutical Statistics.***20**(**1**), 55–67 (2021).
22. Billingham, L., Abrams, K.: Simultaneous Analysis of Quality of Life and Survival Data. *Statistical Methods in Medical Research*. **11**(**1**), 25–48 (2002).
23. Fay, M., Brittain, E., Shih, J., Follmann, D., Gabriel, E.: Causal estimands and confidence intervals associated with Wilcoxon-Mann-Whitney tests in randomized experiments. *Statistics in Medicine*. **37**(**20**), 2923–2937 (2018).
24. Fay, M., Malinovsky, Y.: Confidence intervals of the Mann-Whitney parameter that are compatible with the Wilcoxon-Mann-Whitney test. *Statistics in Medicine*. **37**(**27**), 3991–4006 (2018).
25. Fedorov, V., Mannino, F., Zhang, R.: Consequences of dichotomization. *Pharmaceutical Statistics.***8**(**1**), 50–61 (2009).
26. Permutt, T., Li, F.: Trimmed Means for Symptom Trials With Dropouts. *Pharmaceutical Statistics.***16**(**1**), 20–18 (2017).
27. Keene, O.: Strategies for composite estimands in confirmatory clinical trials: Examples from trials in nasal polyps and steroid reduction, *Pharmaceutical Statistics.***18**(**1**), 78–84 (2019).
28. Lachin, J.: Worst-Rank Score Analysis With Informatively Missing Observations in Clinical Trials. *Controlled Clinical Trials.***20**(**5**), 408–422 (1999).

29. Wang, D., Pocock, S.: A Win Ratio Approach to Comparing Continuous Non-Normal Outcomes in Clinical Trials. *Pharmaceutical Statistics*. **15**(**3**), 238–245 (2016).
30. Bornkamp, B., Rufibach, K., Lin, J., Liu, Y., Mehrotra, D., Roychoudhury, S., Schmidli, H., Shentu, Y., Wolbers, M.: Principal stratum strategy: Potential role in drug development. *Pharmaceutical Statistics*. **20**(**4**), 737–751 (2021).
31. Franklin, J., Patorno, E., Desai, R., Glynn, R., Martin, D., Quinto, K., Pawar, A., Bessette, L., Lee, H., Garry, E., Gautam, N., Schneeweiss, S.: Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative. *Circulation*. **143**, 1002–1013 (2021).
32. Keyhani, S., Cheng, E., Hoggatt, K., Austin, P., Madden, E., Hebert P., Halm, E., Naseri, A., Johanning, J., Abraham, A., Bravata, D.: Comparative Effectiveness of Carotid Stenting to Medical Therapy Among Patients With Asymptomatic Carotid Stenosis. *Stroke*. **53**, 00–00 (2022).
33. Hampson, L., Degtyarev, E., Tang, R., Lin, J., Rufibach, K., Zheng, C.: Comment on "Biostatistical Considerations When Using RWD and RWE in Clinical Studies for Regulatory Purposes: A Landscape Assessment". *Statistics in Biopharmaceutical Research*, 1–4 (2021).
34. Uemura, Y., Shinozaki, T., Nomura, S., Shibata, T.: Comment on "Biostatistical Considerations When Using RWD and RWE in Clinical Studies for Regulatory Purposes: A Landscape Assessment". *Statistics in Biopharmaceutical Research*. 1–3 (2021).
35. Hampson, L., Chu, J., Zia, A., Zhang, J., Hsu, W., Parzynski, C., Hao, Y., Degtyarev, E.: Combining the target trial and estimand frameworks to define the causal estimand: an application using real-world data to contextualize a single-arm trial. *arXiv preprint arXiv:2202.11968* (2022).
36. van der Laan, M., Rubin, D.: Targeted maximum likelihood learning. *The international Journal of Biostatistics*. **2**(**1**), Article 11 (2006).
37. van der Laan, M., Polley, E., Hubbard, A.: Super learner. *Statistical Applications in Genetics and Molecular Biology*. **6**, Article 25 (2007).
38. van der Laan, M., Gruber, S.: Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international Journal of Biostatistics*. **8(1)**, Article 9 (2012).
39. Petersen, M., van der Laan, M.: Causal Models and Learning from Data: Integrating Causal Modeling and Statistical Estimation. *Epidemiology (Cambridge, Mass.).***25**, 418 (2014).
40. FDA (2019): Rare diseases: Common issues in drug development (Draft guidance). US Food and Drug Administration, Silver Spring, MD. https://www.fda.gov/media/119757/download
41. ICH E10 (2010): Choice of control group in clinical trials, https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-10-choice-control-group-clinical-trials-step-5_en.pdf
42. Gökbuget, N., Kelsh, M., Chia, V., Advani, A., Bassan, R., Dombret, H., Doubek, M., Fielding, A., Giebel, S., Haddad, V. Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. *Blood Cancer Journal.***6**(**9**), e473–e473 (2016).
43. Topp, M., Gökbuget, B., Stein, A., Zugmaier, G., O'Brien, S., Bargou, R., Dombret, H., Fielding, A., Heffner, L., Larson, R.: Safety and activity of blinatumomab for adult patients with relapsed or refractory B-precursor acute lymphoblastic leukaemia: a multicentre, single-arm, phase 2 study. *The Lancet Oncology*. **16**(**1**), 57–66 (2015).
44. Kosiborod, M., Cavender, M., Fu, A., Wilding, J., Khunti, K., Holl, R., Norhammar, A., Birkeland, K., Jørgensen, M., Thuresson, M., Arya, N., Bodegard, J., Hammar, N., Fenici, P.: Lower risk of heart failure and death in patients initiated on sodium-glucose cotransporter-2 inhibitors versus other glucose-lowering drugs. *Circulation*. **136**, 249–259 (2017).
45. Hernán, M., Robins, J.: Causal Inference: What If. CRC Press. Tayler & Francis Group. A CHAPMAN & HALL BOOK (2020).
46. Hernán, M., Brumback, B., Robins, J.: Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology.***11**(**5**), 561–570 (2000).

47. Hernán, M., Cole, S., Margolick, J., Cohen, M., Robin, J.: Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Parmacoepidemiology and Drug Safety*. **14**, 477–491 (2005).
48. Robins, J., Rotnitzky, A.: Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*. **90**, 122–129 (1995).
49. Mallinckrodt, C., Lin, Q., Molenberghs, M.: A structured framework for assessing sensitivity to missing data assumptions in longitudinal clinical trials. *Pharmaceutical Statistics*. **12**, 1–6 (2013).
50. Hernán, M., Lanoy, E., Costagliola, D., Robins, J.: Comparison of Dynamic Treatment Regimes via Inverse Probability Weighting. *Basic & Clinical Pharmacology & Toxicology.***98**(**3**), 237–242 (2006).
51. Gruber, S., Lee, H., Phillips, R., Ho, M., van der Laan, M.: Developing a Targeted Learning-Based Statistical Analysis Plan. *Statistics in Biopharmaceutical Research*. 2022 Aug 23, 1–20 (2022).
52. Aalen, O., Cook, R., K. Røysland: Does Cox analysis of a randomized survival study yield a causal treatment effect. *Lifetime Data Analysis*. **21**(**4**), 579–593 (2015).
53. Hernán, M.: The hazards of hazard ratio. *Epidemiology.***21**(**1**), 13–15 (2010).
54. ASA RWE SWG Phase III Team 3: Examples of Applying RWE Causal-Inference Roadmap to Clinical Studies. *Statistics in Biopharmaceutical Research*. (submitted 2022).

Clinical Studies Leveraging Real-World Data Using Propensity Score-based Methods

Heng Li and Lilly Q. Yue

1 Introduction

One of the major contributions that RWD (or more precisely the RWE they generate) can make to the clinical development of medical products is the improvement of efficiency of this process. The subject of this chapter is the leveraging of RWD for this purpose via a type of study design where the study data consists of two parts: (1) those collected on patients prospectively enrolled into a traditional clinical study and (2) RWD. We refer to such a design as a hybrid design and a study so designed as a hybrid study. Here "prospective" means "future" relative to the time when the hybrid study is being planned. Therefore, by definition, when a hybrid study is being planned, patients in the "traditional clinical study" portion of the hybrid study are not yet available. In contrast, the RWD portion of the hybrid study may contain patients who are already available (i.e., the intended treatment has already been administered and/or outcome data already exist) when the hybrid study is being planned. Henceforth, the traditional clinical study portion of a hybrid study will be referred to by the abbreviation "TCS." The use of RWD may be due to ethical or practical considerations and can often save time and reduce cost, which is what motivates these study designs. We discuss the following three kinds of hybrid studies: (1) a non-randomized comparative study in which RWD is used as a comparator group for the TCS. Statistical methods are implemented so that the non-randomized study can be regarded as an approximation of a randomized controlled trial (RCT) in some sense; (2) a single arm, non-comparative hybrid study in which the TCS consists of M prospectively enrolled patients, and these patients are augmented by RWD patients. All patients undergo the same treatment. Statistical

H. Li (✉) · L. Q. Yue
Division of Biostatistics, Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD, USA
e-mail: Heng.Li@fda.hhs.gov

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_10

methods are implemented so that this hybrid study approximates a traditional single arm clinical study consisting of N (>M) prospectively enrolled patients; and (3) a hybrid study in which the TCS is an RCT consisting of prospectively enrolled patients, which is augmented by RWD patients to approximate a traditional RCT of a larger sample size, using similar statistical methods, as in (2). Of course, our premise is that the RWD being considered in a hybrid study are fit-for-purpose, a concept that is discussed elsewhere in the book (ref. chapter "Key Variables Ascertainment and Validation in RW Setting") and hence will not be belabored here. In the rest of this chapter, we assume that this premise holds true and a hybrid design is appropriate given the objectives of the study, which may be to support a marketing application, to seek approval for a labeling expansion, or to inform some other decisions. To properly design and analyze a hybrid study, special statistical methods are needed as mentioned above. These methods, including their statistical underpinnings, will be described in the following sections. But before doing that, we provide an overview of what statistical issues these methods are developed to address.

To justify a hybrid study design, the most fundamental issue to be dealt with is the potential systematic differences between the RWD patients and the TCS patients. These systematic differences are a source of bias, and they are a hindrance to achieving the goal of a hybrid study, which is to approximate a traditional clinical study. Bias may be categorized depending on its source and there is not a standard taxonomy for categories of bias. Before deciding to adopt a hybrid design, it is important to assess the risk of bias from many different sources. If this risk is too high, then a hybrid design may not be appropriate, given that most biases cannot be corrected by statistical means. However, one type of bias, which we will refer to as confounding bias, can be mitigated statistically. This is the bias that can be addressed by the statistical methods to be described in the following sections.

The phrase confounding bias, as used in this chapter, refers to the bias induced by systematic difference between the RWD patients and the TCS patients in terms of the distribution of covariates. It is a familiar concept in the context of non-randomized comparison of treatment (or exposure) groups, which is instanced by the first kind of hybrid study. The objective of such a comparison is almost always causal inference, i.e., the evaluation of the outcome under one treatment relative to that of the other treatment on the same set of patients. Confounding bias is obviously an obstacle to causal inference, insofar as the covariates differentially distributed between the treatment groups are related to the outcome. It is widely known that a standard statistical approach to mitigating such bias is the propensity score methodology. The concept of propensity score was introduced by Rosenbaum and Rubin [1, 2]. The basic idea behind the propensity score methodology as applied to non-randomized comparative studies is to form sets of patients in which the distributions of observed covariates are equalized between the two treatment groups being compared so that "fair" comparison of treatments can be made. This classical application of propensity score methodology to the first kind of hybrid studies will be described in Sect. 2.

To see how confounding bias affects the second kind of hybrid studies, recall that such studies are designed to approximate a traditional single arm study with

a larger sample size. For this goal to be achieved, it's important that the RWD patients "look like" the TCS patients. In statistical terms that means the distributions of observed baseline covariates are similar between the TCS and RWD, so that confounding bias is minimized. In practice, however, there is no reason to expect this to be the case at the planning stage. In other words, confounding bias needs to be addressed in designing the second kind of hybrid studies just as with the first (i.e., non-randomized comparative studies). Given the ability of propensity score methods to equalize the covariate distributions between two groups of patients, it is not surprising that they can be applied to address confounding bias in this context as well, as will be delineated in Sect. 3. Confounding bias affects the third kind of hybrid studies in an analogous fashion to the way it affects the second kind of hybrid studies, and the propensity score methods are used in a similar manner for its mitigation, as will be shown in Sect. 4. Of course, propensity score can only be used to address confounding bias due to observed covariates. Therefore, a key assumption underlying any propensity score-based method is that there is no confounding bias due to unmeasured covariates. Chapter "Sensitivity Analysis in the Analysis of Real-World Data" provides a good discussion of situations where this assumption does not hold.

Another issue to be considered in designing a hybrid study is how to ensure that the amount of information contributed by the TCS and that contributed by the RWD, which may contain a large amount of data, are proportionate so that the latter does not overwhelm the former. This is less of an issue for the first type of hybrid studies where the TCS contributes to the estimation of the parameter of interest associated with the treated group and the RWD are used to estimate the parameter of interest associated with the control group. The objective is to estimate the difference between these two parameters, which means that the information contributed by RWD is not going to dominate the information contributed by the traditional clinical study. On the other hand, in the second and third kind of hybrid studies, the RWD is used to augment the TCS in the estimation of the same parameter of interest. If the sample size of RWD is too large, then too much information for this parameter comes from the RWD relative to that from the TCS, which can sometimes be a concern, depending on specific circumstances. Therefore, in designing a hybrid study of the second or third kind, it is essential to prespecify the maximum amount of information coming from RWD, based on clinical judgment. To make sure that this maximum is not exceeded, the RWD often needs to be down-weighted, or "discounted." Such down-weighting can be achieved using various methods. We will discuss two of these methods for Bayesian and frequentist inference, respectively, namely, power prior and composite likelihood.

The brief discussion above is intended to tell the reader that the propensity score methodology is a useful tool in the design of hybrid studies, and it typically is to be used in conjunction with a discounting method such as power prior or composite likelihood in designing the second and third kind of hybrid studies. The description of these methods and their implementation are the topics of the following sections. One thing to keep in mind in using these methods to design hybrid studies is the integrity of study design. In a traditional RCT, study design

necessarily precedes outcome data collection. However, this is not necessarily true for hybrid studies, where study design is an extensive process, including equalizing covariate distributions between patients in the TCS and those from the RWD, while outcome data may be already available prior to or during this process. To maintain the integrity of study design, thereby enhancing the interpretability of study results, all design activities need to be carried out while blinding to outcome data is administered. Therefore, besides statistical methodology, the practicalities of such blinding will also be discussed.

2 Propensity Score and Type 1 Hybrid Studies

2.1 The Concept of Propensity Score

Suppose a medical product is to be evaluated in a non-randomized comparative study following the type 1 hybrid design, in which the TCS patients constitute the "treated group" (i.e., they undergo the medical intervention being studied) and the RWD patients serve as the control group. A main statistical consideration in designing an observational study like this is minimizing bias due to potential difference in the distributions of observed baseline covariates between the treated and the control groups (confounding bias) and ensuring the objectivity of study design, and propensity score (PS) methodology is standard for handling such challenges. In this subsection, we only provide a summary of the concept of PS and refer the reader to Imbens and Rubin [3] for more details.

The PS $e(X)$ for a patient with a vector X of observed baseline covariates in a comparative study is the conditional probability of being in the treated group ($T = 1$) rather than the control group ($T = 0$) given the vector of baseline covariates X [1, 2]:

$$e(X) = \Pr(T = 1 | X)$$

PS is a balancing score in the sense that conditional on the PS, the distribution of observed baseline covariates is the same between the treated and control patients. Therefore, among patients with the same value of PS, the distribution of observed covariates is the same between these two groups of patients. In other words, the treatment assignment indicator T and the covariate vector X are conditionally independent given the PS $e(X)$, or

$$T \perp X \mid e(X).$$

A practical implication of this balancing property is that, to equalize the distribution of X (or balance X) between the treated and the control groups, one only needs to balance $e(X)$ between these two groups, which is easier since $e(X)$ is a scalar (one-dimensional).

Another property of PS that more directly reveals its utility in causal inference for the treatment effect in any non-randomized comparative study (or observational study) is as follows. Let $Y(1)$ be the potential outcome of a patient if assigned to the treated group and $Y(0)$ be the potential outcome of the same patient if assigned to the control group. Note that some assumptions are needed for the above potential outcomes notation to make sense. However, we will not get into these assumptions here because such potential outcomes notation is commonly used as a starting point for a discussion of causal inference. A comparison between $Y(1)$ and $Y(0)$ defines a causal effect of the investigational treatment relative to the control on a patient.

A comparison of the distribution of $Y(1)$ on a patient population and the distribution of $Y(0)$ on the same patient population defines a causal effect of the investigational treatment relative to the control on this patient population. The treatment assignment mechanism is said to be unconfounded if

$$Y(1), Y(0) \perp T \mid X.$$

From this assumption one can deduce that

$$Y(1), Y(0) \perp T \mid e(X).$$

This property tells us that if the assignment mechanism is unconfounded, then among patients with the same PS, the observational study reduces to an RCT. Hence, if the unconfoundedness assumption holds and the PS of every patient is known, then causal inference for an observational study would conceptually amount to using a valid method for RCT to estimate treatment effect at each distinct value of PS and combining these estimates. In a typical observational study, however, patients' true PSs are unknown and can only be estimated. So, in practice, estimated PSs are used in lieu of true PSs. The strategy is to create sets of patients in which the distribution of estimated PSs in the treated group is similar to that in the control group. This can be achieved in several ways, with the most common ones being matching, weighting, and stratification. Whether good estimates of PSs have been obtained can be directly checked, by examining the distributions of observed covariates in the treated and the control groups to assess balance. If these distributions are not close enough to each other, or, in other words, if some covariates are not adequately balanced according to a pre-specified criterion (more on this later), one may adjust the estimation model for the PSs and obtain a new set of estimates. Thus, PS estimation is an iterative process. In fact, if the PS methodology is to be applied to estimate the treatment effect in an observational study, then this iterative process, called PS design [4], constitutes a major part of the design of this observational study, as will be discussed in the next subsection.

2.2 Estimation of Propensity Score and Assessment of Balance

While a variety of methods for estimating PS have been introduced, logistic regression as suggested in Rosenbaum and Rubin [1] is perhaps still the most widely

used. It postulates that the logit of propensity score is a polynomial in the observed covariates. The linear model (polynomial of degree 1) is often used as the initial PS model. Computationally, in the logistic regression the treatment assignment indicator T is identified as the dependent variable and the observed covariates X are identified as independent variables. After the PS is computed for each patient, covariate balance is carried out via matching, weighting, or stratification. We now give a brief description of each of the three schemes.

PS matching is a method of selection from a pool of control patients such that the selected subset has better covariate balance relative to the treated group than the set of all control patients. The selection can be achieved with a matching algorithm [5], and one of the most common matching algorithms might be the k:1 nearest neighbor matching [6]. In its simplest form, for each treated patient i, 1:1 nearest neighbor matching selects a control patient with the smallest distance from i, where the distance between two patients is usually defined by the absolute difference between the logit of their estimated PSs. Patients in the pool of controls that are not matched to any treated patients are not included in the subsequent estimation of the treatment effect (i.e., "discarded"). Some matching algorithms allow treated patients to be discarded. It should be noted that such matching algorithms are usually not recommended for hybrid studies defined in this chapter. This is because in such studies, patients enrolled into TCS usually represent the population for which the investigational medical product is indicated. Discarding treated patients would risk distortion of patient population and change of indication for use of the medical product.

PS weighting is defined as using one function of PS to weight patients in the treated group and another function of PS to weight patients in the control group so that the weighted distributions of covariates in the two groups are equal [7]. While the choice of this pair of functions is not unique, only a few of them are in common use. Once the choice is made, the weights corresponding to the chosen functions, called balancing weights [8], are then used to weight the outcome variable in the subsequent estimation of the treatment effect. One of the possibilities for balancing weights is:

$$w_1^{(ATE)}(X) = \frac{1}{e(X)} \text{ and } w_0^{(ATE)}(X) = \frac{1}{1 - e(X)}.$$

Here the subscripts "1" and "0" represent the treated group and the control group, respectively. The superscript "(ATE)" stands for Average Treatment Effect. When this pair of weights are applied to the outcome variable, the estimand is the average treatment effect on the population represented by all the patients in the study. Austin [9] refers to these weights as IPTW-ATE weights, where "IPTW" stands for "inverse probability of treatment weighting" [10]. Another possibility for balancing weights is

$$w_1^{(ATT)}(X) = 1 \text{ and } w_0^{(ATT)}(X) = \frac{e(X)}{1 - e(X)},$$

where "(ATT)" stands for Average Treatment effect on the Treated. Austin [9] refers to these weights as IPTW-ATT weights. When this pair of weights are applied to the outcome variable, the estimand is the average treatment effect on the population represented by the patients in the treated group. Since true PSs are usually unknown, estimated PSs are plugged into the expressions for balancing weights to produce estimated weights to be used in the estimation of the treatment effect.

PS stratification forms subsets (strata) of patients within which the treated group and the control group are more similar than they are overall. Specifically, patients are first sorted by their estimated PSs and then stratified based on prespecified cut points (e.g., PS quintiles), so that within each stratum the PS distribution in the treated group is similar with that in the control group. By the balancing property of PS, that means within each stratum, the joint distributions of covariates are similar between the treated and the control groups as well. In terms of the estimation of the treatment effect, PS stratification can be viewed as a variant of PS weighting where the estimated PSs are further smoothed before being plugged into the expressions for the balancing weights [7]. Here is how the smoothing is carried out. Each patient's estimated PS is replaced by another value called coarsened PS, which, for a patient in any given stratum, is equal to the proportion of patients who are in the treated group in that stratum. Hence all the patients in a given stratum have the same coarsened propensity score. To estimate balancing weights, the coarsened PS is plugged into the expression of balancing weights. If the intended estimand is ATE, the coarsened propensity score is plugged into $w_1^{(ATE)}(x) = \frac{1}{e(x)}$ and $w_0^{(ATE)}(x) = \frac{1}{1-e(x)}$ to obtain the estimated balancing weights. If the intended estimand is ATT, the coarsened PS is plugged into $w_1^{(ATT)}(x) = 1$ and $w_0^{(ATT)}(x) = \frac{e(x)}{1-e(x)}$ to obtain the estimated balancing weights. For the ATE and ATT estimands, smoothing the estimated PSs via stratification before they are used to estimate balancing weights may avoid the potential situation where a few subjects have extremely large weights relative to the other subjects, thereby dominating the study results, a problem caused by the unboundedness of IPTW-ATE and IPTW-ATT weights [7]. Of course, coarsening the PS has the potential downside of increasing residual imbalance. Imbens and Rubin [3] also contains a discussion on propensity score stratification as compared to IPTW weighting.

The main purpose of the PS design is balancing observed covariates. Therefore, the estimation of PS is followed by balance assessment, to make sure that the estimated PSs achieve the purpose for which they are intended. Methods for balance assessment, also called balance diagnostics [11], can be divided into two categories: graphical and numerical. A variety of different methods have been used in practice and some may not even be in the literature. In this section, we present one common numerical method. It is formally for PS weighting but can be adapted for PS matching and stratification as well. It uses a metric called absolute standardized mean difference (ASMD), which can be defined as follows for continuous covariates:

$$d = \frac{\mid \overline{x}_{w.treated} - \overline{x}_{w.control} \mid}{\sqrt{\frac{s^2_{w.treated} + s^2_{w.control}}{2}}},$$

where $\overline{x}_{w.treated}$ ($\overline{x}_{w.control}$) is the weighted (using the estimated balancing weights) sample mean of the covariate whose balance is under consideration in the treated (control) group, and $s^2_{w.treated}$ ($s^2_{w.control}$) is the weighted sample variance of the covariate in the treated (control) group [7, 10]. For binary covariates, ASMD can be defined as

$$d = \frac{\mid p_{w,treated} - p_{w,control} \mid}{\sqrt{\frac{p_{w,treated}(1-p_{w,treated})+p_{w,control}(1-p_{w,control})}{2}}},$$

where $p_{w,\,treated}(p_{w,\,control})$ is the weighted proportion corresponding to the binary covariate in the treated (control) group [10]. To apply the metric d to a categorical covariate with more than two categories, we may decompose it into several binary covariates. Balance is considered adequate for this covariate if d is smaller than a prespecified threshold d_0. While there is no clear consensus on the choice of d_0, some researchers have proposed a value of 0.1 [11]. Given its form, the metric d may also be applied to PS stratification if the weights are obtained from the corresponding coarsened PSs. For k:1 nearest neighbor matching, balance can be assessed using the unweighted version of d.

If balance is adequate for all observed covariates, then the PS design is complete. Otherwise, another iteration is started by adjusting the PS model to obtain a new set of estimated PSs. One way to adjust the PS model is to add higher order terms of some covariates to the model. By and large the iterative process of PS design is more of an art than a science involving trial and error. It is possible that despite one's best effort, adequate balance cannot be achieved. This is a risk inherent to the application of propensity score methodology. In practice, if adequate balance cannot be achieved, then one may consider other RWD sources. Another issue to consider is the possibility that multiple PS models can lead to adequate balance. This kind of multiplicity combined with the availability of outcome data prior to or during PS design is of concern. Unless some measures are taken to preempt data dredging, study integrity and objectivity may be compromised. These measures are discussed in the next subsection.

2.3 The Two-Stage Paradigm for Study Design

The two-stage design proposed by Yue et al. [12] is a framework for the practical implementation of the idea of outcome-free design [13, 14] for the application of PS methodology. As pointed out earlier, the goal of PS design is to find a set of PS estimates that can balance all observed covariates through a trial-and-error process, and this set of PS estimates is not unique. Such multiplicity creates an opportunity for data dredging, given that some outcome data, especially those of the RWD, may already exist prior to or during PS design. Therefore, how study integrity and objectivity can be maintained, given this opportunity for data dredging, would be

a critical question. To be more concrete, this is a question about how to preclude the possibility of existing outcome data influencing the PS study design. Rubin's [15] answer to this question is clear: outcome data should not be in sight during PS design. This is what Yue et al. [12] refer to as the outcome-free principle, and their two-stage design puts it to practice.

The essence of outcome-free design is blinding or masking of patient-level outcome data to the process of PS design, which can also be referred to as building a firewall in the biopharmaceutical arena. The scheme that Yue et al. [12] propose is for the investigator of the study to identify an independent statistician to perform the PS design, with no outcome data provided to the independent statistician. The independent statistician shares with the investigator the responsibility of upholding the outcome-free principle [4, 12, 16–18]. This independent statistician is identified in the first design stage of the two-stage design of Yue et al. [12], so are all the covariates to be balanced in the PS design. Otherwise, the first design stage consists of all the elements of the design of an RCT, such as the specification of the study endpoints, the study hypotheses (together with their significance levels), and the initial sample sizes for the treated and the control groups. The reason for the qualifier "initial" is that these sample sizes are revisited in the second design stage and may be revised later prior to the unblinding of the outcome data. The PS design itself constitutes the second design stage, in which the independent statistician identified in the first design stage, who is blinded to outcome data, carries out PS estimation, performs PS matching, weighting, or stratification, and assesses covariate balance. In the next subsection, we give a numerical example to illustrate the implementation of the two-stage design of Yue et al. [12].

2.4 An Illustrative Numerical Example of a Type 1 Hybrid Study

Suppose a type 1 hybrid study is planned to evaluate a medical product. The treated group consists of patients enrolled into a traditional clinical study (the TCS part of the hybrid study) and the control group is to come from an RWD source. Based on clinical and regulatory judgment, an existing national patient registry is thought to be a suitable such RWD source with respect to data quality and availability of patient-level data for both clinical outcomes and baseline covariates of interest. The two-stage design is to be adopted. The first design stage includes the following elements:

1. It is decided that the primary endpoint is the binary variable of treatment success.
2. The primary hypotheses are specified to be those of non-inferiority on the difference scale (i.e., difference of the two probabilities of treatment success) with a margin of 6% and a one-sided significance level of 0.025.
3. Fifteen baseline covariates are identified as needing to be balanced between the treated and the control groups based on clinical considerations. It is verified that

these covariates are collected in the chosen RWD source. The study proposal also includes the stipulation that these covariates will be collected in the TCS.
4. PS stratification with five strata of equal size based on PS quintiles is planned.
5. The procedure described in Yue et al. [12] for calculating the initial sample size (for the PS stratification with five strata of equal size) are directly applied. It is determined that 300 patients in the treated group and 600 patients in the control group may achieve 90% power.
6. An independent statistician is contracted to perform the PS design in the second design stage. These action items are summarized in Table 1.

In this study, the plan for RWD patient acquisition is to extract from the control data source those patients who meet the eligibility criteria of the study and enter the registry between two given dates. It is anticipated that this simple selection rule would yield more control patients than the 600 given by the initial sample size calculation. In general, it is a good idea to have some extra control patients, given the various uncertainties arising in the propensity score design.

The second stage of the two-stage design starts when the patient enrolment into the TCS is complete and so is the patient extraction from the RWD source, at which point baseline covariate data are available for all patients. As planned, 300 patients are enrolled into the TCS. The number of patients extracted from the RWD source happens to be 1000. To build the propensity score model, logistic regression is performed by the independent statistician on those 1300 patients with treatment group membership as the dependent variable and the 15 covariates identified in the first design stage as the independent variables, based on which a PS is calculated for each patient. The patients are then stratified into five PS quintiles. Table 2 shows the number of treated and control group patients in each of the five PS quintiles.

We can see from Table 2 that the first stratum (or PS quintile) contains 250 RWD control patients but no patients from the TCS part of the study, because there were no TCS patients who look like those control patients with respect to propensity

Table 1 Main elements of the first design stage

Primary outcome: treatment success for a patient
Non-inferiority margin: $\delta = 6\%$
Significance level: 0.025 one-sided
Number of baseline covariates considered: 15
Propensity score stratification planned for study design and outcome analysis
Independent statistician identified
Initial sample size for the treated group: $N = 300$
Initial sample size for the control group: $N = 600$

Table 2 Distribution of all 1300 patients across the five propensity score quintiles

	Propensity score quintiles					
	1	2	3	4	5	Total
Control	250	244	234	186	86	1000
Investigational	0	11	20	79	190	300

Table 3 Distribution of the 1050 patients across the five propensity score quintiles

	Propensity score quintiles					
	1	2	3	4	5	Total
Control	196	193	172	128	61	750
Investigational	10	33	67	80	110	300

score and with respect to some covariates. Therefore, it is considered reasonable to discard the 250 RWD control patients in that stratum (i.e., the first PS quintile). Note that any attempt to exclude TCS patients treated with the investigational device is discouraged as the patient population represented by TCS is usually the population for which the medical product is indicated. Discarding treated patients would risk distortion of patient population and could impact the indication for use of the medical product, as pointed out in Sect. 2.2.

After excluding the 250 RWD control subjects in the first PS quintile, the independent statistician continues with the iterative process of PS design based on the remaining 1050 patients. The iterative process consists of fitting a logistic regression to estimate PSs, stratifying patients into five PS quintiles, assessing balance between the treated and the control groups for each covariate within each quintile, and, if balance is not adequate for some covariates, go back to fit a new logistic regression (e.g., by adding quadratic or cross-product terms). The process continues until balance is satisfactory (in this case ASMD <0.1) for all the covariates. At this point, power is revisited and is found to be adequate. The distribution of the 1050 patients (300 in the treated group and 750 in the control group) based on the final logistic regression model is shown in Table 3. This table is added to the statistical analysis plan, along with the final logistic regression equation. The second design stage is now complete. During the entire PS design, only the treatment assignment and baseline covariate data are needed. Any clinical outcome data and follow-up information are neither needed nor accessed by the independent statistician.

After the completion of the entire study design, the outcome data are analyzed. The ATT estimand was specified at the planning stage (as is usually the case for type 1 hybrid studies) and the outcome data analysis is carried out accordingly. As it turns out, the p-value for the non-inferiority hypotheses is 0.021, which means the null hypothesis can be rejected.

3 The Design and Analysis of Type 2 Hybrid Studies

3.1 Definition and Fundamental Statistical Issues

Section 1 introduced the concept of a hybrid study and gave a definition for each of the three types of hybrid studies. Instead of repeating the definition for type 2 hybrid studies, let us use an example to help the reader recall what it is. Suppose a

study is being planned that will provide evidence to support a new indication for an approved medical product. A single-arm traditional clinical study (in this section all traditional clinical studies are single-arm so we may drop the qualifier "single-arm" where there is no confusion) is to be conducted that enrolls patients prospectively. Data from the off-label use of the product have been captured in a high-quality patient registry, forming an RWD source for the evidence. It is determined that these RWD are reliable and relevant and can be leveraged to reduce the sample size of the traditional clinical study. These considerations point to a type 2 hybrid study. More specifically, suppose the evidence required for the labeling expansion can be provided by a traditional clinical study of size N, and such a study can be well approximated by a traditional clinical study of size M ($M < N$) augmented by some RWD patients receiving the same treatment, then these M patients plus the RWD constitute a type 2 hybrid study. The traditional clinical study of size M is the TCS (see Sect. 1 for the meaning of this abbreviation) part of the type 2 hybrid study, while the RWD part contributes a nominal $N - M$ patients. Here we use the word "nominal" to indicate that the actual number of patients that the RWD contain may be much larger than $N - M$. Henceforth, in this section, type 2 hybrid study may be referred to simply as hybrid study where there is no confusion.

Given the above definition, the first decision to be made in planning a hybrid study, if such a study is deemed acceptable from a clinical perspective for the given purpose, is the magnitude of M, or equivalently of $N - M$ (which will be referred to as A). This represents the amount of information to be leveraged from RWD and its determination is based on clinical judgment considering various clinical characteristics of the RWD source. Obviously, if A is too large, then it may not be reasonable to consider the hybrid study an approximation of a traditional clinical study. All subsequent discussion in this section is under the premise that a hybrid study is a viable alternative to a traditional clinical study and the numerical value of A has already been decided. We focus on the statistical issues of (1) how to ensure that the nominal number of RWD patients does not exceed A, and (2) how to mitigate confounding bias so that the hybrid study can better approximate a traditional clinical study. The first issue is essentially one of down-weighting or "discounting," for which two alternative methods will be described, one Bayesian and the other frequentist. The Bayesian method is that of power prior and the frequentist method is that of composite likelihood, and both will be summarized in Sect. 3.2.

To address the second issue, let us recall from Sect. 1 that confounding bias refers to the systematic difference between the RWD patients and the TCS patients in terms of the distribution of covariates. The presence of such confounding bias clearly makes it less convincing that the hybrid study can approximate a traditional clinical study well. In Sect. 3.3, we delineate how the tool of PS methodology can be repurposed to mitigate this confounding bias. Section 3.4 provides a step-by-step description of the two-stage design of a hybrid study to approximate a single-arm traditional clinical study that addresses the above two statistical issues. The second stage of the two-stage design is a PS design analogous to that described in Sect. 2, with an additional element associated with the down-weighting of RWD patients.

3.2 Using Power Prior or Composite Likelihood to Down-Weight RWD Patients

The power prior [19] is originally intended to be an informative prior constructed from historical data [20]. If we substitute RWD for historical data, the method fits our purpose of down-weighting RWD patients perfectly. In our context, a power prior π for a parameter θ associated with an endpoint based on data collected on RWD patients for that endpoint, $\boldsymbol{D_0}$, is constructed as follows:

$$\pi\left(\theta\right) \propto \left[L\left(\theta|\boldsymbol{D_0}\right)\right]^{\alpha} \pi_0\left(\theta\right)$$

where $L(\theta|\boldsymbol{D_0})$ is the likelihood function of θ given the RWD, $\pi_0(\theta)$ is the initial prior distribution for θ, and α $(0 \leq \alpha \leq 1)$ is called the power parameter. This prior is multiplied to the likelihood function of θ given the TCS data $\boldsymbol{D_1}$, $L(\theta|\boldsymbol{D_1})$, to obtain the posterior distribution of θ,

$$\pi\left(\theta|\boldsymbol{D_1}\right) \propto \left[L\left(\theta|\boldsymbol{D_1}\right)\right]\ \pi\left(\theta\right),$$

completing the statistical inference for θ. From this construction, α can evidently be interpreted as the fraction of information RWD patients contribute to the inference for θ. In other words, α is the weight by which the RWD patients are discounted. For example, if $\alpha = 0.1$, each RWD patient contributes 10% of their information, and the total amount of information the RWD patients bring to the statistical inference is equivalent to the information contributed by 0.1 times the total number of RWD patients, which can be interpreted as the nominal number of patients being leveraged for some common distributions such as normal and binomial. If $\alpha = 1$ then the nominal number of patients leveraged is equal to the actual number of RWD patients constituting $\boldsymbol{D_0}$. At the other extreme, if $\alpha = 0$, then no RWD patients are leveraged. In general, if α is equal to the nominal number, RWD patients that one wants to leverage divided by the actual number of RWD patients constituting $\boldsymbol{D_0}$.

The composite likelihood [21] for the parameter of interest θ is a weighted product of probability density functions:

$$L\left(\theta|\boldsymbol{Y}\right) = \prod\nolimits_i f\left(y_i|\theta\right)^{\lambda_i}$$

where each i represents a patient and λ_i is a nonnegative weight. Clearly, when all the λ_i's equal to 1, composite likelihood reduces to ordinary likelihood. To use composite likelihood to serve the purpose of down-weighting RWD patients, we let $\lambda_i = 1$ for TCS patients and $0 \leq \lambda_i \leq 1$ for RWD patients. If statistical inference for θ is conducted based on the composite likelihood after assigning numerical values to λ_i's in this way, then we are essentially down-weighting the RWD patients relative to the TCS patients. For example, if $\lambda_i = 0.1$ for all RWD patients, then each RWD patient contributes roughly 10% of their information, and the nominal number of RWD patients leveraged is 0.1 times the actual number of RWD patients.

If $\lambda_i = 1$ for all i, then the nominal number of RWD patients leveraged is equal to the actual number of RWD patients. If $\lambda_i = 0$ for all RWD patients, then no RWD patients are leveraged. In general, λ_i is equal to the nominal number of RWD patients that one wants to leverage divided by the actual number of RWD patients for all i labeling an RWD patient. We will see that the value of α or λ_i is determined before the unblinding of outcome data.

In this subsection, we provided a summary of the methods of power prior and composite likelihood. However, we will not directly apply them to the entire dataset. Instead, we apply them within PS strata with PS being defined in the Sect. 3.3 for type 2 hybrid studies and Sect. 4.2 for type 3 hybrid studies. Due to the balancing property of PS, within each PS stratum the distributions of observed covariates are similar between the TCS and the RWD. If there is minimal confounding bias due to unmeasured covariates, then the outcome variable is expected to be relatively homogeneous between the TCS and RWD, making the application of power prior or composite likelihood more justified.

3.3 The Propensity Score Redefined

In Sect. 2, we introduced the concept of PS in the context of an observational study comparing two treatment groups: the treated group and the control group. PS is defined as the conditional probability of being in the treated group rather than the control group, given the vector of baseline covariates X. An immediate consequence of this definition is that PS is a balancing score. The (joint) distribution of covariates in the treated group is the same as that in the control group conditional on PS. In other words, in any subset consisting of all patients whose PS is equal to a given value, the covariates X are balanced. Thus, in any such subset confounding bias due to X is removed, thereby removing one obstacle to the fair comparison between the treated group and the control group. Since PSs are generally unknown, one way to take advantage of this balancing property in practice is to estimate PS first and then stratify patients according to the estimated PSs so that within each stratum the PS is relatively homogeneous. Within-stratum treatment effects are estimated and then combined into an overall treatment effect. With estimated PS, it is not expected that confounding bias is completely removed within each PS stratum. However, if good covariate balance is observed, confounding bias is substantially mitigated.

Given that the above scheme is now widely and successfully used to mitigate confounding bias in non-randomized comparative studies, one may ask whether it can be co-opted to mitigate confounding bias in type 2 hybrid studies. The answer is yes, and here is how it can be done. First, in defining PS the treated group and the control group are replaced with TCS and RWD, respectively. In other words, the PS $e(X)$ for a patient with a vector X of observed baseline covariates in a type 2 hybrid study is the conditional probability of being in the TCS ($Z = 1$) rather than the RWD ($Z = 0$), given the vector of baseline covariates X:

$$e(X) = \Pr(Z = 1|X).$$

Second, the estimation of PS is done in an analogous way, e.g., by conducting a logistic regression with the indicator variable Z as the dependent variable and the observed covariates X as independent variables. A similar two-stage outcome-free design is carried out, with the PS design being part of the second stage. PS stratification is performed, and the amount of down-weighting of RWD patients is decided for each stratum. Finally, outcome data are unblinded and statistical inference is conducted, first for the stratum-specific parameter of interest, and then for the overall endpoint parameter by combining the stratum-specific parameters. The rationale for the leveraging of RWD to be done within each PS stratum first is that there is less confounding bias within each stratum thanks to the balancing property of PS, making the leveraging more justified. This strategy is termed propensity score-integrated approach [22–25].

In the next subsection, we provide a detailed description of the propensity score-integrated approach to type 2 hybrid studies through a numerical example.

3.4 The Propensity Score-Integrated Approach for Type 2 Hybrid Studies

Suppose a type 2 hybrid study, as set forth in Sect. 3.1, is proposed, to which the propensity score-integrated approach is applied. The associated two-stage design is described in detail below, which is the same whether Bayesian or frequentist inference is planned. For the first design stage, the primary endpoint of interest is the occurrence of adverse event(s) within 1 year and the parameter of interest θ is the probability of a TCS patient experiencing adverse event(s) within 1 year. The primary endpoint hypotheses are

$$Ho : \theta \geq 36\% vs. Ha : \theta < 36\%,$$

where 36% is called the performance goal. Assuming $\theta = 0.30$, standard sample size calculation tells us that to achieve 80% power at a one-sided significance level of 0.05 (this corresponds to posterior probability threshold of 0.95 for Bayesian inference), 380 patients are needed. The proposal is to enroll 290 patients into the TCS part of the hybrid study and to leverage 90 RWD patients, based on clinical input and regulatory considerations. By leveraging 90 patients, what we mean is that the amount of information leveraged is equivalent to that of 90 patients. The idea is to take all eligible patients from the registry (much more than 90 patients), and down-weight those patients relative to the TCS patients in statistical inference. In this example, eligible means meeting the inclusion/exclusion criteria in the TCS and entering the registry during the time when the TCS is enrolling. Seventeen covariates are identified whose distributions will ideally be similar between the TCS and the

RWD parts of the hybrid study for it to be regarded as a good approximation of a traditional clinical study. All the 17 covariates are collected by the registry serving as the RWD source. The plan is to balance these covariates between the TCS and the RWD of the hybrid study by PS stratification with the PS defined as in Sect. 3.3. An independent statistician is thus appointed for the PS design in the second design stage. The above elements of the first design stage are summarized in Table 4.

After the first design stage is complete, the enrolment in the TCS part of the hybrid study begins. The second design stage starts as soon as all the 290 patients have been enrolled into the TCS and all eligible patients have been extracted from the RWD source, at which time the covariate data for all the patients will be available. The number of eligible RWD patients happens to be 1000. The independent statistician appointed in the first design stage who is blinded to outcome data builds a logistic regression model to estimate the PS for each of the 1290 (290 + 1000) patients. Then 941 RWD patients are selected by excluding those RWD patients whose PSs are not in the range of that of the TCS patients. This step is called trimming. The 1231 patients (290 + 941) are grouped into 5 PS strata in such a way that the same number of TCS patients (58 = 290/5) are in each PS stratum (i.e., using PS quintiles among the 290 TCS patients as cut points). This guarantees that each stratum contains TCS patients. Since within each PS stratum the TCS patients and RWD patients are expected to be more similar than they are overall, the leveraging of RWD patients within stratum is more justified. The numbers of RWD patients and TCS patients in each PS stratum are displayed in Table 5.

Recall that it was decided based on clinical considerations that the total amount of information to be borrowed is equivalent to 90 RWD patients. Since borrowing takes place within each stratum, we need to figure out how to allocate the 90 patients to the 5 PS strata. There are many possible ways to do so. One may allocate equal number of (i.e., 90/5 = 18) patients to each stratum. Our strategy is to make the nominal number of RWD patients to be leveraged in each stratum proportional to the similarity of RWD patients and the TCS patients in terms of baseline covariates

Table 4 Main elements of the first design stage

Primary outcome: probability of adverse event within 1 year
Performance goal: 36%
Significance level: 0.05 one-sided/posterior probability threshold: 0.95
Number of baseline covariates considered: 17
Propensity score stratification planned for study design and outcome analysis
Independent statistician identified
Sample size for the current study: 290
Nominal sample size for RWD patients: 90

Table 5 Sample size in each PS stratum

	1	2	3	4	5	Total
TCS (n)	58	58	58	58	58	290
RWD (n)	281	210	154	187	109	941

Table 6 Overlapping coefficient, standardized overlapping coefficient, nominal number of patients to be borrowed, and power parameter (or composite likelihood exponent) in each stratum

	1	2	3	4	5	Total
Overlapping coefficient	0.87	0.78	0.86	0.84	0.77	
Standardized overlapping coefficient	21%	19%	21%	20%	19%	100%
Patients borrowed (=90 × Std. Overlap Coef.)	19	17	19	18	17	90
α_s (or λ_s) (=Patients Borrowed/RWD (n))	0.07	0.08	0.12	0.10	0.15	

in that stratum. One suggestion is to measure this similarity by an overlapping coefficient [26], the overlapping area of propensity score distributions of the two groups of patients (you may use other reasonable measures). The overlapping coefficients are then standardized so that they add up to 1. The standardized overlapping coefficient times the total nominal number of patients to be borrowed (90) determines the nominal number of RWD patients to be borrowed in each stratum. In this example, the number of RWD patients allocated to each stratum using the suggested strategy is close to that using equal allocation (as shown in Table 6).

The power parameter α_s in the Bayesian approach or the exponent λ_s in the composite likelihood in the frequentist approach in each PS stratum can then be obtained by dividing the nominal number of RWD patients to be leveraged by the total number of RWD patients in that stratum. Having determined α_s (or λ_s) in each PS stratum we know the fraction of information RWD patient contributes, and the study design is complete. The overlapping coefficient, the standardized overlapping coefficient, the nominal number of patients to be borrowed, and the power parameter (or exponent) in each stratum are presented in Table 6. Here, again, all the above design activities are performed by an independent statistician blinded to the outcome data.

After clinical outcomes have been observed from all the patients, the statistical inference is conducted. For the Bayesian approach, apply the power prior method within each stratum to get posterior distributions of stratum-specific parameters of interest θ_s [22], which are then combined to complete the inference for the parameter of interest $\theta = \frac{1}{5}\sum_{s=1}^{5}\theta_s$. Here the number 5 represents the number of strata, and the simple average is because by design there is equal number of TCS patients in each stratum. In general, θ is a weighted average of θ_s with the weight associated with s equal to the number of TCS patients in stratum s [22, 23]. In this example, the posterior probability of $\theta < 36\%$ is 96.9%, which meets the study success criterion. For the frequentist approach, construct the composite likelihood to get stratum-specific maximum likelihood estimates $\hat{\theta}_s$ [23], which are then combined to complete the inference for the parameter of interest $\theta = \frac{1}{5}\sum_{s=1}^{5}\theta_s$. In this example, the combined maximum likelihood estimate $\hat{\theta}$ is 31%, with a one-sided p-value = 0.01.

3.5 More Information on Outcome Analysis

The example in the previous subsection focused on the two-stage outcome-free design for the PS-integrated approach with a very brief description of the outcome analysis. In this subsection, we make some comments on the statistical inference for outcome analysis. Corresponding to the PS stratification, the parameter of interest θ, the probability of a patient experiencing adverse event(s) within 1 year, branches out into S independent stratum specific parameters θ_s $s = 1, \ldots, S$ (S being the number of strata). For Bayesian inference, the idea is to apply the power prior method within each stratum, find the posterior distributions for θ_s, and then combine them to obtain the posterior distribution of θ, via the relation

$$\theta = \frac{\sum_{s=1}^{S} w_s \theta_s}{\sum_{s=1}^{S} w_s}$$

where w_s is number of TCS patients in stratum s. This weighting is chosen because the goal is for the hybrid study to approximate a traditional clinical study represented by the TCS patients. Per the power prior method, the prior distribution of θ_s is

$$\pi\left(\theta_s\right) \propto \left[L\left(\theta_s | \boldsymbol{D}_{\boldsymbol{s},\mathbf{0}}\right)\right]^{\alpha s} \pi_0\left(\theta_s\right)$$

where $\boldsymbol{D}_{\boldsymbol{s},\mathbf{0}}$ represents data collected on the RWD patients in stratum s. The algorithm for obtaining α_s is illustrated in the previous section. The posterior distribution of θ_s is

$$\pi\left(\theta_s | \boldsymbol{D}_{\boldsymbol{s},\mathbf{1}}\right) \propto \left[L\left(\theta_s | \boldsymbol{D}_{\boldsymbol{s},\mathbf{1}}\right)\right] \pi\left(\theta_s\right),$$

where $\boldsymbol{D}_{\boldsymbol{s},\mathbf{1}}$ represents data collected on the TCS patients in stratum s. For frequentist inference, the composite likelihood for θ_s is

$$L\left(\theta_s\right) = \prod_{i \in C} f\left(y_i; \theta_s\right) \prod_{j \in R} f\left(y_j; \theta_s\right)^{\lambda_s}$$

where y_i represents endpoint data collected on a patient, C is the index set for TCS patients and R is the index set for RWD patients; λ_s is obtained in the same way α_s is obtained. Point estimate for θ_s can be obtained by maximizing $L(\theta_s)$, and its variance can be obtained via the methods described in Wang et al. [23]. Finally, a point estimate for θ is

$$\hat{\theta} = \frac{\sum_{s=1}^{S} w_s \hat{\theta}_s}{\sum_{s=1}^{S} w_s}$$

and its variance can be obtained by the fact that $\hat{\theta}_s$ are independent.

Before leaving this section, we would like to point out that, to make our discussion more rigorous, a distinction needs to be made between true PS and estimated PS. In defining stratum-specific parameters θ_s, true PS is used for stratification. On the other hand, in conducting statistical inference for θ_s, including the specification of likelihood functions, estimated PS is used. Therefore, the statistical inference described in this section for θ_s is approximate insofar as the estimated PS is an approximation to the true PS.

4 The Design and Analysis of Type 3 Hybrid Studies

4.1 Definition and Fundamental Statistical Issues

The definition of a type 3 hybrid study has already been given in Sect. 1. Instead of repeating the definition here, let us use an example to refresh our memory. Suppose a 1:1 RCT is being planned to evaluate a medical device by comparing the treatment with the investigational device plus optimal medical therapy to the treatment with optimal medical therapy alone, to find out whether the device has any net benefit. However, there are concerns about slow enrolment because of competing trials, small patient population, etc. So, it is proposed that RWD be leveraged to replace some of the control patients that need to be prospectively enrolled into the RCT. Specifically, a high-quality patient registry where patients are treated with the medical therapy for the control arm is deemed to be an appropriate source of the RWD.

To be consistent with Sect. 3, we use A to denote the nominal number of RWD patients being leveraged to augment the control arm. For the same reason as stated in Sect. 3, it is desired to limit the size of A, and its determination is based on clinical judgment considering various clinical characteristics of the RWD source. It would be convenient to conduct a type 3 hybrid study with the TCS part being a 2:1 RCT and let A be the number of control patients in the 2:1 RCT, so that the entire type 3 hybrid study approximates a 1:1 RCT. But again, it's important that this choice is deemed acceptable from a clinical perspective, given unique circumstances of the existing data and other clinical considerations. The two main statistical issues identified in Sect. 3 still apply: (1) how to ensure that the nominal number of RWD patients does not exceed A, and (2) how to mitigate confounding bias so that the hybrid study can better approximate a traditional clinical study. The tools that can be used to address the first issue, namely, power prior and composite likelihood, have already been introduced (see Sect. 3.2). Just as in Sects. 2 and 3, the second issue is addressed using PS methodology, which is implemented via a two-stage design that is almost the same as in Sect. 3. The only difference is that the TCS is now a 2:1 RCT and therefore has two arms. As will be seen in the following sections, this difference does not add much complexity to the two-stage design. In particular, the definition of PS given in Sect. 3 does not need to be changed.

4.2 The Balancing Property of Propensity Score in Type 3 Hybrid Studies

The PS $e(X)$ for a patient with a vector X of observed baseline covariates in a type 3 hybrid study is the conditional probability of being in the TCS ($Z = 1$) rather than the RWD ($Z = 0$) given the vector of baseline covariates X:

$$e(X) = \Pr(Z = 1 | X).$$

Let T be the indicator variable with $T = 1$ indicating treated patients and $T = 0$ indicating control patients. Chen et al. [24] show that the PS as defined above has the following balancing property:

$$T, Z \perp X \mid e(X).$$

Note that, when RWD is leveraged to augment the control group of an RCT, we have

$$Z = 0 \Rightarrow T = 0,$$

i.e., if $Z = 0$ then $T = 0$. Hence the vector (T, Z) only takes on three values, $(0, 0)$, $(0, 1)$ and $(1, 1)$, which correspond to the three groups of RWD control patients, TCS control patients, and TCS treated patients, respectively. With this in mind, what the balancing property of Chen et al. [24] says is that, among patients with the same value of PS, the distribution of observed covariates is the same in the above three groups of patients. This property is the foundation of the PS-integrated approach for type 3 hybrid studies. It means that, when PS stratification is conducted, we can leverage RWD to estimate the control group parameter within each PS stratum in the same manner as in a type 2 hybrid study due to the balance between the TCS control group and the RWD control group. Furthermore, we can estimate the treatment effect within each stratum as in a type 1 hybrid study due to the balancing property between the TCS treated group and the combined TCS control and RWD control groups. In the next subsection, we provide a detailed description of the propensity score-integrated approach for type 3 hybrid studies through a numerical example.

4.3 The Propensity Score-Integrated Approach for Type 3 Hybrid Studies

Continuing with the example in Sect. 4.1, suppose the PS-integrated approach is considered appropriate for this trial and is implemented in the two-stage design framework. The two-stage design is described in detail below, which is the same whether Bayesian or frequentist inference is planned. For the first design stage,

the primary endpoint is specified to be the binary clinical outcome variable of the occurrence of adverse event(s) within 1 year. The primary endpoint hypotheses are

$$Ho : \mu = 0 \; vs. Ha : \mu \neq 0,$$

where $\mu = \theta^{(1)} - \theta^{(0)}$ is the treatment effect in the RCT with respect to the 1 year adverse event rates $\theta^{(0)}$ (control group) and $\theta^{(1)}$ (treated group). A total of 17 baseline covariates are identified as potential confounders. It is confirmed that these covariates and the outcome variable are collected in the registry referred to in Sect. 4.1. For sample size determination, the expected $\theta^{(0)}$ and $\theta^{(1)}$ are assumed to be 0.29 and 0.165, respectively. At the significance level of 0.05, a power of 80% would be achieved for an RCT with a total of approximately 354 patients at 1:1 randomization ratio. For the planned type 3 hybrid study, 267 patients are to be enrolled into the TCS part at the randomization ratio of 2:1 and a nominal 87 control patients are to be leveraged from the registry. Finally, an independent statistician is identified who is to perform the PS design at the second design stage. Thus, the first design stage of the two-stage design is complete. The main elements of the first design stage are displayed in Table 7.

The second design stage starts when all the 267 patients have been enrolled into the TCS part of the study, which, in this example, includes 183 patients randomly assigned to the treated group and 84 patients randomly assigned to the control group, and when all the eligible patients potentially to be leveraged (totaling 1570) are obtained from the registry. Note that the ratio of the patient numbers is not exactly 2:1, as is often the case in real trials. Using the 267 TCS patients and 1570 RWD patients, PSs are estimated, by the independent statistician identified in the first stage who is blinded to outcome data, via logistic regression with all 17 baseline covariates included in their linear terms as independent variables, and the indicator variable for TCS versus RWD patients as the dependent variable. Excluding those RWD patients whose PSs are not in the range of that of the TCS patients, 1192 RWD patients are selected. Five PS strata are formed for all the patients (267 + 1192) with each stratum containing near equal number of TCS patients (see Table 8 for within stratum patient numbers). Balance is assessed for each covariate and is considered adequate. Then, overlapping coefficients of the

Table 7 Main elements of the first design stage

Primary outcome: probability of adverse event within 1 year (θ)
Hypotheses: $Ho: \; \mu = 0 \;\; vs. \;\; Ha : \mu \neq 0$
Significance level: 0.05 two-sided/posterior probability threshold: 0.975
Number of baseline covariates considered: 17
Propensity score stratification planned for study design and outcome analysis
Independent statistician identified
Sample size for the TCS: 267 (2:1 randomization)
Nominal sample size for RWD control group patients: 87

Table 8 Sample sizes in each PS stratum

	1	2	3	4	5	Total
TCS (n)	54	53	53	53	54	267
Treated	41	28	39	36	39	183
Control	13	25	14	17	25	84
RWD (n)	332	270	233	201	156	1192

Table 9 Overlapping coefficient, standardized overlapping coefficient, nominal number of patients to be borrowed, and power parameter (or composite likelihood exponent) in each stratum

	1	2	3	4	5	Total
Overlapping coefficient	0.85	0.81	0.72	0.74	0.82	
Standardized overlapping coefficient	22%	20%	20%	18%	20%	100%
Patients borrowed (=90 × Std. Overlap Coef.)	19	17	17	16	18	87
α_s (or λ_s) (=Patients Borrowed/ RWD (n))	0.06	0.06	0.08	0.08	0.11	

propensity score distributions defined in Inman and Bradley Jr. [26] are calculated for all strata, and the nominal number of 87 RWD patients being leveraged are allocated to each stratum in proportion to the overlapping coefficients (see Table 9).

After clinical outcomes have been observed from all the patients, the statistical inference is conducted. For the Bayesian approach, the posterior probability of $\mu < 0$ is 97.9%, which meets the study success criterion. For the frequentist approach, the estimate of the overall treatment effect $\hat{\mu}$ is 0.18 with SE equal to 0.04. The p-value from the Wald test is 0.01, which indicates that the adverse event rate of the investigational device is statistically significantly lower than that of the control.

4.4 More Information on Outcome Analysis

The example in the previous subsection focused on the two-stage outcome-free design for the PS-integrated approach with a very brief description of the outcome analysis. In this subsection, we make some comments on the statistical inference for outcome analysis. Corresponding to the PS stratification, the parameters $\theta^{(0)}$and $\theta^{(1)}$ both branch out into S independent stratum specific parameters $\theta_s^{(0)}$and $\theta_s^{(1)}$ $s = 1, \ldots, S$ (S being the number of strata). Accordingly, μ branches out into $\mu_s = \theta_s^{(1)} - \theta_s^{(0)}$. Just as in Sect. 3, true PS is used to define stratum-specific parameters, while estimated PS is used for the Bayesian and frequentist statistical inference described below (see the discussion at the end of Sect. 3). For Bayesian inference [25], the idea is to apply the power prior method for $\theta_s^{(0)}$. Per the power prior method, the prior distribution of $\theta_s^{(0)}$ is

$$\pi_s^{(0)}\left(\theta_s^{(0)}\right) \propto \left[L\left(\theta_s^{(0)} | \boldsymbol{D}_{s,\mathbf{0}}^{(\mathbf{0})}\right)\right]^{\alpha s} \pi_{s,0}^{(0)}\left(\theta_s^{(0)}\right)$$

where $\boldsymbol{D}_{s,\mathbf{0}}^{(\mathbf{0})}$represents data collected on the RWD control patients in stratum s and $\pi_{s,0}^{(0)}$is the initial prior for $\theta_s^{(0)}$.The algorithm for obtaining α_s is illustrated in the previous section. The posterior distribution of $\theta_s^{(0)}$ is

$$\pi_{s,1}^{(0)}\left(\theta_s^{(0)}|\boldsymbol{D}_{s,\mathbf{1}}^{(\mathbf{0})}\right) \propto \left[L\left(\theta_s^{(0)}|\boldsymbol{D}_{s,\mathbf{1}}^{(\mathbf{0})}\right)\right] \pi_s^{(0)}\left(\theta_s^{(0)}\right),$$

where $\boldsymbol{D}_{s,\mathbf{1}}^{(\mathbf{0})}$represents data collected on the TCS control group patients in stratum s. The posterior distribution of $\theta_s^{(1)}$ is

$$\pi_{s,1}^{(1)}\left(\theta_s^{(1)}|\boldsymbol{D}_{s,\mathbf{1}}^{(\mathbf{1})}\right) \propto \left[L\left(\theta_s^{(1)}|\boldsymbol{D}_{s,\mathbf{1}}^{(\mathbf{1})}\right)\right] \pi_s^{(1)}\left(\theta_s^{(1)}\right)$$

where $\boldsymbol{D}_{s,\mathbf{1}}^{(\mathbf{1})}$represents data collected on TCS treated group patients and $\pi_s^{(1)}$is the prior for $\theta_s^{(1)}$. After the posterior distributions of $\theta_s^{(0)}$and $\theta_s^{(1)}$are obtained, the posterior distribution of μ_s is available. Finally, the posterior distribution of μ is obtained via

$$\mu = \frac{\sum_{s=1}^{S} w_s \mu_s}{\sum_{s=1}^{S} w_s}$$

where *ws* is the number of TCS patients in stratum s. For frequentist inference, the composite likelihood for $\theta_s^{(0)}$ is

$$L\left(\theta_s^{(0)}\right) = \prod_{i \in C} f\left(y_i; \theta_s^{(0)}\right) \prod_{j \in R} f\left(y_j; \theta_s^{(0)}\right)^{\lambda_s}$$

where y represents endpoint data collected on a patient, C is the index set for TCS control group patients and R is the index set for RWD patients; λ_s is obtained in the same way α_s is obtained. Point estimate for $\theta_s^{(0)}$ can be obtained by maximizing $L\left(\theta_s^{(0)}\right)$, and its variance can be obtained via the methods described in Chen et al. [24]. Statistical inference for $\theta_s^{(1)}$is carried out based on the ordinary likelihood function. Finally, point estimate for μ is

$$\hat{\mu} = \frac{\sum_{s=1}^{S} w_s \left(\hat{\theta}_s^{(1)} - \hat{\theta}_s^{(0)}\right)}{\sum_{s=1}^{S} w_s}$$

and its variance can be obtained by the fact that the summands are independent.

4.5 Discussion

While the idea of PS has been around since the early 1980s, its application in regulatory studies for the evaluation of the safety and effectiveness of medical products only began in the twenty-first century. Recently, the concept of PS has been expanded so that it can be used not only for causal inference in observational studies but also for leveraging RWD to augment a traditional single-arm study or to augment a traditional RCT. Regarding the latter two applications, more research has been conducted to go beyond the basic data structure and data type covered in this chapter. Chen et al. [27] consider time-to-event endpoints. Li et al. [28] discuss augmenting both arms of the RCT by leveraging RWD. Lu et al. [25, 29] deal with leveraging multiple RWD sources. These extensions and variations all have the same design elements that fit into the templates provided in the previous sections. Specifically, they are all underpinned by the PS methodology and follow the same outcome-free two-stage design framework illustrated above with examples. The down-weighting of leveraged RWD patients may be achieved with Bayesian or frequentist methods. In either case, with the methods introduced in this chapter, the weights assigned to individual patients do not depend on the outcomes of the patients. This feature, outcome-free study design, is very important to maintaining the integrity and objectivity of the study, thereby strengthening the interpretability of study results. The concept of outcome-free design within the two-stage design framework [12] was introduced about 10 years ago and is often used in medical device regulatory studies using all kinds of PS-based methods. In conclusion, we hope that this chapter can serve as a handy reference for all practitioners concerned with the development of medical products in planning, designing, and analyzing a hybrid study.

Disclaimer This chapter reflects the views of the authors and should not be construed to represent FDA's views or policies.

References

1. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. *Biometrika.***70**, 41–55 (1983).
2. Rosenbaum, P.R., Rubin, D.B.: Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*. **79**, 516–524 (1984).
3. Imbens, G., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press, New York (2015).
4. Li, H., Mukhi, V., Lu, N., Xu, Y. & Yue, L.Q.: A note on good practice of objective propensity score design for premarket nonrandomized medical device studies with an example. *Statistics in Biopharmaceutical Research*. **8**, 282–286 (2016).
5. Austin, P.C.: A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*. **33**, 1057–1069 (2014).
6. Stuart, E.A.: Matching methods for causal inference: A review and a look forward. *Statistical Science*. **25**, 1–21 (2010).

7. Li, H., Wang, C., Chen, W.C., Lu, N., Song, C., Tiwari, R., Xu, Y. & Yue, L.Q.: Estimands in observational studies: Some considerations beyond ICH E9 (R1). *Pharmaceutical Statistics*. **21**, 835–844 (2022).
8. Li, F., Morgan, K.L., Zaslavsky, A.M.: Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*. **113**, 390–400 (2018).
9. Austin, P.C.: Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*. **35**, 5642–5655 (2016).
10. Austin, P.C., Stuart, E.A.: Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*. **34**, 3661–3679 (2015).
11. Austin, P.C.: Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*. **28**, 3083–3107 (2009).
12. Yue, L.Q., Lu, N. and Xu, Y.: Designing pre-market observational comparative studies using existing data as controls: challenges and opportunities. J. Biopharm. Stat. **24**, 994–1010 (2014).
13. Langenskold, S., Rubin, D.B.: Outcome-free design of observational studies: peer influence on smoking. *Les Annales d'Economie et de Statistique*. **91/92**, 107–125 (2008).
14. Rubin, D.B.: For objective causal inference, design trumps analysis. *Annals of Applied Statistics*. **2**, 808–840 (2008).
15. Rubin, D.B.: Using propensity score to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*. **2**, 169–188 (2001).
16. Yue, L.Q., Campbell, G., Lu, N., Xu, Y., Zuckerman, B.: Utilizing National and International Registries to Enhance Pre-market Medical Device Regulatory Evaluation. *Journal of Biopharmaceutical Statistics*. **26**, 1136–1145 (2016).
17. Lu, N., Xu, Y., Yue, L.Q.: Good statistical practice in utilizing real world data in a comparative study for premarket evaluation of medical devices. *Journal of Biopharmaceutical Statistics*. **29**, 580–591 (1019).
18. Lu, N., Xu, Y., Yue, L.Q.: Some considerations on design and analysis plan on a nonrandomized comparative study utilizing propensity score methodology for medical device premarket evaluation. *Statistics in Biopharmaceutical Research.***12**, 155–163 (2020).
19. Chen, M-H., Ibrahim, J.G.: Power prior distribution for regression models. *Statistical Science*. **15**, 46–60 (2000).
20. Ibrahim, J. G., Chen, M.-H., Gwon, Y., Chen, F.: The Power Prior: Theory and Applications. *Statistics in Medicine*. **34**, 3724–3749 (2015).
21. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. *Statistica Sinica*. **21**, 5–42 (2011).
22. Wang, C., Li, H., Chen, W-C., Lu, N., Tiwari, R., Xu, Y., Yue, L.: Propensity Score-Integrated Power Prior Approach for Incorporating Real-World Evidence in Single-Arm Clinical Studies. *Journal of Biopharmaceutical Statistics.***29**, 731–748 (2019).
23. Wang, C., Lu, N., Chen, W-C., Li, H., Tiwari, R., Xu, Y., Yue, L.: Propensity Score-Integrated Composite Likelihood Approach for Incorporating Real-World Evidence in Single-Arm Clinical Studies. *Journal of Biopharmaceutical Statistics*. **30**, 495–507 (2020).
24. Chen, W-C., Wang, C., Li, H., Lu, N., Tiwari, R., Xu, Y., Yue, L.Q.: Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data. *Journal of Biopharmaceutical Statistics*. **30**, 508–520 (2020).
25. Lu, N., Wang, C., Chen, W-C., Li, H., Song, C., Tiwari, R., Xu, Y., Yue, L.Q.: Propensity score-integrated power prior approach for augmenting the control arm of a randomized controlled trial by incorporating multiple external data sources. *Journal of Biopharmaceutical Statistics*. **32**, 158–169 (2022).
26. Inman, H.F., Bradley Jr., E.L.: The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics-Theory and Methods*. **18**, 3851–3874 (1989).

27. Chen, W-C., Lu, N., Wang, C., Li, H., Song, C., Tiwari, R., Xu, Y., and Yue, L.Q.: Propensity Score-Integrated Approach to Survival Analysis: Leveraging External Evidence in Single-Arm Studies. *Journal of Biopharmaceutical Statistics*.**32**, 400–413 (2022).
28. Li, H., Chen, W-C., Wang, C., Lu, N., Song, C., Tiwari, R., Xu, Y., and Yue, L.Q.: Augmenting Both Arms of a Randomized Controlled Trial Using External Data: An Application of the Propensity Score-Integrated Approaches. *Statistics in Biosciences*, **14**, 79–89 (2022).
29. Lu, N., Wang, C., Chen, W-C., Li, H., Song, C., Tiwari, R., Xu, Y., Yue, L.Q.: Leverage multiple real-world data sources in single-arm medical device clinical studies. *Journal of Biopharmaceutical Statistics*. **32**, 107–123 (2022).

Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods

Yixin Fang

1 Introduction

In the opening article [1] of *Journal of Comparative Effectiveness Research*, the journal's founding editors pointed out that comparative effectiveness research (CER) "draws from the disciplines of health technology assessment, outcomes research, clinical epidemiology and implementation science, among others, to better answer the fundamental question 'which treatment will work best, in which patient, and under what circumstances?'"

Besides traditional randomized controlled clinical trials (RCTs), CER is looking at alternative real-world study designs [2], including:

- Pragmatic clinical trials such as pragmatic RCTs and large simple trials
- Observational studies such as case–control studies and cohort studies
- Non-randomized single-arm trials with external controls

In CER, causal inference plays an important role in deriving real-world evidence (RWE) from the analysis of real-world data (RWD) that are generated from real-world studies [3]. Research in causality has a long history, but in modern time, different disciplines (e.g., social science, economics, and statistics) took different paths. In this section, we provide a brief history of the development of causal inference in statistics before we move on to recent developments.

In *The Book of Why* [4], Pearl shared his regret that even the founding fathers of modern statistics such as Pearson hindered the development of causal inference in the community of statistics at the early stage of modern statistics. Since Neyman proposed the concept of potential outcomes in his 1923 Master's thesis and

Y. Fang (✉)
Data and Statistical Sciences, AbbVie, North Chicago, IL, USA
e-mail: yixin.fang@abbvie.com

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_11

Rubin in 1974 extended it into a general framework for causal inference in both interventional studies and non-interventional settings [5], we have seen more and more developments of causal-inference methods in the community of statistics. Counterfactual causal inference is the first one on the list of eight most important statistical ideas of the past 50 years selected by a 2021 paper [6]. Here we briefly review three milestones.

The first milestone is propensity-score (PS)-based methods developed by Rubin and colleagues, based upon a fundamental theorem proved in their 1983 paper [7]. The class of PS-based methods includes four methods: (1) matching, (2) stratification, (3) PS as covariate, and (4) weighting. The second milestone is generalized methods (G-methods) developed by Robins and colleagues in 1990s and 2000s, including three major methods: (i) g-formula, (ii) inverse probability of treatment weighting (IPTW), and (iii) G-estimation. Refer to their book [8] for a comprehensive review of G-methods. The third milestone is targeted learning developed by van der Laan and colleagues, starting with their first paper on targeted maximum likelihood estimation [9], leading to two books on targeted learning [10, 11].

The remaining of the chapter is organized as follows. There is a rich literature on reviewing and tutorials of causal inference methods, so we believe we cannot do better in providing another comprehensive review. Instead, in Sects. 2–4, we review some influential methods by making three binary choices: (a) conditional or marginal, (b) weighting or standardization, and (c) time-independent or time-dependent. In Sect. 5, we provide some discussion on the application of these methods to real-world studies with intercurrent events.

2 Conditional or Marginal

2.1 Propensity-Score Methods

We start with a simple point-exposure study, in which A is a binary exposure variable with $A = 1$ being the investigative treatment and $A = 0$ being the comparator (say, the standard of care), Y is an outcome variable that is either continuous or binary, and W is a list of covariates, which are believed to contain all the measured confounders along with effect modifiers. A directed acyclic graph (DAG) for this study is displayed in Fig. 1.

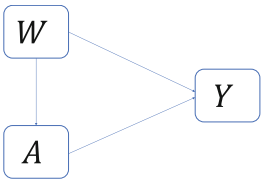

Fig. 1 A directed acyclic graph of a point-exposure study

We conduct causal inference to test the existence and estimate the magnitude of the relationship $A \rightarrow Y$, which is confounded by one back-door path [12], $A \leftarrow W \rightarrow Y$. The randomization feature in RCTs removes the arrow in $W \rightarrow A$, such that

$$A \perp\!\!\!\perp W, \tag{1}$$

leading to removing the confounding bias in the design stage. In non-randomized real-world studies, thanks to the following theorem in [7], we are able to achieve the desirable independence between A and W conditional on the PS function, $e(w) = P(A = 1|W = w)$.

Theorem 1 (Theorem 1 in [7]) *Treatment assignment and the observed covariates are conditionally independent given the propensity score, that is,*

$$A \perp\!\!\!\perp W|e(W). \tag{2}$$

There are four different PS methods based on the above theorem [13]: (1) matching on the PS, (2) stratification on the PS, (3) covariate adjustment using the PS, and (4) IPTW using the PS. Although the validity of all these four methods depends on whether or not PS function $e(w)$ is estimated consistently, in order to understand the pros and cons among them, it is helpful to understand the "conditional" thinking behind PS methods (1)–(3) and the "marginal" thinking behind PS method (4).

The first method, matching on the PS, attempts to mimic an RCT, creating a matched subset conditional on which A and W are independent. The second method, stratification on the PS, stratifies the dataset into several subsets, such that conditional on each subset, A and W are approximately independent. The third method, covariate adjustment using the PS, specifies a regression model of Y against A and $e(W)$, modeling the conditional relationship between Y and A given $e(W)$.

Unlike the first three PS methods that take the conditional thinking, IPTW takes the marginal thinking, creating two pseudo-populations, with one pseudo-population in which all the subjects were treated by $A = 1$ and the other pseudo-population in which all the subjects were treated by $A = 0$. Furthermore, of these four PS methods, IPTW is the only one that can be generalized to methods that can adjust for time-dependent confounding. Hence, we can consider IPTW as the intersection of the class of PS methods and the class of G-methods. IPTW is often discussed with marginal structural models (MSMs) [14], where we use MSMs to define an estimand and use the IPTW method to estimate the estimand.

2.2 Marginal Structural Models

Continue the above point-exposure study. Let $Y^{a=1}$ denote a subject's outcome if treated by the investigative treatment and $Y^{a=0}$ denote the outcome if treated by the comparator. For continuous outcome or 0–1 binary outcome, we can consider the following marginal structural models [14]:

$$E(Y^a) = \alpha + \beta a, \tag{3}$$

which are marginal models because they model the marginal distributions of potential outcomes $Y^{a=1}$ and $Y^{a=0}$ rather than the joint distribution, are structural models because they model the potential outcomes rather than the observed outcomes, and are saturated models because two unknown quantities ($E(Y^1)$ and $E(Y^0)$) are modeled by two parameters (α and β). Note that $\beta = E(Y^1) - E(Y^0)$ for continuous outcome or $\beta = P(Y^1 = 1) - P(Y^0 = 1)$ for binary outcome is the average treatment effect (ATE). In addition, for binary outcome, we may consider different MSMs, for example, $\text{logit}P(Y^a = 1) = \alpha' + \beta' a$, where β' is the log odds ratio between $Y^1 = 1$ and $Y^0 = 1$. Overall, the parameters in these MSMs can be estimated using the IPTW estimators [14].

Because of potential confounding, linear regression analysis of $Y \sim A$ for continuous outcome is biased in estimating β, and logistic regression analysis of $Y \sim A$ for binary outcome is biased in estimating β'. On the other hand, assuming that there is no unmeasured confounding, using weight $\omega = A/e(W) + (1-A)/(1-e(W))$, weighted linear regression analysis and weighted logistic regression analysis are unbiased in estimating β and β', respectively.

The approach of MSM and IPTW can be generalized to analyze studies with multi-level treatment, studies with continuous treatment doses, and studies with time-dependent confounding [14].

3 Weighting or Standardization

There is a rich literature of causal inference methods beyond the PS methods, which are well reviewed in several monographs (e.g., [8, 10, 11, 15, 16]). It is not our intention to review these recent developments comprehensively. Instead, as in [17], in this section, we describe two basic strategies, the weighting strategy and the standardization strategy.

We continue the above point-exposure study, which generates a dataset consisting of $O_i = (W_i, A_i, Y_i)$, $i = 1, \cdots, n$. In (3), the causal quantity is defined as parameter β in the MSM. Here we define the causal quantity of interest as the following ATE directly:

$$\theta^* = E(Y^1) - E(Y^0). \tag{4}$$

In order to construct an estimand, we assume three assumptions [8]: the consistency assumption, the no-unmeasured-confounder (NUC) assumption, and the positivity assumption,

$$\begin{aligned}\text{Consistency}: &\ Y = AY^1 + (1-A)Y^0,\\ \text{NUC}: &\ Y^a \perp\!\!\!\perp A|W, a = 0, 1,\\ \text{Positivity}: &\ P(A = a|W = w) > 0, a = 0, 1; w \in supp(W).\end{aligned}$$

In addition, we may need either or both of the following two functions, the PS function from the propensity-score model of $A \sim W$,

$$g(a|w) = P(A = a|W = w), \tag{5}$$

and the regression function from the outcome-regression model of $Y \sim A + W$,

$$Q(a, w) = E(Y|A = a, W = w). \tag{6}$$

3.1 The Weighting Strategy

3.1.1 Estimand

Under those three identifiability assumptions, we have

$$\begin{aligned}
&E\left\{\frac{I(A=a)}{P(A=a|W)}Y\right\} \quad \because \text{the positivity assumption}\\
&= E\left[E\left\{\frac{I(A=a)}{P(A=a|W)}Y\middle|W\right\}\right] \quad \text{by the double expectation formula}\\
&= E\left[E\left\{\frac{I(A=a)}{P(A=a|W)}Y^a\middle|W\right\}\right] \quad \because \text{the consistency assumption}\\
&= E\left[E\left\{\frac{I(A=a)}{P(A=a|W)}\middle|W\right\}E\{Y^a\middle|W\}\right] \quad \because \text{the NUC assumption}\\
&= E\left[E\{Y^a|W\}\right] \quad \because E(I(A=a|W)=P(A=a|W)\\
&= E(Y^a). \quad \text{by the double expectation formula}
\end{aligned}$$

Hence, we have

$$\theta^* = E(Y^1) - E(Y^0) = E\left\{\frac{I(A=1)}{P(A=1|W)}Y\right\} - E\left\{\frac{I(A=0)}{P(A=0|W)}Y\right\}. \tag{7}$$

This leads to the following estimand,

$$\theta = E\left\{\frac{I(A=1)}{g(1|W)}Y\right\} - E\left\{\frac{I(A=0)}{g(0|W)}Y\right\}. \tag{8}$$

We call this strategy of defining estimand as the weighting strategy because it uses the inverse of $g(a|w) = P(A = a|W = w)$ as the weights in the definition of the estimand. Using these weights, it creates two pseudo-populations: one pseudo-population in which all the subjects would have been treated by $a = 1$, leading to the first term in the right-hand side of (8), and the other pseudo-population in which all the subjects would have been treated by $a = 0$, leading to the second term.

3.1.2 Initial Estimator

If we obtain an estimator of the PS function, $\widehat{g}(a|w)$, using some statistical model, say logistic regression model, then we can obtain an initial estimator of θ, the IPTW estimator,

$$\widehat{\theta}_{IPTW} = \frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=1)}{\widehat{g}(1|W_i)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=0)}{\widehat{g}(0|W_i)}Y_i. \tag{9}$$

3.1.3 Doubly Robust Estimator

Although initial estimator $\widehat{\theta}_{IPTW}$ is asymptotically consistent if the model of $A \sim W$ is correctly specified in the construction of $\widehat{g}(a|w)$, it is not asymptotically efficient. Therefore, it is desirable to develop an augmented estimator that is asymptotically efficient under some model specification requirements.

According to semi-parametric efficiency theory (e.g., [10, 18]), the efficient score of estimating θ is given by

$$D(\theta; g, Q) = \frac{2A-1}{g(A|W)}[Y - Q(A, W)] + Q(1, W) - Q(0, W) - \theta. \tag{10}$$

Based on this efficient score function, we can apply the estimating equation approach to obtain an augmented estimator of θ, $\widehat{\theta}_{AIPTW}$, such that

$$\sum_{i=1}^{n} D(\widehat{\theta}_{AIPTW}; \widehat{g}, \widehat{Q})(W_i, A_i, Y_i) = 0, \tag{11}$$

where estimators $\widehat{g}$ and $\widehat{Q}$ are obtained by specifying some models of $A \sim W$ and $Y \sim A + W$, respectively.

Thus, by solving the estimating equation (11), we obtain the following augmented inverse probability of treatment (AIPTW) estimator [19]:

$$\widehat{\theta}_{AIPTW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{I(A_i=1)}{\widehat{g}(1|W_i)}Y_i - \frac{I(A_i=1)-\widehat{g}(1|W_i)}{\widehat{g}(1|W_i)}\widehat{Q}(1,W_i)\right)$$
$$-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{I(A_i=0)}{\widehat{g}(0|W_i)}Y_i - \frac{I(A_i=0)-\widehat{g}(0|W_i)}{\widehat{g}(0|W_i)}\widehat{Q}(0,W_i)\right). \quad (12)$$

According to the theory of estimating equations [19], $\widehat{\theta}_{AIPTW}$ is a doubly robust estimator; that is, it is asymptotically consistent if either the propensity-score model or the outcome-regression model is correctly specified, and it is asymptotically efficient if both models are correctly specified.

3.2 The Standardization Strategy

3.2.1 Estimand

Under those three identifiability assumptions, we have

$$\begin{aligned}
&E(Y^a)\\
&= E\{E(Y^a|W)\} \quad \text{by the double expectation formula}\\
&= E\{E(Y^a|A=a,W)\} \quad \because \text{the NUC assumption and positivity assumption}\\
&= E\{E(Y|A=a,W)\}. \quad \because \text{the consistency assumption}
\end{aligned}$$

Hence, we have

$$\theta^* = E(Y^1) - E(Y^0) = E_W\{E(Y|A=1,W) - E(Y|A=0,W)\}. \quad (13)$$

This leads to the following estimand:

$$\theta = E_W\{Q(1,W) - Q(0,W)\} = \int [Q(1,w) - Q(0,w)]dP_W(w), \quad (14)$$

where $P_W(w)$ is the probability distribution of W in the study population.

We call this strategy of defining estimand as the standardization strategy because it uses the standardization expectation over the marginal distribution of W of the study population, $E_W\{Q(a,W)\}$, for $a = 0, 1$.

3.2.2 Initial Estimator

If we obtain an estimator of the regression function, $\widehat{Q}(a, w)$, using some regression model, say generalized linear model, then we can obtain an initial estimator of θ,

$$\widehat{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}[\widehat{Q}(1, W_i) - \widehat{Q}(0, W_i)] = \int [\widehat{Q}(1, w) - \widehat{Q}(0, w)]d\widehat{P}_W(w), \quad (15)$$

where $\widehat{P}_W$ is the empirical distribution of W, which is a non-parametric maximum likelihood estimator of P_W. Following [10], we call the above estimator as maximum likelihood estimator (MLE). To understand this, let $\theta = \theta(Q, P_W)$. If $\widehat{Q}$ and $\widehat{P}_W$ are MLEs of Q and P_W, respectively, then $\widehat{\theta}_{MLE} = \theta(\widehat{Q}, \widehat{P}_W)$ is MLE of $\theta = \theta(Q, P_W)$.

3.2.3 Doubly Robust Estimator

Although initial estimator $\widehat{\theta}_{MLE}$ is asymptotically consistent if the model of $Y \sim A + W$ is correctly specified in the construction of $\widehat{Q}(a, w)$, it may not be asymptotically efficient. Therefore, it is desirable to develop a targeted estimator that is asymptotically efficient.

The efficient score of estimating $\theta = \theta(Q, P_W)$ in (10) can be written as $D(Q, P_W, g)(W, A, Y)$, which equals

$$\frac{2A - 1}{g(A|W)}[Y - Q(A, W)] + Q(1, W) - Q(0, W) - \theta(Q, P_W). \quad (16)$$

Based on this efficient score function, [9] develops the targeted learning technique to obtained estimators $(\widehat{Q}^*, \widehat{P}^*_W, \widehat{g}^*)$ such that

$$\sum_{i=1}^{n} D(\widehat{Q}^*, \widehat{P}^*_W, \widehat{g}^*)(W_i, A_i, Y_i) = 0, \quad (17)$$

where $\widehat{P}^*_W = \widehat{P}_W$, the empirical estimator of P_W, and $\widehat{g}^*$ and $\widehat{Q}^*$ are some updated estimators of initial estimators $\widehat{g}$ and $\widehat{Q}$, respectively. Thus, we can construct the targeted maximum likelihood estimator (TMLE),

$$\widehat{\theta}_{TMLE} = \theta(\widehat{Q}^*, \widehat{P}_W) = \int [\widehat{Q}^*(1, w) - \widehat{Q}^*(0, w)]d\widehat{P}_W(w). \quad (18)$$

3.3 Implementation and Comparison

Consider the implementation of the aforementioned four estimators: IPTW, AIPTW, MLE, and TMLE. We can use SAS procedure "CAUSALTRT" to implement $\widehat{\theta}_{IPTW}$, $\widehat{\theta}_{MLE}$, and $\widehat{\theta}_{AIPTW}$, along with their statistical inferences. Please see the following skeleton of the SAS procedure:

```
PROC CAUSALTRT;
MODEL outcome = covariate_1 covariate_2 ... ;
PSMODEL treatment = covariate_1 covariate_2 ...;
RUN;
```

In the above SAS procedure, there are "PSMODEL" and "MODEL" statements: (1) if only a generalized linear model (GLM) of $A \sim W$ is specified in the "PSMODEL" statement, it implements $\widehat{\theta}_{IPTW}$, (2) if only a GLM of $Y \sim A + W$ is specified in the "MODEL" statement, it implements $\widehat{\theta}_{MLE}$, and (3) if two GLM models are specified in the "PSMODEL" and "MODEL" statements, respectively, it implements $\widehat{\theta}_{AIPTW}$.

Furthermore, we can consider flexible models other than GLM (say, super learner [20]) to obtain initial estimators $\widehat{Q}$ and $\widehat{g}$ to improve the chance of consistency in estimating functions Q and g. For this aim, we can use R function "tmle" in R package "tmle" [21] to implement $\widehat{\theta}_{TMLE}$, along with its standard error for conducting statistical inference. Please see the following skeleton of the R function:

```
tmle(Y, A, W,
  Q.SL.library = c("SL.glm", "tmle.SL.dbarts2", "SL.glmnet"),
  g.SL.library = c("SL.glm", "tmle.SL.dbarts.k.5", "SL.gam"),
  family = "gaussian", ...)
```

In the above R function, we see that we adopt the same set of notations for variable names and function names in this chapter (e.g., Y, A, W, g, Q) from the R package "tmle," which makes it easy for us to plug in values into the arguments. For example, the "Q.SL.library" argument allows us to specify a flexible super learner model for the Q function, with a default library consisting of generalized linear model (glm), discrete Bayesian additive regression tree (dbart), and glm model regularized by elastic net (glmnet), while the "g.SL.library" argument allows us to specify a super learner model for the g function, with a default library consisting of glm, dbart, and generalized additive model (gam). Besides there default options, we can prespecify other options for the super learner libraries, including highly adaptive lasso. In addition, the "family" argument can take on default value "gaussian" for continuous outcome and other value "binomial" for binary outcome.

Chapter 6 of [10] provides both theoretical comparisons and numerical comparisons (extensive simulations and case studies) between these four methods. Here we only summarize some comparisons briefly. First, AIPTW and TMLE are doubly robust versions of IPTW and MLE, respectively. Second, AIPTW relies on parametric modeling of Q and g, while TMLE allows for flexible modeling of Q and g using super learner. Third, MLE and TMLE are plug-in estimators, which

are more stable than the weighted estimators. Fourth, all the four methods are G-methods, which can be generalized to analyze longitudinal data with time-dependent confounding.

4 Time-Independent or Time-Dependent

In the above point-exposure study, the treatment status is determined at a single time (time zero) for all the subjects and the treatment effect does not need to make references to the time at which treatment occurs [8]. On the other hand, in longitudinal studies with time-dependent treatments or intercurrent events, we need to incorporate time explicitly [8].

Chapter "Personalized Medicine with Advanced Analytics" of this book will review statistical methods for personalized medicine and dynamic treatment regimes. In this chapter, we focus on longitudinal studies with static treatment regimes and intercurrent events.

Assume that there is one longitudinal study starting with baseline $t = 0$, along with follow-up visits, $t = 1, \cdots, T$. Assume that the primary endpoint Y is the outcome variable at the final visit T. Let $\overline{A} = (A_0, \cdots, A_{T-1})$ be the actually received treatment sequence and $\overline{A}_t = (A_0, \cdots, A_t)$ be the treatment up to t, $t = 0, \cdots, T-1$. Let W_0 be baseline covariates, W_t be the vector including time-dependent covariates and intermediate outcome, and $\overline{W}_t = (W_0, \cdots, W_t)$ be the vector consisting of all the observed history up to time t including baseline covariates, time-dependent covariates, and intermediate outcomes.

Let $\overline{a} = (a_0, \cdots, a_{T-1})$ be a given static treatment regime. At each time t, $a_t = 1$ stands for treated by the investigative treatment, 0 for the comparator, `NA` for treatment discontinuation, and 2 for some rescue medication. Two examples are $\overline{a} = \overline{1} = \text{rep}(1, T)$, which means the subject is initially treated by $a_0 = 1$ and throughout, and $\overline{a} = \overline{0} = \text{rep}(0, T)$, which means the subject is initially treated by $a_0 = 0$ and throughout.

Let $Y^{\overline{a}^0}$ be the potential outcome if the subject follows the static treatment regime $\overline{a}^0 = (a_0^0, \cdots, a_{T-1}^0)$. The population summary of $Y^{\overline{a}^0}$ is referred to as the value of $\overline{a}^0$ in [16]. For continuous or binary outcome variable, we define the value of $\overline{a}^0$ as

$$v^*(\overline{a}^0) = E\{Y^{\overline{a}^0}\}. \tag{19}$$

In order to construct an estimand for the evaluation of the value, $v^*(\overline{a}^0)$, we also need three identifiability assumptions [8], the consistency assumption,

$$Y^{\overline{a}^0} = Y \text{ if } \overline{A} = \overline{a}^0, \tag{20}$$

the static sequential exchangeability assumption (a.k.a., the NUC assumption),

$$Y^{\overline{a}^0} \perp\!\!\!\perp A_0 | W_0,$$
$$Y^{\overline{a}^0} \perp\!\!\!\perp A_t | (\overline{A}_{t-1}, \overline{W}_t), \text{ for } t = 1, \cdots, T-1, \tag{21}$$

and the positivity assumption,

$$P\left(\overline{A} = \overline{a}^0 | W_0 = w_0\right) > 0, \text{ for } w_0 \in supp(W_0). \tag{22}$$

Consider a longitudinal study that generates a dataset consisting of $O_i = (W_{0i}, A_{0i}, \cdots, W_{T-1,i}, A_{T-1,i}, Y_i), i = 1, \cdots, n$. In the following two subsections, we will describe four major estimators, IPTW, AIPTW, MLE, and TMLE, that are respectively generalized from those four G-estimators described in Sect. 3. For this aim, we define two series of functions.

Propensity-Score Modeling
Let $H_0 = W_0$ and $H_t = (\overline{W}_t, \overline{A}_{t-1})$ be the history up to t before making decision $A_t, t = 1, \cdots, T-1$. Define the PS functions from modeling $A_t \sim H_t$,

$$g_t(a|h_t) = P(A_t = a | H_t = h_t), t = 0, \ldots, T-1. \tag{23}$$

We can obtain an estimator of $g_t(a|h_t)$, $\widehat{g}_t(a|h_t)$, using some statistical model such as logistic regression model.

Outcome-Regression Modeling
We attempt to define regression functions from modeling $Y \sim A_t + H_t$, $t = 0, \ldots, T-1$. However, the outcome variable Y is measured after the final decision point $T-1$, which depends on decisions made between $t+1$ and $T-1$. Therefore, we should apply some special approach to define them. The most popular approach is the backward induction approach [16], which defines regression functions recursively from decision point $T-1$ to decision point 0.

At decision point $T-1$, define

$$Q_{T-1}(H_{T-1}, A_{T-1}) = E(Y | H_{T-1}, A_{T-1}), \tag{24}$$

which can be estimated using some regression model such as GLM, with its estimator denoted as $\widehat{Q}_{T-1}(h_{T-1}, a_{T-1})$. Note that $(h_{T-1}, a_{T-1}) = (\overline{w}_{T-1}, \overline{a}_{T-1})$. Next, define $\widetilde{Q}_{T-1}(H_{T-1}) = Q_{T-1}(H_{T-1}, a^0_{T-1})$, which is the expected outcome if the treatment at $T-1$ is consistent with the static treatment regime $\overline{a}^0$ at $T-1$ and which can be used as the model outcome variable at decision point $T-2$.

At decision point $t = T-2, \ldots, 1$, define

$$Q_t(H_t, A_t) = E(\widetilde{Q}_{t+1}(H_{t+1}) | H_t, A_t), \tag{25}$$

which can be estimated using some regression model such as GLM, with its estimator denoted as $\widehat{Q}_t(h_t, a_t)$. Note that $(h_t, a_t) = (\overline{w}_t, \overline{a}_t)$. Next, define

$\widetilde{Q}_t(H_t) = Q_t(H_t, a_t^0)$, which is the expected outcome if the treatments at decision points from t to $T-1$ are consistent with the static treatment regime $\overline{a}^0$ at decision points from t to $T-1$.

Finally, at decision point $t = 0$, define

$$Q_0(W_0, A_0) = E(\widetilde{Q}_1(H_1)|W_0, A_0), \tag{26}$$

which can be estimated using some regression model such as GLM, with its estimator denoted as $\widehat{Q}_0(w_0, a_0)$. Define $\widetilde{Q}_0(W_0) = Q_0(W_0, a_0^0)$, which is the expected outcome if the subject takes the static treatment regime $\overline{a}^0$ at all decision points from 0 to $T-1$.

4.1 The Weighting Strategy

4.1.1 Estimand

By the weighting strategy, we can define the corresponding estimand for the value of $\overline{a}^0$. That is, under those three identifiability assumptions, $\nu^*(\overline{a}^0)$ is equal to

$$\nu(\overline{a}^0) = E\left\{\frac{I[\overline{A} = \overline{a}^0]Y}{g_0(a_0^0|W_0)\prod_{t=1}^{T-1} g_t(a_t^0|\overline{W}_t, \overline{A}_{t-1})}\right\}, \tag{27}$$

where propensity-score functions g's are defined in (23).

4.1.2 Initial Estimator

If we obtain estimators of propensity-score functions, $\widehat{g}_t$, $t = 0, \cdots, T-1$, then we can obtain an initial estimator of $\nu(\overline{d})$,

$$\widehat{\nu}_{IPTW}(\overline{a}^0) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I[\overline{A}_i = \overline{a}^0]Y_i}{\widehat{g}_0(a_0^0|W_{0i})\prod_{t=1}^{T-1}\widehat{g}_t(a_t^0|\overline{W}_{ti}, \overline{A}_{t-1,i})}\right\}. \tag{28}$$

4.1.3 Double-Robust Estimator

According to semi-parametric efficiency theory (e.g., [11, 18]), the efficient score of estimating $\nu(\overline{a}^0)$ is given by

$$D(\nu(\overline{a}^0); P) = \sum_{t=0}^{T} D_t(\nu(\overline{a}^0); P), \tag{29}$$

where P is the true underlying distribution of observation O_i and

$$
\begin{aligned}
D_0(\nu(\overline{a}^0); P) &= Q_0(W_0, a_0^0) - \nu(\overline{a}^0), \\
D_t(\nu(\overline{a}^0; P); P) &= \frac{I[\overline{A}_{t-1} = \overline{a}_{t-1}^0]}{\prod_{s=0}^{t-1} g_s(a_s^0 | \overline{W}_s, \overline{A}_{s-1})}[Q_t(\overline{W}_t, \overline{A}_t) - Q_{t-1}(\overline{W}_{t-1}, \overline{A}_{t-1})], \\
& \qquad t = 1, \ldots, T-1, \\
D_T(\nu(\overline{a}^0; P); P) &= \frac{I[\overline{A}_{T-1} = \overline{a}^0]}{\prod_{t=0}^{T-1} g_t(a_t^0 | \overline{W}_t, \overline{A}_{t-1})}[Y - Q_{T-1}(\overline{W}_{T-1}, \overline{A}_{T-1})].
\end{aligned}
$$

Therefore, if we further obtain estimators of regression functions, $\widehat{Q}_t(h_t, a_t)$ (which can be rewritten as $\widehat{Q}_t(\overline{w}_t, \overline{a}_t)$), then we can obtain the following doubly robust estimator for $\nu(\overline{a}^0)$, by solving the estimating equation $D(\nu(\overline{a}^0); \widehat{P}) = 0$,

$$
\begin{aligned}
\widehat{\nu}_{AIPTW}(\overline{a}^0) = \frac{1}{n}\sum_{i=1}^{n} \Bigg\{ & \frac{I[\overline{A}_i = \overline{a}^0]Y_i}{\overline{g}_{T-1}(\overline{W}_{T-1,i})} + \left[1 - \frac{I[A_{0i} = a_0^0]}{\widehat{g}_0(a_0 | W_{0i})}\right]\widehat{Q}_0(W_{0i}, a_0^0) \\
& + \sum_{t=1}^{T-1}\left[\frac{I[\overline{A}_{t-1,i} = \overline{a}_{t-1}^0]}{\overline{g}_{t-1}(\overline{W}_{t-1,i})} - \frac{I[\overline{A}_{ti} = \overline{a}_t^0]}{\overline{g}_t(\overline{W}_{ti})}\right]\widehat{Q}_t(\overline{W}_{ti}, \overline{a}_t^0)\Bigg\}, \quad (30)
\end{aligned}
$$

where $\overline{g}_t(\overline{W}_{ti}) = \widehat{g}_0(a_0 | W_{0i}) \prod_{s=1}^{t} \widehat{g}_s(a_s^0 | \overline{W}_{si}, \overline{a}_{s-1}^0)$.

4.2 The Standardization Strategy

4.2.1 Estimand

By the standardization strategy, we can define the corresponding estimand for the value of $\overline{a}^0$. That is, under those three identifiability assumptions, $\nu^*(\overline{a}^0)$ is equal to

$$
\nu(\overline{a}^0) = E\{Q_0(W_0, a_0^0)\}, \quad (31)
$$

where regression function Q_0 is defined in (26).

4.2.2 Initial Estimator

If we obtain an estimator of $Q_0(w_0, a_0)$, $\widehat{Q}_0(w_0, a_0)$, then we can obtain the following estimator for $\nu(\overline{a}^0)$:

$$
\widehat{\nu}_{MLE}(\overline{a}^0) = \frac{1}{n}\sum_{i=1}^{n} \widehat{Q}_0(W_{0i}, a_0^0). \quad (32)
$$

4.2.3 Double-Robust Estimator

If we further obtain estimators of propensity-score functions, $\widehat{g}_t$, $t = 0, \cdots, T-1$, we can construct the corresponding doubly robust estimator. For this aim, we apply the backward induction approach. At each decision point $t = T-1, T-2, \ldots, 0$, we first obtain an initial estimator of regression function, $\widehat{Q}_t(\overline{w}_t, \overline{a}_t)$, then we update the initial estimator into $\widehat{Q}_t^*(\overline{w}_t, \overline{a}_t)$ via the targeted learning theory based on the efficient score $D_{t+1}(\nu(\overline{a}^0); P)$, where $\widehat{Q}_t^*(\overline{w}_t, \overline{a}_t)$ is on the least favorable submodel that passes through $\widehat{Q}_t(\overline{w}_t, \overline{a}_t)$. At the end, we obtain $\widehat{Q}_0^*(W_{0i}, a_0^0)$ and thus the doubly robust estimator,

$$\widehat{\nu}_{TMLE}(\overline{a}^0) = \frac{1}{n}\sum_{i=1}^{n} \widehat{Q}_0^*(W_{0i}, a_0^0). \tag{33}$$

4.3 Implementation and Comparison

Similar to Sect. 3.3, here we provide some brief comparison. First, these four methods are generalized from those four methods with the same names in Sect. 3. Second, AIPTW and TMLE are doubly robust versions of IPTW and MLE, respectively. Third, AIPTW relies on parametric modeling of Q_t's and g_t's, while TMLE allows for flexible modeling of Q_t's and g_t's using super learner. Fourth, MLE and TMLE are plug-in estimators, which are more stable than the weighted estimators.

In practice, we can use R package "DTR" [16] to implement $\widehat{\nu}_{IPTW}$, $\widehat{\nu}_{MLE}$, and $\widehat{\nu}_{AIPTW}$, along with their statistical inferences, by specifying GLMs for Q_t's and g_t's. We can use R package "ltmle," with "l" standing for "longitudinal," to implement $\widehat{\nu}_{TMLE}$, by specifying either GLMs or super learner for Q_t's and g_t's. Refer to [22] for a detailed description of R package "ltmle." In the below, we provide an example of using it to estimate the ATE for longitudinal studies with intercurrent events.

Assume that we are interested in estimating the following ATE:

$$\theta = \nu(\overline{1}) - \nu(\overline{0}), \tag{34}$$

which measures the treatment effect of the investigative static treatment regime $\overline{a}^0 = \overline{1}$ compared against the reference static treatment regime $\overline{a}^{0\prime} = \overline{0}$. In order to understand this estimand, we should envisage one hypothetical world in which all the patients follow $\overline{a}^0 = \overline{1}$ throughout the study and the other hypothetical world in which all the patients follow $\overline{a}^{0\prime} = \overline{0}$ throughout the study. That is, in the construction of estimand (34), we apply the hypothetical strategy of ICH E9(R1) [23] to handle intercurrent events (e.g., treatment discontinuation, treatment changing, and rescue medication).

Table 1 The structure of the dataset in one example

Argument	Variable names[a]
Baseline covariates	c("L0.a", "L0.b", "L0.c")
Lnodes[b]	c("L1.a", "L1.b")
Anodes	c("A0", "A1")
Cnodes	c("C0", "C1")
Ynodes	c("Y1", "Y2")

[a] The order of the variables in the dataset: data.frame(L0.a, L0.b, L0.c, A0, C0, L1.a, L1.b, Y1, A1, C1, Y2)

[b] L_t in the Lnodes is the same as W_t in the context

In order to estimate θ in (34), we define the censoring variable C_t, which is a factor variable with two levels, "uncensored" or "censored," at each time t, $t = 0, \cdots, T-1$. If for time t, while $A_s = A_0$ for $s = 0, \cdots, t$, an intercurrent event occurs between t and $t+1$, then $C_t = \cdots = C_{T-1} =$ "$censored$." Note that in this setting we consider the event that directly leads to censoring as the intercurrent event. For example, assume that an adverse event leads to treatment discontinuation, which directly leads to data censoring, and then we consider the treatment discontinuation as an intercurrent event.

To demonstrate the use of R function "ltmle," we look at one example where there are two follow-up visits ($T = 2$), three baseline covariates at $t = 0$ ("L0.a", "L0.b", "L0.c"), two time-dependent covariates at $t = 1$ ("L1.a", "L1.b"), treatment variable measured at $t = 0, 1$ ("A0", "A1"), censoring variable measured at $t = 0, 1$, and outcome variable measured at $t = 1, 2$ ("Y1", "Y2"). Note that the L-node variables form the time-dependent covariates W_t; that is, $W_t = L_t, t = 0, 1$. Table 1 displays the structure of the dataset to be defined in R.

Here is an excerpt of R codes presented in [22] used to implement the TMLE estimator in the above example, providing the point estimate and 95% confidence interval of θ in (34):

```
data <- data.frame(L0.a, L0.b, L0.c, A0, C0, L1.a, L1.b, Y1, A1,
                   C1, Y2)
Lnodes <- c("L1.a", "L1.b")
Anodes <- c("A0", "A1")
Cnodes <- c("C0", "C1")
Ynodes <- c("Y1", "Y2")
ltmle(data = data, Anodes = Anodes, Cnodes = Cnodes, Lnodes = Lnodes,
Ynodes = Ynodes,  survivalOutcome = NULL,
  abar = list(treament = c(1, 1), control = c(0,0)))
```

Here is a remark on how these methods can be extended to survival outcome (a.k.a., time-to-event outcome). In the above R function, "survivalOutcome = NULL" indicates that the outcome variable is either continuous variable or binary having single Ynodes. We set "survivalOutcome = FALSE" for binary outcome variable with multiple Ynodes. For survival outcome, we set "survivalOutcome = TRUE" to indicate that Y_t nodes are indicators of an event, and if Y_t at some time point t is 1, then $Y_s, s = t+1, \cdots, T-1$, should be 1.

5 Discussion

In this chapter, we briefly review some recent statistical development of causal inference methods beyond PS methods. Instead of providing a comprehensive review, we investigate three checkpoints, which may be helpful for guiding us to select an appropriate approach for any study at hand.

If we want to consider one of the four PS methods, then the first checkpoint is whether the conditional approaches (matching, stratification, PS as covariate) or the marginal approach (IPTW). IPTW is a G-method, which can be generalized from point-exposure studies to longitudinal studies.

If we want to consider one of the G-methods, then the second checkpoint is the weighting approaches (e.g., IPTW and AIPTW) or the standardization approaches (e.g., MLE and TMLE). AIPTW is the doubly robust version of IPTW and TMLE is the doubly robust version of MLE.

The third checkpoint is to consider the problem as a time-independent problem or a time-dependent problem. Every G-method has two versions, one simple version for time-independent problem and the other complex version for time-dependent problem. Therefore, all the four methods (IPTW, AIPTW, MLE, and TMLE) have versions for time-dependent problem.

We conclude the chapter with a brief discussion on how to apply these methods to studies with intercurrent events (ICEs). ICH E9(R1) defines ICEs as "events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest. It is necessary to address intercurrent events when describing the clinical question of interest in order to precisely define the treatment effect that is to be estimated." Therefore, we should specify how to handle ICEs in the definition of the estimand and then select an appropriate causal inference method to estimate the estimand.

There are five ICH E9(R1) strategies for handling ICEs: (1) hypothetical strategy, (2) treatment-policy strategy, (3) composite-variable strategy, (4) while-on-treatment strategy, and (5) principal-stratum strategy.

5.1 Hypothetical Strategy for ICEs in Estimand Definition

Applying this strategy, we envision a scenario in which ICEs would not occur and define the estimand of interest as in (34), comparing the treatment regime of taking $A_0 = 1$ throughout against the treatment regime of taking $A_0 = 0$ throughout. The methods described in Sect. 4.3 can be applied to estimate this estimand.

5.2 Treatment-Policy Strategy for ICEs in Estimand Definition

This strategy requires that we collect data even after the ICE occurrence. Applying this strategy, we can use the value of the outcome variable regardless of whether

or not the ICE occurs and define the estimand of interest as in (8) or (14). All the methods described in Sect. 3 can be applied to estimate this estimand without revising the definition of outcome variable.

5.3 *Composite-Variable Strategy for ICEs in Estimand Definition*

Applying this strategy, we need to revise the definition of outcome variable. The new outcome variable is a composite variable of the original outcome variable and the ICE occurrence, and the estimand of interest can be defined as in (8) or (14) with the new outcome variable. All the methods described in Sect. 3 can be applied to estimate this estimand using the new outcome variable.

5.4 *While-on-treatment Strategy for ICEs in Estimand Definition*

Applying this strategy, we need to revise the definition of outcome variable as well. The new outcome variable is a function of the outcome variable measured prior to the ICE occurrence and the time of ICE occurrence (e.g., the rate of change). The estimand of interest can be defined as in (8) or (14) with the new outcome variable. All the methods described in Sect. 3 can be applied to estimate this estimand using the new outcome variable.

5.5 *Principal-Stratum Strategy for ICEs in Estimand Definition*

Applying this strategy, as proposed by ICH E9(R1), "the target population might be taken to be the principal stratum in which an ICE event would occur. Alternatively, the target population might be taken to be the principal stratum in which an ICE would not occur." The estimand of interest can be defined as in (8) or (14), with the outer expectation taken over the principal stratum of interest. To estimate this estimand, we need to estimate the membership of the principal stratum. Then, all the methods described in Sect. 3 can be applied, considering the estimated principal stratum as the target population.

References

1. Greenfield, S., Rich, E.: Welcome to the journal of comparative effectiveness research. *Journal Of Comparative Effectiveness Research*. **1**, 1–3 (2012)

2. Framework for FDA's Real-World Evidence Program, https://www.fda.gov/media/120060/download
3. Fang, Y., Wang, H., He, W.: A statistical roadmap for journey from real-world data to real-world evidence. *Therapeutic Innovation & Regulatory Science*. **54**, 749–757 (2020)
4. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic books, New York (2018)
5. Rubin, D.: Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. **66**, 688–701 (1974)
6. Gelman, A., Vehtari, A.: What are the most important statistical ideas of the past 50 years?. *Journal Of The American Statistical Association*. **116**, 2087–2097 (2021)
7. Rosenbaum, P., Rubin, D.: The central role of the propensity score in observational studies for causal effects. *Biometrika*. **70**, 41–55 (1983)
8. Hernan, M., Robins, J: Causal Inference: What If. Boca Raton: Chapman & Hall/CRC (2020).
9. van der Laan, M.; Rubin, D. Targeted maximum likelihood learning. *The International Journal Of Biostatistics*. **2** (2006)
10. van der Laan, M., Rose, S.: Targeted Learning: Causal Inference for Observational and Experimental Data. Springer (2011)
11. van der Laan, M., Rose, S.: Targeted Learning in Data Science. Springer (2018)
12. Pearl, J.: Causality. Cambridge university press (2009)
13. Austin, P.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*. **46**, 399–424 (2011)
14. Robins, J., Hernan, M., Brumback, B.: Marginal structural models and causal inference in epidemiology. *Epidemiology*. **11** pp. 550–560 (2000)
15. Imbens, G., Rubin, D.: Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press (2015)
16. Tsiatis, A., Davidian, M., Holloway, S., Laber, E.: Dynamic Treatment Regimes: Statistical Methods for Precision Medicine. Chapman (2020)
17. Fang, Y. Two basic statistical strategies of conducting causal inference in real-world studies. *Contemporary Clinical Trials*. **99** pp. 106193 (2020)
18. Bickel, P., Klaassen, C., Ritov, Y., Wellner, J.: Efficient and Adaptive Estimation for Semiparametric models. Springer (1993)
19. Bang, H., Robins, J.: Doubly robust estimation in missing data and causal inference models. *Biometrics*. **61**, 962–973 (2005)
20. van der Laan, M., Polley, Hubbard, A.: Super learner. *Statistical Applications In Genetics And Molecular Biology*. **6** (2007)
21. Gruber, S., van der Laan, M.: tmle: An R package for targeted maximum likelihood estimation. *Journal Of Statistical Software*. **51** pp. 1–35 (2012)
22. Lendle, S., Schwab, J., Petersen, M., van der Laan, M.: ltmle: an R package implementing targeted minimum loss-based estimation for longitudinal data. *Journal Of Statistical Software*. **81** pp. 1–21 (2017)
23. ICH E9(R1) (2021): Statistical Principles for Clinical Trials; Addendum: Estimand and Sensitivity Analysis in Clinical Trials, https://www.fda.gov/media/148473/download

Innovative Hybrid Designs and Analytical Approaches Leveraging Real-World Data and Clinical Trial Data

Lisa V. Hampson and Rima Izem

1 Introduction

There are a variety of ways in which real-world data (RWD) can enhance clinical trials in hybrid designs and associated analytical methods. The different design strategies are illustrated in Fig. 1. Pragmatic randomized controlled trials are designed to address the question of whether an intervention works under usual conditions with some, or all, outcomes captured in routine care settings. These trials therefore inform point-of-care clinical decision-making with evidence that targets health-care systems and payers. The approaches shown in the middle row of Fig. 1 span a range of design options and are further illustrated in Figs. 2 and 3. We may integrate a conventional randomized controlled trial (RCT) with pragmatic design aspects to leverage RWD or remotely collected data outside of routine care on patients whilst preserving randomization. Alternatively, we may enrich patients in the RCT with real-world (RW) patients. Finally, incorporating external control data into the study design to contextualize the results of a single arm interventional trial can be a more ethical, feasible, or efficient way forward than conducting an RCT in certain settings. In this chapter, we refer to all the approaches illustrated in Fig. 1 as hybrid approaches, in the sense that they prospectively plan to incorporate RWD and clinical trial data in the evaluation of new medicines. More specifically, we first review the use of external controls to complement and augment clinical trials in Sect. 2. Then, we review approaches using RWD, in place of some more traditional methods of data capture in clinical trials, in pragmatic and decentralized randomized controlled trials in Sect. 3.

L. V. Hampson (✉) · R. Izem
Statistical Methodology, Novartis Pharma AG, Basel, Switzerland
e-mail: lisa.hampson@novartis.com

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_12

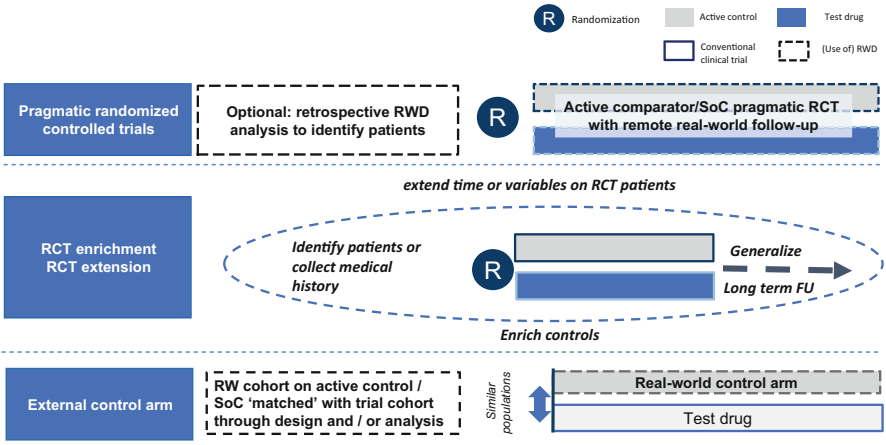

Fig. 1 Hybrid designs for clinical trials-RWD. Abbreviations: SoC = Standard of care; FU = Follow-up

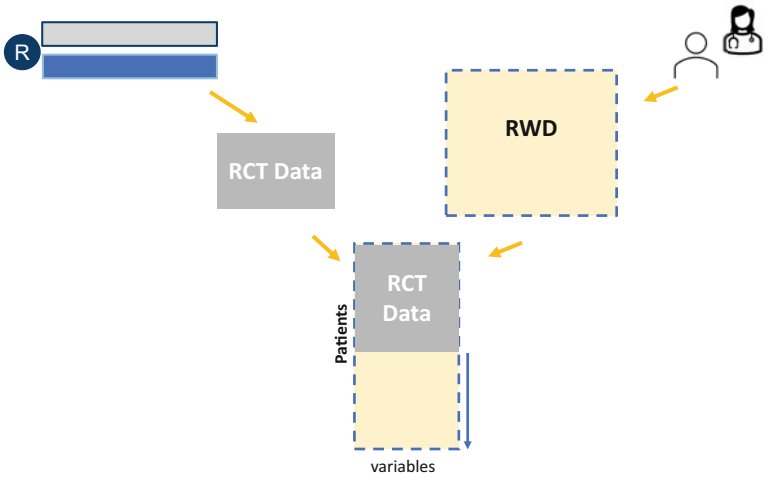

Fig. 2 RWD augmenting RCT data by adding more patients

2 Hybrid Designs and Analytical Approaches Leveraging Real-World External Controls and Clinical Trial Data

External controls are characterized by several features including their (a) source (other clinical trials or RWD); (b) type (patient-level or synthetic, where the latter are data generated from a synthesis of trial-external evidence on control); (c) level of the data (aggregate or individual patient data); and (d) timing relative to the new clinical trial (historical or concurrent). In this section, we will discuss

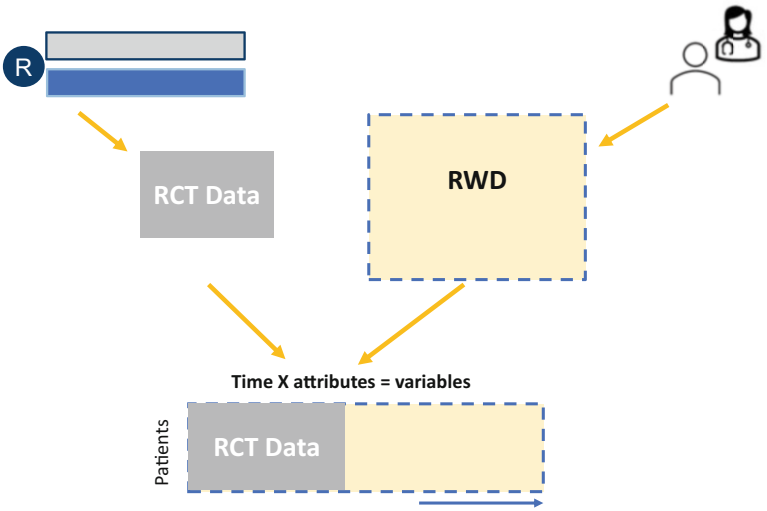

Fig. 3 RWD augmenting RCT data by adding more attributes on the same patients (e.g., through linkage)

how patient-level external controls can be leveraged to support the analysis and interpretation of a new clinical trial and how these uses of external controls, as well as prior uncertainty about their relevance, can be reflected in the design of the trial. Discussion of simulated patient-level data, sometimes referred to as synthetic controls, "in silico" data or virtual twins, will be out of scope for this chapter. Instead, we will assume external controls are drawn from a combination of RWD and clinical trials, where [1] refer to this collection of trial-external complementary data on control as "co-data." To mitigate selection bias, external control cohorts and patients should be identified through a systematic review that is planned, and ideally completed, before outcome data from the new clinical trial becomes available [2].

2.1 An Overview of Approaches for Leveraging External Control Data to Support Drug Development

When outcomes are available on our external controls, Figs. 1 and 4 show that use of this patient-level data to augment or replace the control arm of an RCT lies somewhere on a continuum. We can use the external controls to create a "hybrid" control arm, which is a mixture of internal controls drawn from an RCT and external controls. Or we can use them to create an "external control arm" (ECA), which is then compared indirectly with data from a single-arm trial (SAT). As we move from augmenting trial patients with external controls to replacing hypothetical internal controls with observed external controls, our reliance on the appropriateness of this

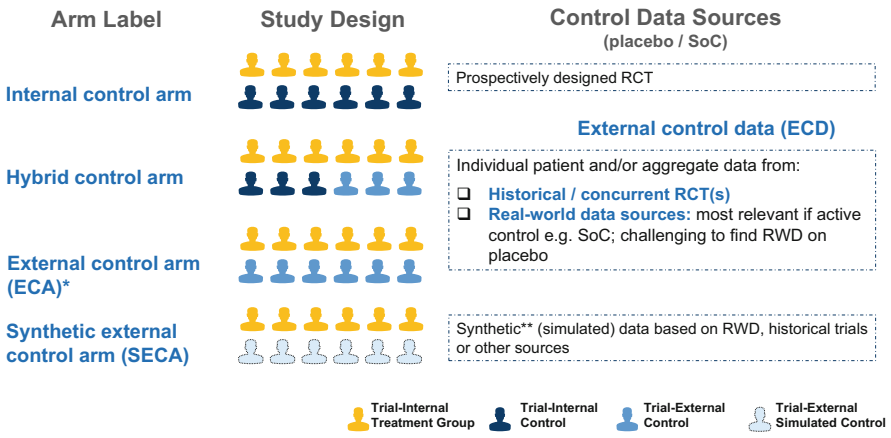

Fig. 4 From internal control arm to synthetic external control arm, different ways of leveraging trial external data

external data increases [3]. One advantage of choosing a hybrid control arm over an ECA is that there is still the opportunity to detect and react to a conflict between the internal and external controls.

The feasibility of leveraging external controls to create a hybrid or external control arm will depend on several scientific and strategic factors, with their quality, completeness, relevance to the research question, and "similarity" to the clinical trial data being crucial [4, 5]. From a strategic perspective, settings where the use of external controls in any guise is currently more acceptable include early phase clinical trials and scenarios where traditional stand-alone RCTs are less practical, ethical, or relevant [6]. Concern about the impact of biases driven by residual differences between the clinical trial data and external controls may be smaller when transformational treatment effects are anticipated. Six stringent criteria have been proposed by Pocock [7] for selecting external controls drawn from completed RCTs to augment a new RCT performed by the same investigators. Pocock's criteria stipulate that the external controls should be contemporaneous with the controls in the new trial and comparable in terms of baseline patient characteristics, while both groups should also be the same with respect to the treatment received and method of treatment evaluation. Similar principles also apply when selecting patients to create an ECA. While it is rare to strictly apply all six of Pocock's criteria in practice, they do shed light on scenarios when using external controls is most likely to be scientifically feasible. For example, the requirement for external controls to be drawn from a "recent" clinical trial or data source should be interpreted in the context of how rapidly standard of care has evolved prior to the new clinical trial: in a quickly advancing field, we may only be willing to consider strictly concurrent controls, whereas if standard of care has remained stable for several years, a wider timeframe may be regarded as sufficiently contemporaneous. More recently, updates to Pocock's criteria have been proposed which reflect advances in

statistical techniques which can control for some differences between external and internal controls and the greater variety of use cases for external controls [8].

As an aside, we note that not all uses of external controls require outcome data on these patients. One such application highlighted in Fig. 1 is the use of baseline covariate data on external controls to "generalize" the findings of an RCT [9, 10] to a target population. More specifically, this approach requires data on key baseline prognostic and/or predictive variables from external controls who are considered to be representative of a target population. For example, this target population might be real-world "treatment-eligible patients" who would meet the inclusion/exclusion criteria of the original RCT and will be prescribed the test therapy after regulatory approval and reimbursement coverage. In this case, baseline covariates could be distributed rather differently in the target population compared with the RCT population due to the selection bias introduced by the trial recruitment process. To address this discrepancy, we can use the covariate data on the external controls to "generalize" [11] the average causal treatment effect from the original RCT to the target population, so long as covariate distributions for the target and trial populations share a common support or at least an adequate overlap after trimming of the external controls if necessary. More discussion of the statistical methods for generalization will be provided later in Sect. 2.4.4.

Let us return now to focus on uses of external controls for whom outcome data are available. At the time of designing a new clinical trial with a hybrid or external control arm, we can first use the ICH E9(R1) estimand framework [12] to define the causal estimand of interest, and then use the target trial framework [13] to define the corresponding RCT we would have performed in an ideal world to estimate the estimand. We can then design our new clinical trial to try to emulate the target RCT. However, this process may reveal that the trial we *can* emulate differs from the trial we *want* to emulate due to potential limitations in the external controls. For example, it may be apparent early on that it will be infeasible to apply all of the inclusion/exclusion criteria defining the target population to select our external controls. Even if this process doesn't reveal any serious limitations in the external controls given what is known at the time of designing our new trial, uncertainties will likely remain about just how comparable our external controls will be with the future trial participants in terms of measured prognostic factors. This will translate into uncertainty about how we should design the clinical trial and the degree of reliance we should place on the external controls. In the next section, we explore adaptive hybrid designs which can accommodate prior uncertainty about their relevance.

2.2 *Adaptive Designs That Mitigate Uncertainty About the Relevance of External Controls*

Typically, if it is feasible to conduct an RCT then it will be highly preferable to incorporate some randomization into the design of a future trial, and the role of any

external controls will be to augment the trial control arm. For example, this will be the case if we are working in late phase clinical development and seeking to generate substantial evidence of efficacy, where obtaining an unbiased estimate of the causal effect of the test drug relative to control is of paramount importance. One strategy for planning RCTs with hybrid control arms is to design the trial sample size and randomization ratio so that the amount of statistical information for the control parameter contributed by the internal and external controls combined is equal to what would have been generated by a conventional RCT. However, if there is uncertainty about the relevance of the external controls and the amount of information they will contribute, this will translate into uncertainty about how many patients we should randomize to control in the new trial and thus the target sample size. Section 2.2.1 discusses adaptive designs as solutions to this challenge.

Alternatively, in phase II, there are certain therapeutic areas, such as oncology, where single-arm trials have historically been common practice, and there has been debate about the pros and cons of randomization [14]. In this context, where the evidence generated by the early phase trial is not intended to be used for confirmatory purposes, a different trade-off between costs, complexity, time, and the risks of potential biases may be tolerated by the sponsor. Consequently, there are scenarios, such as cancers which are uncommon but not rare, where a SAT could be countenanced as a design option even though an RCT, while potentially challenging operationally, is not infeasible. Of course, even in these cases, the acceptability of a SAT will depend on having access to a comparable group of external controls so that decisions of whether to invest in large-scale phase III trials can be based on estimates of more interpretable causal estimands rather than single-group parameters. A priori uncertainty about the relevance of the external controls may lead to uncertainty about whether to design the phase II trial as an RCT or SAT. In a recent paper, Götte and co-authors [15] propose an adaptive approach which can be used to mitigate this uncertainty in a quantitative and pre-planned way. We provide an overview of this adaptive design in Sect. 2.2.2.

2.2.1 Adaptive Approaches to Determining the Sample Size of an RCT with a Hybrid Control Arm

The effective sample size (ESS) of a Bayesian distribution for an unknown parameter quantifies the amount of statistical information it represents in terms of an equivalent number of observations. While it is relatively straightforward to calculate the ESS of a conjugate distribution for a single parameter, this is not the case for non-conjugate distributions. Indeed, several alternative information-based approaches to defining and evaluating the ESS of a non-conjugate distribution have been proposed, which can return markedly different results. For example, see the approach proposed by Morita and co-authors [16] in their seminal paper on calculating the ESS of a parametric prior distribution. More recently, the expected-local-information-ratio (ELIR) method has been proposed for calculating the ESS of a distribution of a single parameter [17]. The ELIR method, which is implemented

in the RBesT R package [18], has the advantage that it is predictively consistent in the sense that the ESS of a posterior distribution after collecting N observations is equal to the prior ESS + N. In what follows, we describe an adaptive strategy for determining the sample size of an RCT with a hybrid control arm, which uses the ESS to calculate how much information we have on a control-group parameter when leveraging external controls. The adaptive strategy is not prescriptive as to which method should be used to calculate the ESS, although we assume the ELIR method will often be preferred where possible.

When planning RCTs with hybrid control arms, we typically begin by calculating the sample size and decision rule needed for a conventional RCT to control the probabilities of erroneous decisions at specified levels. Denote the corresponding sample sizes on the test treatment and control by N_E and N_C, respectively: these can be thought of as targets for the ESS of the posterior distributions of key parameters (e.g., response probabilities on test and control) at the end of the hybrid trial. If our prior ESS for the parameter on control based on the external controls is $ESS_{C,0}$, a naive approach would be to plan the hybrid trial to randomize $(N_C - ESS_{C,0})$ internal controls. However, when using a Bayesian dynamic borrowing approach [19] to combine the internal and external controls, the external controls will be severely downweighted should we find outcomes are distributed very differently in these two groups, in which case, the ESS for the control parameter at the final analysis will miss our target N_C, inflating the risk of a false negative conclusion. In response to this concern, Schmidli et al. [20] propose the two-stage adaptive design given below. Note that while the authors' proposal assumes the external controls are incorporated via a robust Bayesian meta-analytic-predictive (MAP) prior, in principle the adaptive strategy could be applied using other Bayesian dynamic borrowing approaches (see Sect. 2.4 for further details). The adaptive design proceeds as follows:

Stage 1: Randomize $N_{E,1}$ patients to test treatment and $N_{C,1}$ to control.

Interim analysis (IA): Using the Stage 1 controls, update the robust MAP prior and calculate the ESS of the posterior, denoted by $ESS_{C,1}$. If the Stage 1 controls are consistent with the robust MAP prior, $ESS_{C,1} \approx ESS_{C,0} + N_{C,1}$ [20].

Stage 2: Randomize $N_{E,2} = N_E - N_{E,1}$ patients to test treatment and $N_{C,2} = \max(N_C - ESS_{C,1}, n_{C,\min})$ to control. Setting $n_{C,\min} > 0$ will facilitate blinding and identification of drifts in the control outcome distribution over the study.

When using this adaptive approach, the Stage 1 group sizes $N_{E,1}$ and $N_{C,1}$ need to be chosen carefully on a case-by-case basis. If $N_{C,1}$ is too small, it will be difficult to distinguish a prior data conflict from sampling variability. However, if we time the IA too late, the adaptation will have little impact, and we may find a smaller number of internal controls would have sufficed. Simulations can be used to identify values of $N_{E,1}$ and $N_{C,1}$ which adequately balance these risks.

2.2.2 Adaptive Clinical Trial Designs Mitigating the Risk of an External Control Arm

As mentioned earlier in the preamble to Sect. 2.2, phase II proof-of-concept (PoC) trials in oncology are often run as SATs [21], even when an RCT would be ethical and feasible. Götte et al. [15] propose an adaptive two-stage design for determining whether to conduct a PoC study as an RCT or as a SAT with an external control arm when there is uncertainty about the relevance of the N_C external controls. The proposed design is reasonably complex, requiring careful pre-specification of several design parameters which could be further tuned using simulation. Despite this, we find this systematic and pre-planned approach to evaluating the adequacy of a single-arm design to be interesting and offering potential advantages for improving the reliability of early phase decision-making. For ease of presentation, we do not provide details on all specifications for the adaptive design, and instead refer the interested reader to [15].

The adaptive phase II study proceeds in two stages:

Stage 1: Proceed as a SAT allocating $N_{E,1}$ patients to the test treatment.

Interim analysis: Quantify how comparable the $N_{E,1}$ patients from Stage 1 are with the N_C external controls with respect to measured baseline prognostic factors. Note that access to outcome data is not needed for this evaluation.

Stage 2: If the Stage 1 patients and external controls are deemed to be sufficiently similar according to a pre-specified decision rule, continue in Stage 2 as a SAT allocating patients only to the test treatment. Otherwise, randomize patients between treatment and control.

Final analysis: If Stage 2 proceeded using randomization, at the final analysis use data from Stages 1 and 2 on the test treatment and the Stage 2 internal controls to estimate the causal estimand. Note that as no outcome data were used to inform the interim adaptation, a standard analysis can be used with minimal impact on type I error rate control. If Stage 2 proceeded as a SAT, use all available data and a propensity score (PS)-based approach to emulate randomization and estimate the average effect of treatment on the treated.

One clarifying comment on the adaptive design is necessary. To evaluate the comparability of trial participants with the external controls at the IA, the design stipulates that a preference score (P_i) [22] be calculated for patient i, for each $i = 1, \cdots, N_{E,1} + N_C$, which is given by:

$$\log\left(\frac{P_i}{1-P_i}\right) = \log\left(\frac{e_i}{1-e_i}\right) - \log\left(\frac{\pi_1}{1-\pi_1}\right)$$

where $\pi_1 = N_{E,1}/(N_{E,1} + N_C)$ and e_i is the estimated propensity score (PS) for patient i, defined as the probability they are enrolled in the trial and exposed to the test drug given their baseline covariates. The definition of the preference score

implies that $P_i = 0.5$ when $e_i = \pi_1$. Therefore, a preference score of 0.5 indicates that given a patient's baseline covariates, they are no more or less likely to be in the clinical trial than they would have been had the external controls and Stage 1 patients been randomly allocated to groups in an R:1 ratio, with $R = N_{E,1}/N_C$, agnostic to their baseline covariates. If a patient's preference score deviates from 0.5, this indicates that they have a different propensity to be in the trial than simple randomization would suggest. If this is replicated across many patients, it suggests a lack of overlap between the distributions of measured covariates between groups. Therefore, one can define criteria for similarity between the Stage 1 trial patients and external controls in terms of the proportions of patients in each group who have preference scores in the neighborhood of 0.5.

2.3 Hybrid Adaptive Clinical Trials Using External Controls to Support Interim Decision-Making

So far, we have discussed designs for hybrid clinical trials which intend to leverage external controls in the final analysis. However, when planning a group sequential or adaptive trial, we may look to use external controls to support early stopping decisions or adaptations at an IA, and then use only trial-internal data at the final analysis. For example, external controls could be used to inform early stopping decisions for futility; a sample size reassessment; a population enrichment decision; or a dose-selection decision. Intuitively, by combining the trial-internal and trial-external controls, we should be able to increase the reliability of our interim decision-making (thus improving the trial operating characteristics) or alternatively, we can time the IA earlier and still make reliable decisions.

To give a flavor of these approaches, we share a recent proposal for a Bayesian scheme to leverage patient-level external controls to support a futility IA [23]. The aim of leveraging the external controls is to increase the probability of correctly stopping the trial early when the test treatment is inferior to control and to reduce the risk of erroneous stopping when it is superior. While frequentist approaches to combining the trial-internal and trial-external controls are possible, for this use-case Bayesian approaches may be preferred since they also facilitate the calculation of Bayesian metrics such as the posterior predictive probability of trial success and the posterior probability of a clinically relevant advantage of the test treatment versus control, which can be useful for supporting quantitative decision-making. By combining resampling and Bayesian multiple-imputation techniques, the authors in [23] sample from the predictive distribution of Z_2, the usual standardized test statistic at the final analysis, given the external controls and interim data. Then, the trial is stopped early for futility if the Bayesian predictive probability of statistical significance at the final analysis is sufficiently low [24]. The stopping threshold itself must be tailored to each trial and reflects how much power we are willing to sacrifice in order to be able to stop for futility with a high probability under the

null hypothesis. To compute Bayesian predictive power, several multiple imputation approaches are proposed which differ by how they make use of information on patient baseline covariates. For example, assuming a binary endpoint, the trial data and external controls can be used to fit a Bayesian logistic outcome model adjusting for treatment and baseline covariates. Sampling from the joint posterior distribution of the model parameters, and resampling with replacement baseline covariate vectors from trial-internal patients, we can impute the "missing" baseline profiles and outcomes of patients yet to be recruited and/or followed-up at the time of the IA and thus calculate Z_2. Repeating this process N times and counting the proportion of times Z_2 exceeds the final success threshold reveals the predictive probability of statistical significance given the data available at the IA. It is possible to extend this procedure by fitting different outcome models, for example, replacing the baseline covariates in the logistic regression by the logit of the PS or adjusting for both the covariates and the logit of the PS.

Returning now to hybrid clinical trials which intend to leverage our external controls in the final analysis, the authors in [6] highlight two popular schools of analytical techniques which can combine various sources of control data while accounting for between-source heterogeneity: PS approaches and Bayesian meta-analytic approaches. In the following section, we introduce the Bayesian meta-analytic approach to borrowing, compare and contrast this with PS-based methods, and discuss recent advances which have fused the two approaches together to leverage patient-level external controls.

2.4 Analytical Approaches for Combining External Controls and Clinical Trial Data

2.4.1 Comparing Bayesian Dynamic Borrowing and Propensity Score Analytic Approaches

PS methods are widely applied in the epidemiological literature for the analysis of patient-level data from observational studies. In the context of clinical trials leveraging external controls, a patient's propensity score is interpreted as the conditional probability they are in the clinical trial given their vector of baseline covariates. The goal of PS methods is to draw inferences about causal treatment effects by emulating randomization. This can be done by using the PS to match, stratify or weight patients, or adjust for the PS as a covariate in an outcome regression, to ensure groups are balanced with respect to measured baseline confounders. We refer the interested reader to chapter "Clinical Studies Leveraging RWD using Propensity Score-Based Methods" for more details.

Meanwhile, in recent years, a rich literature has emerged in the clinical trials community exploring how Bayesian methods can be used to augment an RCT with external controls. While in principle these approaches can be applied with external controls drawn either from clinical trials or RWD, there are few published examples

of their use with RWD [6]. Of particular interest are Bayesian approaches which facilitate dynamic borrowing, that is, inconsistencies between observed outcomes among the external and internal controls are taken to imply the external and internal controls differ with respect to key parameters of their outcome distributions, which prompts the external controls to be discounted. Note that priors for Bayesian model parameters and other parameters influencing the borrowing behavior are specified ahead of time. Therefore, while the weight attributed to the external controls in the posterior is outcome adaptive, the method determining this is pre-specified. We see that Bayesian dynamic borrowing characterizes between-source heterogeneity in terms of differences between parameters of the outcome distribution (such as the log-odds of response), whereas PS methods focus on differences between the distributions of measured baseline confounders.

Bayesian dynamic borrowing approaches include modified power priors (which raise the likelihood of the historical data to an unknown power, regarded as a random variable); commensurate priors; and the (robust) MAP prior [25]. When applied to external controls, these three Bayesian dynamic borrowing approaches differ in terms of how they relate parameters in the different sources of controls. However, there are equivalencies between the methods for the case of a single source of external controls, with each method assuming that source-specific parameters in the external controls and new clinical trial are similar but not identical [1]. Note that the power prior which raises the likelihood of the historical data to a fixed power does not facilitate dynamic borrowing, only static borrowing, since the weight attributed to the external controls is pre-specified and therefore cannot react to an observed prior-data conflict. Several simulation studies have been reported comparing different dynamic and static approaches to leveraging external controls [19, 25].

In the authors' experience, Bayesian dynamic borrowing approaches are popular in practical applications. To be concrete, we describe the MAP prior for the case that we have access to K sources of external control data which we want to leverage in the analysis of a new study *. The approach can accommodate $K \geq 1$, and thus can be applied even if there is only one source of external controls. While the MAP prior was first proposed with aggregate data, for several common types of outcome data it is straightforward to accommodate patient-level data or a mixture of patient-level and aggregate data [26]. For the purposes of illustrating the MAP approach, suppose each patient provides a binary response so that the ith patient on control in source k is distributed as $Y_{ik} \mid p_k \sim \text{Bern}(p_k)$, for $i = 1, \cdots, n_k$. Furthermore, define $\theta_k = \log\{p_k/(1-p_k)\}$, for each $k = 1, \cdots, K, *$, and the source-specific log-odds on control are connected via a random-effects distribution, that is,

$$\theta_1, \cdots, \theta_K, \theta_* \mid \mu, \tau \sim N\left(\mu, \tau^2\right), \tag{1}$$

where τ captures the heterogeneity in the source-specific log-odds parameters and μ is interpreted as the global average log-odds of response on control. Typically, the parameter which is of primary interest in this model is θ_*, the log-odds of

response on control in the new trial, rather than μ. The Bayesian hierarchical model is completed by placing priors on μ and τ. Random-effects model (1) can be extended to accommodate several strata of external controls such as controls from RCTs, prospectively generated RWD, secondary-use RWD, each of which are characterized by a different heterogeneity parameter [1]. One can use the control data from sources 1, ..., K, to fit the Bayesian hierarchical model and derive a MAP prior for θ_*.

In practice, a robust version of the MAP prior for θ_* is usually taken forward for the analysis of study *, which is a mixture of a weakly informative prior and the MAP prior. It is referred to as "robust" because its heavier tails means it discounts the external control information more quickly in the event of a prior-data conflict. If trial * is an RCT, the robust MAP prior will be updated with the internal controls once they become available. If, instead, trial * is designed as a SAT, the robust MAP prior will not be updated. We see that in contrast to PS approaches, where the contribution of external controls is determined at the patient-level given their baseline covariates, Bayesian approaches discount at the level of the data source.

The Bayesian hierarchical model defined above does not incorporate information on baseline covariates which will often be available with patient-level external controls. However, this covariate data may be useful for explaining between-source heterogeneity. With this in mind, several approaches combining PS and Bayesian dynamic borrowing approaches are available and are reviewed below.

2.4.2 Combining PS Matching and Bayesian Dynamic Borrowing

A case-study in oncology submitted as part of the FDA's Complex Innovative Design (CID) pilot program proposed using a two-stage approach to leverage patient-level controls from a partially concurrent RCT to augment the analysis of a key secondary endpoint (Overall Survival) in a planned RCT [27]. First, in Step 1, PS matching is used to identify similar external controls. Then in Step 2, a Bayesian commensurate prior is used to dynamically borrow information from the trial-external controls [28]. A recent simulation study has evaluated the performance of several Bayesian approaches to leveraging external controls when preceded by a matching procedure to identify relevant external controls [29].

2.4.3 Combining PS Stratification and Bayesian Dynamic Borrowing

The authors in [30] propose an alternative approach to combining PS with the Bayesian MAP approach to leverage external controls from K sources to augment a new RCT or SAT. The approach begins by using the trial participants and external controls to fit a PS model, and then trims the external controls to discard those with estimated PS outside the range seen in the trial. S strata are then defined for the PS, which are chosen to contain approximately equal numbers of trial patients. Within each stratum, the data are assumed to follow a Bayesian hierarchical model

which stipulates that parameters, such as the log-odds of response, in the trial controls and each source of external controls are samples from a normal random-effects distribution with a stratum-specific mean and standard deviation. The target parameter is defined as the weighted average of the stratum-specific parameters for the trial controls, which we denote by θ*. Hyperparameters of the prior distributions for the stratum-specific standard deviations are calibrated to ensure the ESS of the MAP prior for θ* is equal to a pre-specified target.

2.4.4 Combining PS Weighting or Parametric g-Estimation with the Bayesian Meta-analytic Approach

To account for differences in the distribution of baseline covariates (referred to as "case mix") between studies, the authors in [31] propose a novel approach to the meta-analysis of individual patient data from comparative RCTs which proceeds in two stages. First, the analyst generalizes the treatment effect estimate for each RCT to the covariate distribution of the target trial of interest. Second, these generalized estimates are combined via a (frequentist) random-effects meta-analysis. Through this process, we can disentangle the heterogeneity due to differences in the distribution of baseline covariates and heterogeneity due to other (perhaps unmeasured) factors such as differences in trial protocol, definition of treatments, etc.

While [31] assumes each study is an RCT providing a comparative treatment effect estimate, it is possible to extend their ideas to the scenario where we want to leverage K sources of external controls in the analysis of a new hybrid clinical trial. In this case, first we generalize estimates of single-group parameters (e.g., response probability) from each control data source to the case-mix in the new trial, which is regarded as the target population. Generalization can either be achieved through outcome regression and standardization or through PS weighting. Second, we combine these estimates via a Bayesian random-effects meta-analysis to create a robust MAP prior for the control parameter of interest in the new study (i.e., θ_*).

The outcome regression and standardization approach for generalizing results from one population to another has advantages and disadvantages compared with the weighting by odds technique. One advantage is that if external control dataset k is large, it might still be possible to obtain reliable estimates for the prognostic effects of covariate levels which occur with a low prevalence in the external controls but are common in the new trial. However, one disadvantage is that without careful consideration, it is easy to extrapolate beyond the range of observation [31]. Therefore, the choice of approach and their relative merits need to be considered on a case-by-case basis.

Clearly this two-stage approach has advantages over a one-step procedure using only PS weighting, only outcome modelling and standardization, or only meta-analysis to combine marginal parameter estimates. By first generalizing estimates to a common target population, we ameliorate between-source heterogeneity, which should facilitate borrowing from the external controls. However, the method is flexible enough to accommodate between-source differences which go beyond

differences in case mix. As an aside, techniques such as outcome regression and standardization or weighting can also be used to generalize study results from an RCT to a more representative real-world population.

3 Randomized Controlled Studies Incorporating Real-World Data

3.1 Pragmatic Randomized Designs and Decentralized Randomized Designs

Multiple novel hybrid designs aim to maintain the high level of evidentiary standards of the RCT, for example by incorporating randomization and pre-specifying hypotheses of interest, while trying to add flexibility in the design to achieve other aims. The evidentiary standards are maintained high to inform effectiveness or safety of medical products to multiple stakeholders including patients, clinicians, healthcare administrators, and policy-makers. Those designs are either named by the type of flexibility they use or the additional aims they are trying to achieve.

More specifically, *pragmatic randomized studies* or *point-of-care studies (PCTs)* maintain randomization while incorporating more clinical-practice-like design strategies [32, 33]. Similarly, *virtual clinical trials*, *direct-to-patient* or *decentralized clinical trials (DCTs)* maintain randomization while incorporating more patient-centric design strategies [34, 35]. Relative to traditional RCTs, PCTs and DCTs share similar goals of decreasing burden on patient or investigator participation in the trial, increasing diversity of the cohort participating in the trial, and accelerating evidence generation in research.

Both designs are considered hybrid because they can leverage existing RWD, for example by linkage to electronic healthcare systems collected as part of a patient's standard of care. In addition, DCTs may incorporate home visits, or other sources of RWD such as patient's self-report, or use of digital technology, at discrete times or continuously over the trial's follow-up period. Thus, in the hybrid design spectrum, DCTs and PCTs augment the data as shown in Fig. 4 by allowing RWD to add more variables on the same patients recruited into the randomized clinical trials.

While PCTs and DCTs may share similar goals and methods of recruitment and consent of patients, they are distinct in their philosophies about outcome data collection. On one hand, fully pragmatic studies take a minimalist approach to data collection as they aim to seamlessly integrate in a clinician's workflow and their interactions with their patients (e.g., match frequency and timing of visits). On the other hand, fully decentralized studies are maximalist as they try to have a more comprehensive view of the patient health journey from multiple patient-centric sources, including but not limited to their interactions with their clinicians.

In therapeutic development, PCT designs are more common in the post-market setting to evaluate effectiveness or safety of approved therapies. For example, the

PCT study ACHIEVE Control investigated the safety and effectiveness of insulin glargine 300 U/ml, after it was approved, relative to first-generation basal insulin analogues in patients with uncontrolled type 2 diabetes mellitus [36]. Similarly, the large DAPA-MI trial is embedded in routine care and registries in Sweden and the United Kingdom, to support a label expansion of dapagliflozin [37]. The large Salford Lung Study was the first PCT of its kind, in collaboration with the UK healthcare system, and was successful at investigating the effectiveness of a new inhaler combination against standard of care [38].

When DCTs incorporate digital technology elements explored as biomarkers, they are used in the proof of concept setting to validate new endpoints. For example, a new phone application assessing visual activity at home instead of the clinic could be used in clinical trials [39, 40]. Also, the use of the actigraphy tracking device to measure moderate to vigorous physical activity is gaining regulatory acceptance as a primary endpoint in clinical trials [41, 42]. With the COVID-19 pandemic disruption of on-site clinical care, many studies across therapeutic areas incorporated decentralized elements and thereby increased the interest in using DCT designs [43–46] beyond early and late development.

Pragmatic study designs fall in a spectrum from more controlled to more pragmatic and closer to clinical practice. The PRECIS-2 tool [47] can help guide discussions among the research team or stakeholders to balance the sometimes conflicting goals of optimizing flexibility, feasibility, and fitness-for-purpose. The PRECIS-2 tool has nine domains (eligibility criteria, recruitment, setting, organization, flexibility of therapy delivery, flexibility in therapy adherence, follow-up, primary outcome, and primary analysis) and scores the acceptable level of flexibility of each domain from 1 (very explanatory) to 5 (very pragmatic).

Decentralized studies also fall in a spectrum from fully on-site procedures for all patients to fully off-site or patient-centric for all procedures and for all patients. The procedures can fall anywhere in the patient journey including outreach, determining eligibility, accessing therapy, or collecting clinical outcomes over time until the end of the study. The scope of decentralization is also multi-factorial including having decentralized outcomes for only a subset of patients, or for a subset of visits, or for a subset of outcomes, or any combination of subsetting in the above.

3.2 Scientific Considerations with PCT and DCT Hybrid Designs

3.2.1 Real-World Considerations, the Scientific Question/Estimand, and the Study Hypotheses

Novel designs like PCT or DCT can increase the breadth of questions that can be answered with a randomized study. Thus, one strategic consideration of interest is whether these novel designs are answering a scientific question that could not be answered otherwise or whether they are answering these questions more efficiently.

The timing of when to answer specific scientific questions in the development program is also important because while generalizability of findings from a trial to the overall indicated population in real-world setting is of interest for all trials, demonstrating a good benefit-risk profile for a new compound takes priority. Also, less frequent on-site monitoring may not protect patient safety for a new molecular entity but may be acceptable for a molecular entity with a better known benefit-risk safety profile. Also, because effect sizes in efficacy are typically higher than for effectiveness, one would typically demonstrate the former before the latter. Similarly, one would demonstrate safety in the short term before investigating safety in the long term.

To tease out the importance of real-word considerations on the scientific question of interest, it is helpful to use the estimand framework outlined in the ICH-E9 addendum and also in chapter "Estimand in Real-World Evidence Study: From Frameworks to Application" of this book. This framework helps spell out the scientific questions in detail regarding all five attributes (population, the treatment, variable, intercurrent event, and summary measure) [12]. For example, are novel designs more promising at targeting the indicated population? Will they enable better mimicking of the indicated decision to initiate or use the treatment? Will the designs enable one to target novel endpoints or endpoints that matter to different decision-makers? Will the designs minimize the concern for some intercurrent events?

Real-world considerations may impact the hypotheses of interest and whether these will aim for superiority, non-inferiority, or equivalence. Pragmatic safety studies may also aim for a hypothesis to rule out an excess risk. Planning for a superiority hypothesis to standard of care in real-world settings may be different than planning for superiority against placebo in a more controlled RCT. More specifically, the input and considerations in a sample size calculation to power the study to detect a change will need to be tailored to the expected recruitment rate, variability, effect size against an active control, and the handling of intercurrent events to reach the estimand of interest. Also, justifying a non-inferiority margin, a rule-out margin, or an equivalence margin based on past performance of the comparator drug in traditional RCTs when one plans for a novel design in real-world conditions may prove difficult.

3.2.2 Real-World Endpoints and Statistical Methods Used to Support Validity and Fitness-for-Purpose

When the PCTs or DCTs are using novel endpoints, one may have to demonstrate fitness-for-purpose of these endpoints in label-enabling studies [48–50]. For a clinical outcome assessment, this process starts with a justification for the target construct that the endpoint is purported to measure and a purpose for using this endpoint in the study or the clinical development (e.g., for a labeling claim or a marketing claim). The construct justification includes information on why the endpoint would be meaningful to the indicated patients and how they feel or live their life, for example through survey of patients, their families, or their healthcare

support network. Then, showing fitness-for-purpose requires evidence of the validity of the clinical outcome assessment and that the strength of the evidence matches the stated purpose. Validating an assessment's use in a clinical study typically requires showing that the use is accurate at capturing the target construct with reproducible results and small measurement error.

Being clear about the target construct and the regulatory purpose is therefore very important in a fitness-for-purpose assessment of the endpoint. For example, codes for myocardial infarction in insurance claims databases accurately reflect the physician's diagnosis of this event as reported in the patient's medical records with a positive predictive value above 90% [51]. While this accuracy may be sufficient for the purpose of post-market safety assessment, it may not be so for marketing approval where the target construct goes beyond real-world assessment of a cardiovascular event to an adjudication of each event by the same independent committee of cardiologists.

Although the setting of data capture of all study outcomes in PCTs and DCTs may be different than traditional RCTs, this does not necessarily imply a loss of accuracy or an increase in variability relative to the target construct. For example, real-world data capture in real time of patient-reported outcomes in a DCT may be less prone to recall bias than a capture at set visits in an RCT [52]. Similarly, electronic healthcare records or insurance claims records related to a particular hospitalization may have more comprehensive information on procedures preceding or following hospitalization to inform assessment of causality to therapy than the protocol pre-specified variables reported in a case report form in an RCT.

Another element of accuracy is timeliness of the data sources for a given target construct. For example, death or cause of death may be missing or inaccurately captured in electronic healthcare records used in PCTs or DCTs and linkage to death records that are generally more accurate may have a time gap of a few years. Thus, complementing these sources with additional data capture from patients or other data sources may be necessary to increase accuracy and timeliness.

When there are multiple sources for the same target construct, for example if the same measure is collected on-site and off-site between or across subjects, then one has to establish equivalence and exchangeability or prioritization in case of conflict. This evaluation can be within the same study or in a separate study where a subset of patients received both methods of measurements. Several psychometric analyses and measures developed to evaluate inter-rater agreement, such as Cronbach's alpha and Kappa statistic for dichotomous measures and intra-class correlation for continuous measures, can help establish equivalence or evaluate differences. In addition, interoperability of data sources that may be using different ontologies of data collection and storage may be necessary to integrate information from these multiple sources.

Data mining and machine-learning methods are also important in handling the large volume of wearable device data in clinical trials. For example, artificial intelligence methods have been used to establish authenticity of this data in risk-based monitoring and verify that the data comes from the patient and not another source. In addition, dimensionality reduction methods can help compress the

complex time series to the relevant features (e.g., mean, peak, or area under the curve, over a time window).

Lastly, in the design of a DCT or PCT, one has to balance the convenience of recruitment and participation with a potential increase in variability in the endpoint. On the one hand, relative to a typical RCT, PCTs and DCTs are expected to recruit more patients and/or shorten the study duration, by design, because of greater convenience to patients. On the other hand, PCTs and DCTs could suffer from a larger measurement error, a larger between source variability or a larger between subject variability in the endpoint of interest that will impact study power to detect a change.

4 Discussion

This chapter reviewed innovative hybrid designs and analytical strategies that integrate RWD with clinical trials. We have particularly focused on designs combining the advantage of planning and randomization in RCTs with leveraging existing or conveniently collected RWD.

Based on the examples discussed in this chapter, we believe that the use of external controls to create hybrid control arms in RCTs should be routinely considered for early phase clinical trials, as well as in pivotal settings where conventional stand-alone RCTs are less practical or relevant, such as in the treatment of rare disorders, in pediatric studies, or in epidemics [10]. More broadly, external controls can also play a useful role to support interim decision-making in an adaptive early or late phase clinical trial, supporting futility stopping decisions or mid-study adaptations. Conversely, PCTs are more promising as label expansion or post-market safety studies, when prescribing the product is easier to implement at point-of-care and less frequent monitoring of patients to increase convenience does not put these patients at risk. Fully virtual DCTs have been used in early development to validate new digital endpoints or new methods of recruitment of patients but have not been fully tested in clinical development. There are however some indications that tomorrow's patients will expect flexibility in their interactions with the healthcare system and a mix of in-person, remote interaction, and data sharing through wearables may become the norm [53].

We have seen in this chapter that adaptive clinical trial designs can play an important role in mitigating some of the risks associated with leveraging external controls where there is a prior uncertainty about comparability between the trial and real-world patients (which we speculate will typically be the case in practice). However, even with this additional flexibility, careful evaluation is still needed at the design stage of the strategic and scientific feasibility of proposed design options in light of the scientific question the study is intended to address and the level of evidence required [54].

Lastly, real-life follow-up is prone to intercurrent events that may be hard to plan for or minimize and that will impact the main analyses, and the interpretation

of a treatment policy treatment effect estimate of a PCT or DCT relative to more controlled clinical trials. Similarly, compared to a typical RCT, missed visits, loss of follow-up, or departure from randomized treatment may be more common in an external control data source or in study arms of a PCT. Wearable devices used in DCTs offer the opportunity of more frequent longitudinal assessments on the same patient but also have novel sources of measurement error such as inaccuracy or missingness due to loss of connectivity on the device [55]. All of the designs we have discussed in this chapter put a large emphasis on time 0 and comparability at start of follow-up for leveraging external control RWD or randomization in PCTs and DCTs. However, once the hurdle of comparability at time 0 is overcome, we need to increase our understanding of the impact of different RWD-specific intercurrent events on the question of interest and the main analyses.

For any hybrid design to be successful, planning, recruitment, and data capture in clinical trials need to be nimble, easily integrating evidence from different sources. Similarly, electronic healthcare systems need to more easily be used and accessed for answering research questions. Stronger collaborations between healthcare systems, regulators, and industry, such as was seen in the development of treatments for COVID-19, can facilitate these designs in answering questions that go beyond a particular development program to be in the realm of public health.

References

1. Neuenschwander, B., S. Roychoudhury, and H. Schmidli, *On the use of co-data in clinical trials.* Statistics in Biopharmaceutical Research, 2016. **8**(3): p. 345-354.
2. Higgins, J.P. and J. Thomas. *Cochrane Handbook for Systematic Reviews of Interventions*. 2022 [cited 2022 September]; Available from: www.training.cochrane.org/handbook.
3. Concato, J. and J. Corrigan-Curay, *Real-World Evidence — Where Are We Now?* New England Journal of Medicine, 2022. **386**(18): p. 1680-1682.
4. Izem, R., et al., *Real-World Data as External Controls: Practical Experience from Notable Marketing Applications of New Therapies.* Therapeutic Innovation & Regulatory Science, 2022: p. 1-13.
5. Levenson, M., et al., *Statistical Consideration for Fit-for-Use Real-World Data to Support Regulatory Decision Making in Drug Development.* Statistics in Biopharmaceutical Research, 2022: p. 1-8.
6. Schmidli, H., et al., *Beyond randomized clinical trials: Use of external controls.* Clinical Pharmacology & Therapeutics, 2020. **107**(4): p. 806-816.
7. Pocock, S.J., *The combination of randomized and historical controls in clinical trials.* Journal of chronic diseases, 1976. **29**(3): p. 175-188.
8. Hatswell, A., et al., *Summarising salient information on historical controls: A structured assessment of validity and comparability across studies.* Clin Trials, 2020. **17**(6): p. 607-616.
9. Dahabreh, I.J., et al., *Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals.* Biometrics, 2019. **75**(2): p. 685-694.
10. Cole, S.R. and E.A. Stuart, *Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial.* American journal of epidemiology, 2010. **172**(1): p. 107-115.
11. Colnet, B., et al., *Causal inference methods for combining randomized trials and observational studies: a review.* arXiv preprint arXiv:2011.08047, 2020.

12. The International Council of Harmonisation. *ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials*. 2020; Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf.
13. Hernán, M.A. and J.M. Robins, *Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available.* Am J Epidemiol, 2016. **183**(8): p. 758-64.
14. Grayling, M.J., et al., *A Review of Perspectives on the Use of Randomization in Phase II Oncology Trials.* JNCI: Journal of the National Cancer Institute, 2019. **111**(12): p. 1255-1262.
15. Götte, H., et al., *An adaptive design for early clinical development including interim decision for single-arm trial with external controls or randomized trial.* Pharmaceutical Statistics, 2022. **21**(3): p. 625-640.
16. Morita, S., P.F. Thall, and P. Müller, *Determining the effective sample size of a parametric prior.* Biometrics, 2008. **64**(2): p. 595-602.
17. Neuenschwander, B., et al., *Predictively consistent prior effective sample sizes.* Biometrics, 2020. **76**(2): p. 578-587.
18. Weber, S., et al., *Applying Meta-Analytic-Predictive Priors with the R Bayesian Evidence Synthesis Tools.* Journal of Statistical Software, 2021. **100**(19): p. 1 - 32.
19. Viele, K., et al., *Use of historical control data for assessing treatment effects in clinical trials.* Pharmaceutical Statistics, 2014. **13**(1): p. 41-54.
20. Schmidli, H., et al., *Robust meta-analytic-predictive priors in clinical trials with historical control information.* Biometrics, 2014. **70**(4): p. 1023-1032.
21. Langrand-Escure, J., et al., *Quality of reporting in oncology phase II trials: A 5-year assessment through systematic review.* PLoS One, 2017. **12**(12): p. e0185536.
22. Walker, A., et al., *A tool for assessing the feasibility of comparative effectiveness research.* Comparative Effectiveness Research, 2013. **3**: p. 11-20.
23. Ventz, S., et al., *The use of external control data for predictions and futility interim analyses in clinical trials.* Neuro-oncology, 2022. **24**(2): p. 247-256.
24. Jennison, C. and B.W. Turnbull, *Group sequential methods with applications to clinical trials*. 1999: CRC Press.
25. van Rosmalen, J., et al., *Including historical data in the analysis of clinical trials: Is it worth the effort?* Statistical methods in medical research, 2018. **27**(10): p. 3167-3182.
26. Gsteiger, S., et al., *Using historical control information for the design and analysis of clinical trials with overdispersed count data.* Statistics in Medicine, 2013. **32**(21): p. 3609-3622.
27. The US Food and Drug Administration. *Complex and Innovative Design Case Study: External control in Diffuse B-Cell Lymphoma*. 2022 [cited 2022 October]; Available from: https://www.fda.gov/media/155405/download.
28. Lewis, C.J., et al., *Borrowing from historical control data in cancer drug development: a cautionary tale and practical guidelines.* Statistics in biopharmaceutical research, 2019. **11**(1): p. 67-78.
29. Shan, M., et al., *A Simulation-Based Evaluation of Statistical Methods for Hybrid Real-World Control Arms in Clinical Trials.* Statistics in Biosciences, 2022: p. 1-26.
30. Liu, M., et al., *Propensity-score-based meta-analytic predictive prior for incorporating real-world and historical data.* Statistics in medicine, 2021. **40**(22): p. 4794-4808.
31. Vo, T.T., et al., *A novel approach for identifying and addressing case-mix heterogeneity in individual participant data meta-analysis.* Research synthesis methods, 2019. **10**(4): p. 582-596.
32. Schwartz, D. and J. Lellouch, *Explanatory and pragmatic attitudes in therapeutical trials.* Journal of clinical epidemiology, 2009. **62**(5): p. 499-505.
33. Lentz, T.A., et al., *Designing, Conducting, Monitoring, and Analyzing Data from Pragmatic Randomized Clinical Trials: Proceedings from a Multi-stakeholder Think Tank Meeting.* Therapeutic Innovation & Regulatory Science, 2020. **54**(6): p. 1477-1488.
34. National Academies of Sciences Engineering Medicine and Health, et al., *The National Academies Collection: Reports funded by National Institutes of Health*, in *Virtual Clinical*

Trials: Challenges and Opportunities: Proceedings of a Workshop, C. Shore, E. Khandekar, and J. Alper, Editors. 2019, National Academies Press (US) Copyright 2019 by the National Academy of Sciences. All rights reserved.: Washington (DC).
35. Khozin, S. and A. Coravos, *Decentralized trials in the age of real-world evidence and inclusivity in clinical investigations.* Clin Pharmacol Ther, 2019. **106**(1): p. 25-27.
36. Anderson, J., et al., *Target attainment in insulin-naive patients at high risk for hypoglycemia: Results from ACHIEVE Control.* J Diabetes Complications, 2021. **35**(4): p. 107831.
37. AstraZeneca. *Farxiga granted Fast Track Designation in the US for heart failure following acute myocrdial infarction leveraging an innovative registry-based trial design*. 2020 [cited 2022 October]; Available from: https://www.astrazeneca.com/media-centre/press-releases/2020/farxiga-granted-fast-track-designation-in-the-us-for-heart-failure-following-acute-myocardial-infarction-leveraging-an-innovative-registry-based-trial-design.html#!
38. The National Institute for Health and Care Research. *Case study: Delivering real world research- The Salford Lung Study*. 2019 [cited 2022 October]; Available from: https://www.nihr.ac.uk/documents/case-study-delivering-real-world-research-the-salford-lung-study/11555.
39. Kaiser, P.K., et al., *Feasibility of a novel remote daily monitoring system for age-related macular degeneration using mobile handheld devices: results of a pilot study.* Retina, 2013. **33**(9): p. 1863-1870.
40. Novartis. *Novartis launches FocalView app, providing opportunity for patients to participate in ophthalmology clinical trials from home*. 2018 [cited 2022 October]; Available from: https://www.novartis.com/news/media-releases/novartis-launches-focalview-app-providing-opportunity-patients-participate-ophthalmology-clinical-trials-from-home.
41. Bellerophon Pulse Technologies. *A Study to Assess Pulsed Inhaled Nitric Oxide in Subjects with Pulmonary Fibrosis at Risk for Pulmonary Hypertension*. 2017 [cited 2022 October]; Available from: https://clinicaltrials.gov/ct2/show/NCT03267108.
42. Bellerophon Therapeutics. *Bellerophon Announces FDA Acceptance of Change to Ongoing Phase 3 REBUILD Study of INOpulse for Treatment of Fibrotic Interstitial Lung Disease*. 2022 [cited 2022 October]; Available from: https://investors.bellerophon.com/news-releases/news-release-details/bellerophon-announces-fda-acceptance-change-ongoing-phase-3.
43. Normand, S.-L.T., *The RECOVERY Platform.* New England Journal of Medicine, 2020. **384**(8): p. 757-758.
44. Loucks, T.L., et al., *Clinical research during the COVID-19 pandemic: The role of virtual visits and digital approaches.* Journal of Clinical and Translational Science, 2021. **5**(1): p. e102.
45. Skipper, C.P., et al., *Hydroxychloroquine in Nonhospitalized Adults With Early COVID-19: A Randomized Trial.* Ann Intern Med, 2020. **173**(8): p. 623-631.
46. Alemayehu, D., et al., *Perspectives on Virtual (Remote) Clinical Trials as the "New Normal" to Accelerate Drug Development.* Clin Pharmacol Ther, 2022. **111**(2): p. 373-381.
47. Loudon, K., et al., *The PRECIS-2 tool: designing trials that are fit for purpose.* BMJ : British Medical Journal, 2015. **350**: p. h2147.
48. The US Food and Drug Administration. *Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims, Guidance for Industry*. 2009 [cited 2022 October]; Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-reported-outcome-measures-use-medical-product-development-support-labeling-claims.
49. The US Food and Drug Administration. *Biomarker qualification evidentiary framework guidance for industry*. 2018 [cited 2022 October]; Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/biomarker-qualification-evidentiary-framework.
50. The US Food and Drug Administration. *Qualification Process for Drug Development Tools, Guidance for Industry and FDA Staff*. 2020 [cited 2022 October]; Available from: https://www.fda.gov/media/133511/download.
51. Cutrona, S.L., et al., *Validation of acute myocardial infarction in the Food and Drug Administration's Mini-Sentinel program.* Pharmacoepidemiol Drug Saf, 2013. **22**(1): p. 40-54.

52. Hassan, E., *Recall bias can be a threat to retrospective and prospective research designs.* The Internet Journal of Epidemiology, 2006. **3**(2): p. 339-412.
53. Perry, B., et al., *Patient preferences for using mobile technologies in clinical trials.* Contemporary Clinical Trials Communications, 2019. **15**: p. 100399.
54. Ghadessi, M., et al., *A roadmap to using historical controls in clinical trials – by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG).* Orphanet Journal of Rare Diseases, 2020. **15**(1): p. 69.
55. The US Food and Drug Administration. *Digital Health Technologies for Remote Data Acquisition in Clinical Investigations*. 2022 [cited 2022 October]; Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/digital-health-technologies-remote-data-acquisition-clinical-investigations.

Statistical Challenges for Causal Inference Using Time-to-Event Real-World Data

Jixian Wang, Hongtao Zhang, and Ram Tiwari

1 Introduction

RWD have been increasingly used in drug development, particularly in combination with clinical trial data. As treatments are not randomized in real-world settings, a major challenge is how to adjust for population difference for subjects receiving different treatments to eliminate or reduce confounding biases; hence, causal inference is a key component in using RWD. Several approaches applicable to RWD have been nicely summarized in Levenson et al. [22] and Ho et al. [15], as well as in chapters "Causal Inference with Targeted Learning for Producing and Evaluating Real-World Evidence", "Clinical Studies Leveraging Real-World Data Using Propensity Score-based Methods", and "Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods". Hernan and Robins [14] gave more details, in particular, time-to-event (TTE)-related methods in Chapter 17 in [14]. Our focus here is on some specific challenges in using RWD for TTE outcomes and to offer some practical solutions.

RWD can be used for multiple purposes. The following list is by no means exhaustive:

1. Indirect comparisons in which the RWD form the external control for a single-arm trial
2. Augmenting a small internal control arm in an RCT

J. Wang (✉)
GBDS, Bristol Myers Squibb, Boudry, Switzerland
e-mail: jixian.wang@bms.com

H. Zhang
Merck & Co., Inc., North Wales, PA, USA

R. Tiwari
GBDS, Bristol Myers Squibb, Lawrenceville, NJ, USA

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_13

3. Estimating treatment effects based on RWD, e.g., for designing a trial or as a reference
4. Generalizing evidence, e.g., from an RCT to a population specified by RWD

In fact, the key step in these tasks is adjusting for the population difference between RWD and the trial or between subgroups receiving different treatments in comparison in RWD. Many general approaches such as propensity score (PS) matching and weighting [32, 33], covariate balancing [9, 53], and direct adjustment using models [34], also known as g-formula [14] in more general situations can be used, although the analysis of TTE presents extra challenges. Therefore, we will emphasize on the first two tasks and specific considerations on TTE analysis. The last topic, evidence generalization, that we will only briefly mention is closely related to the population adjustment above and shares the same methodology toolbox. For example, using RWD as external control can be considered as generalizing the evidence in RWD to the trial population, if they are different.

Recent development on causal estimand in the Neyman–Rubin framework [27, 35] especially in pharmaceutical context [5, 23] also has strong influence on using RWD, since the analysis using RWD should take into account all elements that determine an estimand (chapter "Estimand in Real-World Evidence Study: From Frameworks to Application"), which is more complex for TTE RWD. Here, we use estimands in the narrow sense of being a quantity to be estimated, rather than the more general sense used in the E9(R1) guidance [17]. The estimation of causal estimands using TTE presents additional challenges due to the impact of censoring. The intercurrent event that either changes the treatment or causes missing information is often more common in real-world settings. Although the hazard ratio (HR) is the most commonly used measure of treatment effect of TTE data [6], alternative estimands have been increasingly used. Among them, the restricted mean survival time (RMST) is a common choice, since it does not need a strong assumption such as proportional hazard [6]. Its estimation, especially with adjustment for confounding biases, can be simplified by using pseudo-observations (POs) [2].

Another issue related to causal estimands for TTE RWD is the choice of starting time (time zero). Often, there are multiple choices for time zero; however, a choice may change attributes of an estimand, e.g., the population in which the estimand is defined. The trial emulation approach is a novel approach [13], which we will explore in Sect. 4. The concept of emulating a hypothetical target trial was developed for epidemiology. In our context, the trial to be compared/combined naturally forms the target of emulation, and hence their approaches can be adapted easily. Classical causal inference was mainly developed under the frequentist framework. Nevertheless, Bayesian approaches also have its advantage in its own perspective. For using RWD as external controls, the Bayesian borrowing using power priors [16] has been a hot research area. In addition, some frequentist approaches have their Bayesian interpretation or can be adapted as an approximate Bayesian approach. The adjustment for confounders relies on some untestable assumptions, e.g., no unobserved confounders, which cannot be verified by the data. When using RWD to

augment an internal control arm, a Bayesian borrowing approach can provide extra protection when these assumptions are not valid.

This chapter provides an overview of challenges and some solutions to using TTE RWD in combination with trial data, mainly focusing on using RWD in combination with clinical trial data, and practical implementation of these solutions. We examine additional issues of using HR for RWD and assumptions needed for HR being a valid casual estimand. Then, we explore using alternatives such as RMST as an alternative to HR in the well-developed causal estimand framework. The use of approaches for causal inference such as the PS matching and weighting and g-formula for TTE will be briefly described. We also give a brief overview of recent development on using Bayesian approaches in combination with classical confounding adjustment when using RWD for indirect comparisons and augmenting an internal control arm. The issue of time zero selection, as well as other topics such as using aggregated RWD, will also be discussed.

2 Causal Estimands, Confounding Bias, and Population Adjustment When Using RWD

For notations in this chapter, we denote the event and censoring times, treatment, and covariates of patient i by T_i, C_i, D_i, and $\mathbf{X}_i$, respectively. For simplicity, we assume that T_i and C_i take on integer values 0,1,...; e.g., in days, as they are often recorded in clinical data. Some other advantages of doing so can be seen later. Here, we assume that $i = 1, ..., n$ include patients from the trial and as well as from RWD. The two populations and the treatments are labeled by two indicators H_i and D_i, with $H_i = 1$ if i belongs to RWD and $H_i = 0$, otherwise, and with $D_i = 1$ for the test treatment and $D_i = 0$ for the control. When the RWD is the only control for a single-armed trial, we have $D_i = 1 - H_i$. To focus on population adjustment, we assume non-informative censoring. Therefore, standard survival analysis methods such as the Kaplan–Meier (KM) estimator [19] and Cox regression [6] can be used, with adjustment for covariate-dependent censoring using, e.g., inverse probability of censoring weighting (IPCW) [31], if necessary.

Causal estimands for treatment effects in TTE endpoints are an important but also difficult topic and has been playing an increasingly important role in the estimand framework. The most commonly used estimand for treatment effect for TTE is the HR [6]. It is a single number as an overall measure of the treatment effect and has been widely used for a long time. It may also be used as a population summary: one of the five attributes in an estimand [17]. Nevertheless, it has a number of drawbacks as a causal estimand, even for RCTs [1, 11]. At a given time t, the hazard function under treatment d is defined as $h^d(t) = P(T_i = t | T_i \geq t, D_i = d)$ (since t is an integer), and the HR is $\gamma(t) = h^1(t)/h^0(t)$. The proportional hazard assumption assumes a constant HR at all time, hence $\gamma(t) = \gamma$. Under this assumption, γ can be estimated by the Cox regression model

$$h^d(t) = h_b(t)\exp(\gamma d), \tag{1}$$

where $h_b(t)$ is the baseline hazard, known as the marginal hazard model [25]. However, the assumption is rarely true. A typical scenario is that since $h^d(t)$ is defined in a subpopulation at risk with $T_i \geq t$, we may have $\gamma(t) \neq 1$ because of not only treatment effects but also the population difference between treated and controls in the at-risk set. Suppose that patients can be classified into high- and low-risk groups and the treatment only reduces the risk at $t = 1$ in the high-risk group; hence we have $\gamma(1) < 1$. However, the treated at-risk set at $t = 2$ will have more high-risk patients than that of control group, hence $\gamma(2) > 1$, although the treatment has no effect after $t = 1$. When the assumption of $\gamma(t) = \gamma$ is violated, what the estimated HR based on a Cox model represents is not clear; furthermore, it depends on the censoring mechanism. The latter can be fixed by a weighting approach, but the former remains an unsolved issue. The issue is more complex when using RWD, since the assumption is population dependent. This is because, intuitively, at any event time, the risk of event should be constant among those still at risk, which obviously depends on the population. This issue also occurs in hazard difference as it also depends on comparison based on $h^d(t)$.

Given these issues, we examine causal estimands in the classical Neyman–Rubin framework to look for alternatives. This framework has been well developed in terms of the concept as well as methods of estimation. The reader is referred to chapter "Causal Inference with Targeted Learning for Producing and Evaluating Real-World Evidence" and [5] for details. The most common estimand is the average treatment effect (ATE) defined as

$$\Delta = E(Y_i^1 - Y_i^0), \tag{2}$$

where $Y_i^d, d = 0, 1$, is the (counterfactual) outcome when subject i takes treatment d, but we only observe $Y_i = D_i Y_i^1 - (1 - D_i)Y_i^0$ [35]. Here, the expectation is taken over the whole population of the trial and RWD subjects. Another commonly used estimand is the average treatment effect among treated (ATT), for which the expectation in (2) is taken over the treated ones. ATT is appropriate when we use the RWD as an external control and will be our main focus here. For ATT, $E(Y_i^1)$ can be estimated by the sample mean among the treated. It is $E(Y_i^0)$ that needs population adjustment, if the population of $D_i = 0$ is different from that of $D_i = 1$. Several approaches have been developed for the inference of Δ; see references [14, 15, 22] for details of estimating causal estimands using different approaches.

For TTE, due to censoring, it is not possible to replace Y_i^d with T_i^d in (2) to estimate Δ, the mean survival time difference, without a distributional assumption. Nevertheless, we can use the framework for other TTE estimands. For example, with survival functions $S^d(t) = E(T_i^d > t), d = 0, 1$, where the expectation is taken over the target population, we can take the difference between them

$$\Delta S(t) = S^1(t) - S^0(t) \tag{3}$$

as a causal estimand at a specific time t and it can be estimated without distributional assumptions. As an ATT estimand in (3), $S^1(t)$ can be easily estimated, but the estimation of $S^0(t)$ needs population adjustment. Some other causal estimands can be derived from $S^d(t)$. A commonly used is the difference between RMST for a given end time t

$$\Delta R(t) = \int_0^t S^1(u)du - \int_0^t S^0(u)du. \tag{4}$$

A similar summary to γ is the cumulative hazard ratio:

$$\Gamma(t) = \frac{H^1(t)}{H^0(t)}, \tag{5}$$

where $H^d(t) = \int_0^t h^d(u)du$. $\Gamma(t)$ is always defined and is also a causal estimand. But it depends on the time t and hence has the same problems as the difference in survival rate or RMST at t. One may take t sufficiently large to capture all the cumulative differences in the hazard, but in this way early differences may not be captured. For example, $\Gamma(t)$ at $t = 10$ years could be close to 1 for patients with advanced cancer, and hence it would not be a useful global estimand. Although RMST does not have this problem, it is still t-dependent even when the proportional hazard assumption holds.

In summary, although HR is problematic as a causal estimand, no alternatives have the same simplicity as a single measure. The most appropriate causal estimand depends on multiple factors including clinical relevance, the nature of the data, and often more than one estimand may be needed. For their estimation, some approaches in the rest of the chapter are provided as general tools, and some are specific to a specific choice of estimand.

3 Adjustments for Causal Inference

The estimation of a causal estimand often needs adjustment since treatments are not randomized in real-world settings. Adjustment methods for other types of outcomes can be found in chapters "Clinical Studies Leveraging Real-World Data Using Propensity Score-based Methods" and "Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods". Here, we will concentrate on those related to TTE. We classify the methods into two classes: those specifically for TTE and those independent of the outcomes, although combinations between them are also commonly used.

The first uses an outcome model to directly adjust population differences, also known as g-formula [14]. As it depends on modeling the outcome–covariate relationship, it is sensitive to model misspecification. Here, we describe the method with a Cox model for the outcome. Other models such as the additive hazard models

or parametric models can also be used. Suppose that we would like to estimate the ATT estimand $\Delta S(t)$ (i.e., in the treated population). As $S^1(t)$ can be estimated with normal KM method [19] without adjustment, we only need to adjust control patients in RWD to estimate $S^0(t)$. To this end, we take the following steps:

1. Fit a Cox model $h(t|X_i) = h_b(t)\exp(\boldsymbol{\beta}_0^T X_i)$ to data of control patients and obtain $\hat{\boldsymbol{\beta}}_0$, where $h_b(t)$ is the baseline hazard with a corresponding baseline survival function $S_b(t)$.
2. Calculate $\hat{S}_i^0 = \hat{S}_b(t)^{\exp(\hat{\boldsymbol{\beta}}_0^T X_i)}$ for X_i of all treated ($H_i = 0$) patients, where $\hat{S}_b(t)$ is the estimated baseline survival function from the fitted Cox model in Step 1. Note that $\hat{S}_i^0$ is t-specific.
3. Calculate $\hat{S}^0(t) = \sum_{i=1}^n (1 - H_i)\hat{S}_i^0/n_1$, where n_1 is the number of treated patients in the trial.

In the presence of covariate-dependent censoring, IPCW may be needed to adjust for it by fitting the Cox model with the inverse of the probability of censoring as a function of covariates at a given time as weights [31]. The probability of censoring may be estimated by fitting another Cox model to time to censoring data. A parametric model such as a Weibull model instead of the Cox model can also be used in step 1, which is known as parametric g-formula [14]. The advantage of this approach is that none of the above steps need more than conventional regression tools. However, the uncertainty of estimating $\boldsymbol{\beta}$ should be taken into account. An R-package *stdreg* [40] performs the above steps and provides the SE of $\hat{S}^0(t)$ based on the delta approximation. An alternative is to run a bootstrap over the above steps to obtain either an estimate of the SE or a bootstrapped confidence interval (CI) directly.

In practice, some multivariate analyses involving treatments and covariates are based on a Cox model which includes X_i and d_i

$$h(t|X_i, d_i) = h_b(t)\exp(\boldsymbol{\beta}^T X_i + \gamma^* d_i) \tag{6}$$

in which γ^* is often reported as adjusted treatment effect. However, γ^* is not the same as the marginal HR γ, as the model is conditional on X_i. A correct approach to estimating γ, given the assumption holds, is to fit a marginal Cox model to the predicted survival curves after adjustment for X_i. Suppose that our estimand is the ATT HR, and the following algorithm can be used:

1. Fit model $h(t|X_i) = h_b(t)\exp(\boldsymbol{\beta}_0^T X_i)$ to control patients and obtain $\hat{\boldsymbol{\beta}}_0$.
2. Calculate $\hat{S}^0(t_i) = \hat{S}_b(t_i)^{\exp(\hat{\boldsymbol{\beta}}_0^T X_i)}$ for t_i and X_i of all treated patients.
3. Obtain $\hat{S}^1(t_i)$ from the survival curve of treated patients.
4. Fit a generalized linear model (GLM) with a log–log link function to data $\hat{S}^1(t_i), d_i = 1$, and $\hat{S}^0(t_i), d_i = 0$, with d_i as the only independent variable.

The first two steps estimate $\hat{S}_0(t)$ at multiple time points. The last step utilizes a well-known relationship between the HR (if exists) and survival function, and that will also be used in Sect. 5. One needs to check (e.g., by comparing the

survival curves $\hat{S}^1(t_i)$ and $\hat{S}^0(t_i)$) the proportional hazard assumption. This approach depends on a correct Cox model at the first step too and hence is less robust than the other approaches below.

The second class includes a wide range of standard causal inference methods. The following methods are commonly used:

- Propensity score matching (PSM) and stratification
- Inverse probability weighting (IPW)
- Covariate balancing

The first two [32, 33] are based on PS defined as the probability of subject i being in the RWD, given $\mathbf{X}_i$: $p_i \equiv P(H_i = 1|\mathbf{X}_i)$. In most of our context, this means $p_i = P(D_i = 0|\mathbf{X}_i)$. A logistic regression can be used to obtain $\hat{p}_i$: an estimate of p_i. The above methods have multiple variants. For example, the IPW has a stable version and a truncated version that may perform better under some situations. Matching and stratification can also be based on prognostic factors if feasible. Also we assume covariate-independent censoring, although IPCW can be used for covariate-dependent censoring. We will focus on the basic version and covariate-independent censoring due to space limit.

Using these approaches to estimate some estimands such as $\Delta S(t)$ is straightforward. But the implementation depends on the target population of the estimand. Again, we will consider the ATT estimand with the trial population as the target and RWD are the only source of control. For IPW, we use $w_i = (1 - \hat{p}_i)/\hat{p}_i$ to "generalize" subjects in RWD to the trial population; that is, we weight the RWD KM curve by w_i and take the difference between it and the (unweighted) treated KM curve. For PSM, we match each trial patient with one or more RWD patients based on their PS and then compare the KM curves of the matched patients. Note that weights may be needed if more than one RWD patient are matched to a treated patient. For example, if subject i in the trial population is matched to n_i RWD patients, then the n_i patients should contribute as one, and hence each patient has weight n_i^{-1}.

Covariate balancing [9, 53] is also a weighting approach. For ATT estimands, the weights are determined so that for RWD patients their weighted summary statistics of key prognostic factors match those of trial patients. That is, we find weights w_i such that

$$\sum_{i=1}^{n} H_i w_i X_i = \bar{X}^1, \tag{7}$$

where $\bar{X}^1$ is the mean of X_i over the trial population. Then, the outcome data are weighted with the same weights in the KM estimate [50]. The weights that satisfy (7) are not unique. Hence, we can find the optimal weights that are closest to the uniform weights $w_i = 1/n_0$, where n_0 is the number of patients in RWD depending on the distance to measure the closeness. One common option is the squared distance, while another is the entropy $w_i \log(w_i)$. The latter leads to an

estimator also known as the matching-adjusted indirect comparison (MAIC) [39], which is commonly used in health technology assessment. Also, the latter results in unique non-negative weights and such weights that satisfy (7) may not exist if the distributions of $\mathbf{X}_i$ in the RWD and trial population do not overlap.

For changing the target population to the RWD population, one just needs to switch the role of the two populations. If the target is the joint population of both, the above procedures need some adaptation. For IPW, each patient is weighted by the inverse of the propensity of being in the actual population, i.e., an RWD patient is weighted by $1/p_i$ and a trial patient $1/(1-p_i)$. For matching, a full matching [10] that allows matching one trial patients to one or more RWD patients and vice versa is needed. The matched data should also be properly weighted so that the weighted population represents the target one. For covariate balancing, weights are determined for all patients so that weighted summary statistics of both trial and RWD populations all match those of the target population, that is,

$$\sum_{i=1}^{n} H_i w_i X_i = \sum_{i=1}^{n} (1-H_i) w_i X_i = n^{-1} \sum_{i=1}^{n} X_i. \tag{8}$$

We have only described approaches for $\Delta S(t)$. For other estimands such as RMST, the difference in RMST $\Delta R(t)$ can be estimated by weighted mean difference using the same weights as above. The same approaches can be applied for the estimation of HR, although the assumption of constant HR should hold in the weighted population. Again we only describe the estimation of treatment effect in the trial population here. To estimate HR with IPW, we fit a Cox model with D_i as the only covariate and weight subjects with weights $w_i = (1-\hat{p}_i)/\hat{p}_i$ for those with $H_i = 1$ and $w_i = 1$ for those with $H_i = 0$. The covariate balancing approach can also be applied in the same way as before. For using PSM, we match in the same way as for the KM estimation and then fit the Cox model to the matched dataset. For each method, one should also check the proportional hazard assumption by the weighted or matched Kaplan–Meier curves. Note that for PSM adjustment, one may have the choice of stratifying by matched pair or not. As Austin [4] pointed out, stratification changes the target estimand, although it may gain some power of hypothesis testing. Therefore, he suggests estimating the HR unstratified, while using stratification for hypothesis testing.

Some theoretical issues dealing with censoring and asymptotic properties can be avoided by considering discrete survival time. With this setting, TTE is converted into longitudinal data with survival and censoring times represented by indicators at discrete time points, which fits our assumption of integer t, and hence a wide range of methods for causal inference for longitudinal data can be applied. For example, one can define the hazard and model it at each time point. In more complex situations, a general framework using IPW and g-formula to deal with complex data including competing risks has also been developed [51]. This approach models the events at time $t+1$, given the information upon time t, and then uses either IPW or g-formula to adjust time-varying confounders.

4 The Selection of Time Zero

The starting time or time zero is key to defining the estimand for survival analyses. Although in many situations its selection is straightforward, it could be a real challenge for using TTE RWD (or simply RWD). The following hypothetical example illustrates the difficulty it involves. Suppose that we compare a test treatment in a clinical trial with control treatments (e.g., the standard of care (SoC), which is nonspecific and may consist of multiple treatments) in a disease (e.g., cancer) registry. Similar scenarios can be found in chronic diseases (e.g., type II diabetes and resistant hypertension) for which multiple treatment options are available depending on the progress of the disease. Table 1 shows hypothetical patient records and Fig. 1 presents some examples (see below) graphically. For example, trial patient 1 was diagnosed and treated with control A at day 1 and then switched to control C on day 30. He/she was randomized on day 75, and the trial treatment T started on day 90. The time of randomization could also be the time of enrollment in a single-armed trial. In this scenario, there are two issues with time zero selection. First, if the SoC consists of all controls A,B, C, and D, how should time zero be set for the SoC for registry patients? Second, even if our control treatment is specific; e.g., B, so we may take the start of first dose of B (here, 30 and 120 days for the two RWD patients) as the time zero, since the waiting period between randomization and treatment start (e.g., day 75 to day 90 for RWD patient 1) has no counterpart in RWD, how should we treat them, as treated or untreated with the trial treatment?

Here, we follow the approach of emulating the trial [12], that is, to make the real-world setting as similar as the trial as possible. We will first consider the waiting period issue since it closely relates to the classical immortal time bias [41] with extensive research work done on how to deal with it. There are multiple ways, but none is prefect: (1) excluding it, so time zero is the time of trial treatment start (hence trial patient 1 starts from day 90 rather than day 75), (2) including it, so time zero is the time of randomization, and (3) including it but counting it not under the trial treatment. For (3), a time-varying Cox model approach has been proposed. These approaches change the estimand of the trial, e.g., (1) estimates the ITT estimand and (2) the as-treated one. An approach similar to the trial emulation approach

Table 1 A hypothetical dataset for trial and RWD patients

		Randomization or treatment start and change/days since diagnosis			
Source	ID	1	2	3	4
Trial	1	Control A/1	Control C/30	Randomization/75	Trial treatment/90
Trial	2	Control A/1	Randomization/45	Trial treatment/60	Control C/100
...					
RWD	1	Control A/1	Control B/30	Control C/80	Control D/100
RWD	2	Control C/1	Control A/60	Control B/120	
...					

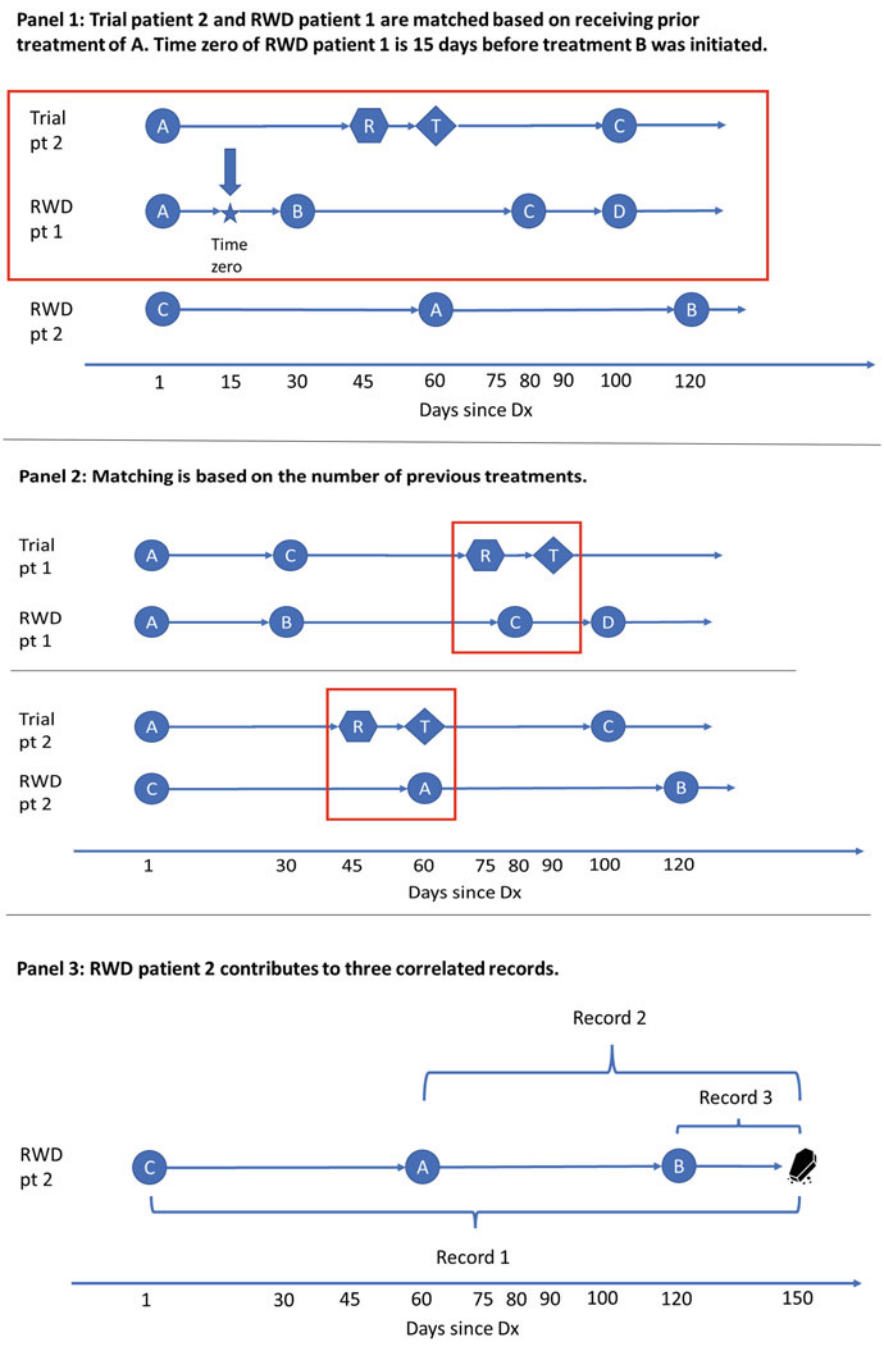

Fig. 1 Three different scenarios of determining time zero

adds comparable waiting times to the RWD data so that they are comparable to the trial data [52]. To this end, for each RWD patient, we sample such a period from the waiting period distribution of trial patients with similar prognostic factors. For example, RWD patient 1 matches to trial patient 2 by previous treatment (A only). Hence, we assume a waiting time 15 days before B and set starting time 15, rather than 30, as shown in Panel 1, Fig. 1. Another approach is landmarking [26]. It uses a fixed landmark time as time zero for all patients. Although a simple approach, its clinical relevance is often questionable.

The issue of waiting time has impact beyond the specific period; it may also induce selection bias because more vulnerable patients in the trial are more likely to become unsuitable to the treatment during the waiting time, and hence those receiving the treatment may have better prognosis than those in the RWD, even if the two populations were similar otherwise. To adjust for this bias, we have to assume no unobserved factor that affects the probability (propensity) of receiving the trial treatment. With this assumption, we can apply the approaches such as PS weighing or matching described in the previous sections. For example, the prevalent new-user cohort designs [42], although proposed for drug safety studies, can be adapted for comparison between RWD and trial data. More details and alternatives will be described later.

Next, we consider the situation of using nonspecific SoC as the control in which there are multiple choices of time zero, that may lead to ambiguity, but may also provide flexibility of making the best choice. For example, patients 1 and 2 in the trial can be matched to patients 1 and 2 in the RWD, respectively, with time zero of 80 and 60 by the number previous treatments and, approximately, days since diagnosis (Panel 2, Fig. 1), while in the situation when the comparator is fixed, previous disease/treatment history may be more difficult to match. Even with the flexibility, adjusting for all individual factors, including previous treatment/disease history, may be impossible. In this case, PS plays a key role here. Again, we consider the trial as a part of the whole cohort, with some patients receiving an additional option: the trial treatment in addition to the controls. Therefore, the PS can be defined as the probability of receiving the trial treatment $q_i = P(\text{Trial treat.}|X_i)$, given all prognostic factors X_i for patient i in the whole cohort. However, when X_i includes time since diagnosis or the number of lines of previous treatments, the propensity depends on the choice of time zero. Hence, we need to define $q_{ik} = P(\text{Trial treat. after kth control}|X_{ik})$, where X_{ik} includes all factors upon the kth control. For example, for RWD patient 1, the days since diagnosis for propensities q_{i0}, q_{i1}, and q_{i2} are 1, 30, and 80, respectively (Table 1).

Based on propensity q_{ik}s, two approaches can be used. The first one matches each trial patient with one or more RWD patients. For example, an optimal 1:1 match method is to find one match for each trial patient among all candidates to minimize a distance between the PS values of the trial and RWD patients, with the restriction that each RWD patient can only be used once. After matching, analysis methods for PSM can be used. This matching approach was first proposed by Suissa et al. [42] for drug safety studies.

Another approach uses multiple starting points for each RWD patient, adjusted by q_{ik} as well as the correlation within patient [12]. For this purpose, n_i records are generated with different time zeros for patient i. For example, as shown in Panel 3, Fig. 1, for RWD patient 2, suppose he had an event on day 150, and the three records have TTE 150, 90, and 30, respectively, with varying treatment history among them. If HR is the estimand of interest (assuming its validity), the Wei, Lin, and Weissfeld approach [49] can be used to take the correlation between, e.g., the three TTEs of RWD patient 2 into account. For this purpose, one can use patient's ID to specify correlated records in SAS PROC PHREG or R-function coxph, weighted by $w_{ik} = q_{ik}/(1 - q_{ik})$ for all RWD patients and 1 for trial patients. Note the difference in weights from those in Sect. 3 is due to the difference in the definition of the PS. If the survival function is the estimand, the weighted KM approach [50] is easy to use, although the CI band needs separate estimation. One practical approach is bootstrapping which resamples at patient level and repeats the estimation B times. Then, the $\alpha/2$ and 1-$\alpha/2$ quantiles of the B-bootstrap sample can be used as the CI at level α.

5 Pseudo-observations: An Approach for Easy Use of Complex Causal Inference Methods

The pseudo-observation (PO) approach is a powerful tool to simplify causal inference using TTE data. [2, 3]. The PO for the survival function of subject i is estimated as

$$\hat{S}_i^d(t) = n_d \hat{S}^d(t) - (n_d - 1)\hat{S}_{-i}^d(t),$$

where n_d is the number of subjects with treatment $D_i = d$ and $\hat{S}^d(t)$ and $\hat{S}_{-i}^d(t)$ are the KM estimators using all n_d subjects and that leaving out subject i, respectively. With POs, $\hat{S}_i^d(t)$, we can derive POs of other estimands. For example, the PO for RMST can be calculated as

$$\hat{R}_i^d(t) = \int_0^t \hat{S}_i^d(u)du. \tag{9}$$

The calculation has been implemented in R-package *pseudo* [20] and is very easy to use. Although covariate-dependent censoring can be adjusted by IPCW for POs [29], *pseudo* does not implement it.

The asymptotic properties of the PO have been examined in [8, 28]. They showed that, with large sample size,

$$E(\hat{S}_i^d(t)|X_i) \approx S^d(t|X_i) \tag{10}$$

Hence, $\hat{S}_i^d(t)$ can be considered as a measure of the counterfactual survival rate at t, for subject i, had (s)he received treatment d. Therefore, general approaches for the estimation of mean difference (2) can be used for $\Delta S(t)$ and $\Delta R(t)$. We again consider using RWD as external control to estimate the ATT in survival rate. As discussed, the key step is to estimate $S^0(t)$ using RWD. For example, with $p_i \equiv P(H_i = 1|\mathbf{X}_i)$, the IPW estimator [3] is

$$n_1^{-1} \sum_{i=1}^{n} \frac{H_i(1-\hat{p}_i)\hat{S}_i^0(t)}{\hat{p}_i}, \tag{11}$$

where $\hat{S}_i^0(t)$ is the PO of subject i in RWD and $\hat{p}_i$ is an estimate of p_i. The same approach can be used for matched or stratified population, with weights determined by the number of matched or stratified patients.

The PO approach can also be used for direct adjustment. For the outcome model for POs, the Cox model can be converted to a GLM with the complementary log–log link function: $g(x) = \log(-\log(x))$ [2]:

$$g^{-1}(E(\hat{S}_i^d(t)|X_i)) \approx \beta^d(t) + X_i^T \boldsymbol{\beta}^d, \tag{12}$$

where $\beta^d(t) = -\int_0^t h_0^d(u)du$ is to be estimated as an unknown parameter. This model can be used to estimate $\boldsymbol{\beta}^d$ in the Cox model for adjustment.

The above model can be used to adjust for X_i for using RWD to estimate $S^0(t)$ in an ATT estimand. First, we calculate $\hat{S}_i^0(t)$ for each i in RWD and fit model (12). Then, a direct adjustment estimator is

$$n_1^{-1} \sum_{i=1}^{n} (1-H_i)\hat{\mu}_i^0, \tag{13}$$

where $\hat{\mu}_i^0 = \exp(-\exp(\hat{\beta}^0(t) + X_i^T \hat{\boldsymbol{\beta}}^0))$ is an estimate of $E(S_i^0(t)|X_i)$ for all trial patients. Combining (11) and (13), one can construct a PO-based DR estimator [46] which is valid when either the PS model or the outcome model is correct. The DR estimator for $S^0(t)$ is

$$n_1^{-1} \sum_{i=1}^{n} \hat{p}_i^{-1}[H_i(1-\hat{p}_i)\hat{S}_i^0(t) - (H_i - \hat{p}_i)\hat{\mu}_i^0]. \tag{14}$$

The model (12) can also be used for estimating the HR by marginal survival function 1 for a specific population. For the estimation of HR in the trial population using IPW, assuming it exists, we calculate $S_i^d(t_k)$, $d = 0, 1$, at multiple time points $t_1, ..., t_K$, then fit the following model with log–log link function:

$$g^{-1}(E(\hat{S}_i^d(t_k)|X_i)) \approx \gamma_0(t_k) + \gamma d_i \tag{15}$$

with weights $w_i = 1$ for all trial patients and $w_i = (1 - \hat{p}_i)/\hat{p}_i$ for RWD patients. This approach uses the same model as the HR estimation algorithm in Sect. 3, but with IPW, rather than the outcome model, for adjustment.

6 Bayesian Approaches for Indirect Comparisons and Augmenting an Internal Control Arm

This section gives a brief review on Bayesian methods for indirect comparisons with RWD as the only control source, as well as using RWD to augment a small trial with an internal control arm, with emphasis on the latter. The estimand can be, e.g., $\Delta S(t)$, defined in the trial population and hence is an ATT estimand. Note that for both tasks, the key step is to use both control sources to estimate, e.g., $S^0(t)$. Bayesian borrowing approaches [7, 16, 44] are powerful tools to properly discount the historical data and to mitigate the impact of insufficient adjustment.

First, we introduce the power prior in Bayesian modeling in a general form in the setting of augmenting trial control arm. Let D^0 and D^h denote data including T_i, C_i, and $\mathbf{X}_i$ from the internal and RWD control data, respectively, θ denotes model parameter, which may be, or contain, the estimand to be estimated, e.g., $S^0(t)$, and $L(\theta; D)$ denote the likelihood function given data D. Our goal is to estimate θ given control data D^0 and D^h. The power prior, conditional on D^h, is formulated as

$$\pi(\theta|D^h, a_0) \propto L(\theta|D^h)^{a_0}\pi_0(\theta), \tag{16}$$

where $0 \leq a_0 \leq 1$ is the power prior (discount) parameter in the likelihood of historical data, and $\pi_0(\theta)$ is the initial prior for θ. The corresponding posterior distribution can be written as

$$\pi(\theta|D^h, D^0, a_0) \propto L(\theta|D^0)\pi(\theta|D^h, a_0) \propto L(\theta|D^0)L(\theta|D^h)^{a_0}\pi_0(\theta). \tag{17}$$

The parameter a_0 allows us to control the contribution of historical data in (17).

The likelihood functions in (17) can be replaced by a partial likelihood function, so the approaches can be applied to TTE analysis using a Cox regression model. We can also use a quasi-likelihood for POs. The following is an example for estimating $S^0(t)$, hence $\theta = S^0(t)$, with POs $\hat{S}_i^h, i = 1, ..., n_h$ from RWD as D^h and $\hat{S}_i^0, i = 1, ..., n_0$ from the trial control arm as D^0. Also, we drop t in the POs to simplify the notation. Although the POs are not exactly binary variables, we can use the binomial distribution as a quasi-likelihood. Therefore, with a prior distribution $S^0(t) \sim Beta(1, 1)$, an approximate posterior distribution is

$$\pi(S^0(t)|D^h, D^0, a_0) \sim Beta(a_0 S_\cdot^h + S_\cdot^0 + 1, n_0 + a_0(n_h - S_\cdot^h) - S_\cdot^0 + 1), \tag{18}$$

where $S_\cdot^0$ and $S_\cdot^1$ are sums of $\hat{S}_i^h$ and $\hat{S}_i^0$, respectively, and a_0 controls the contribution of the RWD data.

The power parameter a_0 can be fixed so that the amount of information borrowed from historical data is independent of the trial data. However, intuitively, one should choose a small a_0 when the survival rates of the trial control arm and RWD are very different and a large one when they are similar. This intuition leads to dynamic borrowing that determines a_0 based on $S_{\cdot}^0$ and $S_{\cdot}^h$. A full Bayesian approach is to include a_0 as a parameter with a prior distribution. Then, the joint posterior distribution can be derived but often needs intensive computation. Empirical Bayes approach is an alternative that estimates a_0 with a marginal likelihood function [7]. For binary outcomes, the marginal likelihood for a_0 is

$$\pi(a_0|D^h, D^0) \propto \frac{Beta(a_0 S_{\cdot}^h + S_{\cdot}^0 + 1, n_0 + a_0(n_h - S_{\cdot}^h) - S_{\cdot}^0 + 1)}{Beta(a_0 S_{\cdot}^h + 1, a_0(n_h - S_{\cdot}^0) + 1)}. \tag{19}$$

Then, we can use a grid search for the a_0 that maximizes (19). Intuitively, a_0 should be high when $S_{\cdot}^h/n_h$ and $S_{\cdot}^0/n_0$ are similar. This is indeed so with (19), although not explicitly in the formula. Dynamic borrowing should be used with care, due to its outcome-dependent nature.

Till now, we have not considered the difference in prognostic factors between the trial and external populations. If the difference in the outcome is mainly due to the difference in these factors, one may determine a_0 based on the latter. For this, we have an easy measure of similarity based on the PS. The determination of a_0 can be combined with adjustment for these factors. To this end, the composite (also known as weighted) likelihood function approach [43] allows subject level weights. The RWD part (16) can be written as

$$\pi(\theta|D_i^h, a_0, w_i) \propto L(\theta|D_i^h)^{a_0 w_i} \pi_0(\theta), \tag{20}$$

where D_i^h represents the data of the ith patient in D^h, and w_i is the corresponding weight. As a frequentist approach, Wang et al. [45] propose to use stratification based on the PS and, within stratum k, the weights are proportional to r_k: the overlapping area between the PS distributions of RWD patients and internal control patients within the stratum k [18]. A Bayesian approach [44] follows the same route, but instead of using fixed r_k, it draws samples from a Dirichlet distribution with the parameters proportional to r_k for the weight. In this way, the amount of borrowing depends on the similarity after the adjustment (here, stratification). Other adjustment methods such as weighting and balancing can also be used. For example, Sachdeva et al. [37] utilize sampling importance resampling method based on propensity score p_i and weight RWD patients with the odds ratio of the probability of being in the study population $(1 - p_i)/p_i$. This approach has a close link to the IPW adjustment, as both use the same weights.

The covariate adjustment approaches can be combined with dynamic borrowing. Intuitively, if the adjustment can reduce the difference between the trial and RWD controls, more can be borrowed. For example, we can adjust $S_{\cdot}^h$ with IPW weighting first and then replace the $S_{\cdot}^h$ in (19) with the adjusted one to determine a_0. An efficient adjustment should lead to a higher a_0. As for statistical inference of the

causal estimand with borrowing, recent development on Bayesian methods without a fully specified model, in particular, weighted/Bayesian bootstrap provides some simple alternatives to the full Bayesian approaches [24]. Wang et al. [48] showed that one can combine the estimation of a_0 together with IPW to obtain approximate posterior distribution of the causal estimand using Bayesian bootstrap.

The Bayesian borrowing approaches generally do not control the frequentist type I error. Due to the potential biases and extra variability induced by dynamic borrowing, the type I error is inflated if the ordinary tests at a given level are applied. Although it is possible to approximately control type I errors by recalibrating the test under restrictive assumptions, in general, it is not possible to always gain power if exact type I error control is needed [21].

7 On the Use of Aggregated RWD

In many situations, the RWD information is given in an aggregated form in publications, e.g., the KM curves or survival rates at given times. Meta-analysis or network meta-analysis, when there are multiple sources of RWD with different treatments, are well-developed methods for aggregated data analysis. Early work was not aimed at causal inference and confounding adjustment, but recently research [30, 38] in this direction is increasing. Although none of them is specifically for TTE analysis, it is possible to adapt these methods for estimating causal estimands similarly as described above, but with assumptions such as no unobserved confounding factors at population level. Schnitzer et al. [38] give a systemic approach to causal inference with meta-analyses.

We start from simple ones that can directly use the methods above and then explore approaches adapted for using aggregated data. Suppose that the RWD information includes a summary of TTE, e.g., a KM curve or an estimate of RMST, the means of covariates $\bar{X}$, and that we would like to compare it with a trial. Then, the covariate balancing method can be used directly to find weights to balance the weighted covariates in the trial. If a collaborator owns the RWD but cannot transfer individual data due to, e.g., data privacy concern, it is also possible to use PSM with aggregated data with Fisher's discriminant as a quasi-propensity score [47]. Both approaches are based on the assumption that aggregated data such as the mean and pooled variance of $\mathbf{X}_i$ of a population determine the probability of the population being the trial or RWD. If aggregated RWD are from publication, in general, RWD are the only feasible target population, since one can only weight trial data to make them similar to the RWD, not vice versa.

When there are multiple ($K > 1$) RWD sources with mean covariates $\bar{X}_k$, it might be possible to target the trial population, depending on not only the population level assumptions but also the distribution of $\bar{X}_k$ and the mean covariates of the trial $\bar{X}_0$. For using IPW, the key assumption is that the propensity of data source $k, k = 0, ..., K$, belonging to RWD depends only on $\bar{X}_k$. For covariate balancing, we can balance $\bar{X}_k$ between the trial and RWD by finding weights w_k so that

$$\sum_{k=1}^{K} w_k \bar{X}_k = \bar{X}_0. \tag{21}$$

For categorical X_is, we balance the proportion of each category, which is the mean of its indicator. This approach not only needs that balancing $\bar{X}_k$s is sufficient but also that $\bar{X}_0$ is not outside of the distribution of $\bar{X}_k$s of RWD, so that non-negative w_ks exist for (21).

The g-formula can also be used with a meta-regression model at the population level. Let $\hat{R}_k$ be the RMST of the kth source, and we can fit model

$$g(E(\hat{R}_k)|\bar{X}_k) = \boldsymbol{\beta}^T \bar{X}_k \tag{22}$$

to the data from the k sources. Then, the RMST adjusted to the trial population can be estimated as

$$\hat{R} = \sum_{k=1}^{K} w_k g^{-1}(\hat{\boldsymbol{\beta}}^T \bar{X}_0), \tag{23}$$

where w_k depends on the variance of $\hat{R}_k$. Although we allow $g(.)$ being a nonlinear model, the adjustment at population level may not be sufficient if $g(.)$ is highly nonlinear. See [38] for details of population level assumptions.

If aggregated RWD data are used for augmenting the control arm of a trial, we may also use dynamic Bayesian borrowing with a power prior, as described in Sect. 6. The amount of borrowing will depend on the similarity of the (adjusted) outcome (e.g., RMST) between the RWD and the control arm. As we have shown with an example, in Sect. 6, the estimation of a_0 only uses aggregated data. This approach provides an extra safeguard against unobserved confounding bias and insufficient adjustment.

8 Other Topics

Several areas we have not touched upon include using RWD only for treatment comparisons and evidence generalization from or to the RWD population. Here, we provide a brief summary on other topics, in particular, evidence generalization, closely related to the key approaches in previous sections.

Evidence generalization provides further insight into approaches above. Suppose that we have evaluated the treatment effect based on RCT, so that the internal validity is guaranteed. But a remaining question is how can the evidence be generalized to a target population? This question is often asked by a payer or a health authority when their population of concern, as described by RWD, is not

represented in the trial. To generalize the RCT evidence to another population involves population difference adjustment used throughout this chapter.

Some general rules for weighting and matching can be summarized as follows: for IPW, (1) weighting by the inverse PS generalizes the effect out of the specific population, (2) weighting by the PS generalizes the effect into a specific population. Combining (1) and (2), we have (3) weighting by the inverse odds of PS generalizes the effect in one population to the other. As an application of (3), the IPW approach for indirect comparison to estimate ATT includes weighting RWD patients out of the RWD population and into the trial population in one step, and hence the weight needed is the inverse odds ratio of the PS of belonging to the RWD. For covariate balancing by weighting, we weight the source population so that their weighted summary of prognostic factors matches to that of the target population. Matching approaches use the same principle; i.e., a patient in the target population is matched to one or more patients in the source population.

All these rules depend on the assumption of overlapped populations. Empirically, it means that the PS, $p_i > 0$, for patients in the source population for IPW, and that for each p_i in the source population, there is at least one $p_{i'}$ that is considered similar to p_i for PSM. For covariate balancing, the lack of overlap between the populations is indicated by no solution to (7) with non-negative w_i.

The g-formula is another way of evidence generalization. For example, after fitting model (12): $g^{-1}(E(\hat{S}_i^d(t)|X_i)) = \hat{\beta}_t^d + X_i^T \hat{\boldsymbol{\beta}}^d$ to POs of the source population, we can use $\hat{\beta}_t^d$ and $\hat{\boldsymbol{\beta}}^d$ to generalize the survival rate under treatment, d, to any population represented with sample $Z_i, i = 1, ..., m$. The generalized survival rate under treatment d is given by the g-formula

$$m^{-1} \sum_{i=1}^{m} g(\hat{\beta}_t^d + Z_i^T \hat{\boldsymbol{\beta}}^d). \tag{24}$$

Unlike the other approaches that rely on overlapped populations, this approach instead relies on a correct model (12).

These approaches not only apply directly to estimators such as (11) but also to likelihood functions and estimating equations. In general, weighting an estimating equation

$$S(\gamma) = \sum_{i=1}^{n} H_i S(T_i, D_i, X_i, C_i) = 0 \tag{25}$$

by $w_i = (1 - p_i)/p_i$, we generalize the RWD to the trial population. When $S(.)$ is the score function of Cox model, this approach gives the HR estimated by RWD but generalized to the trial population. Note that a constant HR may hold in the trial population, but not in the RWD population, and hence the proportional hazard assumption should be checked after weighting.

Another use of RWD is to estimate the effect of a treatment of interest using RWD only. Even for treatment comparison within RWD, as treatments are rarely randomized in RWD, population adjustment is necessary for treatment comparison. Some of the above approaches apply directly by considering the subgroup of patients treated with the test treatment as the trial and those treated with the comparator as the RWD. However, it is important to specify the target population. For example, if a trial is designed to compare treatment A vs C, while treatment B and C have been used in RWD. To make a fair comparison, one may set the trial population as the target and estimate the treatment effect of B vs C in this population such that the B vs C and A vs C effects are comparable.

Some of these approaches can also be used to form a hypothetical estimand of given treatment regimes. For example, if cancer patients in a disease registry start with treatment A, with an option to switch to B upon disease progression, it is often of interest to know the effect of continuing with treatment A in the population that have switched to B. The PS-based and g-formula approaches can be used, but the point to apply adjustment is not the start of treatments, but the time of event (e.g., disease progression) that may trigger the switching. The validity of these approaches depends on the assumption of no unobserved confounding factor upon progression, or that there is an outcome model to correctly predict post-progression survival, none can be verified by the data.

Another important approach is targeted maximum likelihood estimation (TMLE). It is a powerful method for causal inference [15], although initially not developed as so. As it is based on a rather different framework than that this chapter is based, we cannot cover it here and refer the reader to [15] and its references.

9 Summary and Areas of Further Researches

We have given a brief review of the challenges and approaches to using TTE RWD, in particular, for the adjustment of confounding biases due to population differences. The choice of causal estimand for TTE as well as time zero are difficult tasks and should be dealt with care. Multiple approaches for causal inference can be used. Among them, some are universal and outcome independent, such as IPW, PSM, and covariate balancing. However, we should pay attention to using them for specific estimand, e.g., the assumption of proportional hazard. Some other approaches depend on outcome modeling such as the g-formula approach and hence are TTE specific. The doubly robust approaches are a combination of the two, and the use of POs makes the combination easy.

The challenges of using TTE RWD indicate further work is needed in multiple areas. We only covered Bayesian approaches in the last section. Although the majority of population adjustment methods are frequentist and are non-likelihood based, recent development shows Bayesian adaptation of such approaches, such as Bayesian IPW and DR estimators [36]. Approximate posterior sample can be obtained via some simple approaches such as the Bayesian bootstrap. Further simulation for evaluation as well as practical experiences on using them is needed.

Dealing with missing data is another important topic that we did not cover here, and it is a topic worth further research. See chapter "Assessment of Fit-for-Use Real-World Data Sources and Applications" for details. Missing is more common in real-world settings than in clinical trials, and hence its impact should be considered carefully. Some issues may be more difficult to deal with for TTE analysis. Although general approaches based on assumptions such as missing at random can still apply, it may be more difficult to justify due to the limitation of data capture by RWD.

The use of POs allows simple implementation of causal inference approaches to TTE data. To our best knowledge, comparison between PO-based and standard survival analysis approaches has not been extensively performed, especially in RWD settings. Multiple alternatives to HR have been proposed and increasingly used in practice. As none of them is a single number measure, multiple estimands such as survival rate or RMST differences at multiple time points may be needed. Although methods of estimation and inference have been developed, further practical experience is needed for practical use of them.

In summary, there are multiple statistical challenges for using TTE RWD, which also present opportunities for developing and implementing more advanced methods in practice. This chapter only covered a small fraction of them, but we hope it will motivate the reader to look into these important topics from practical and methodological aspects.

References

1. Aalen, O. O., Cook, R. J., & Røysland, K. (2015). Does Cox analysis of a randomized survival study yield a causal treatment effect?. Lifetime data analysis, 21(4), 579–593.
2. Andersen, P. K., Klein, J. P., & Rosthøj, S. (2003). Generalised Linear Models for Correlated Pseudo-Observations, with Applications to Multi-State Models. Biometrika, 90(1), 15–27.
3. Andersen, P. K., E. Syriopoulou, and E. T. Parner (2017). Causal inference in survival analysis using pseudo-observations. Statistics in Medicine 36 (17), 2669–2681.
4. Austin P. C. (2014). The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. Statistics in medicine, 33(7), 1242–1258.
5. Chen, J., Scharfstein D., Wang, H. et al. (2022). Estimand in real-world evidence studies. Statistics in Biopharmaceutical Research. under review.
6. Cox, D. R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2), 187–220.
7. Gravestock, I., Held, L., & COMBACTE-Net consortium (2017). Adaptive power priors with empirical Bayes for clinical trials. Pharmaceutical statistics, 16(5), 349–360.
8. Graw, F., Gerds, T. A., and Schumacher, M. (2009), On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis* 15: 241–255.
9. Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political analysis, 20(1), 25–46.
10. Hansen, B. B. (2004). Full Matching in an Observational Study of Coaching for the SAT. Journal of the American Statistical Association, 99(467), 609–618.
11. Hernán M. A. (2010). The hazards of hazard ratios. Epidemiology, 21(1), 13–15.

12. Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R., & Shrier, I. (2016). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. Journal of clinical epidemiology, 79, 70–75.
13. 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.
14. Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.
15. Ho, M et al. (2021): The Current Landscape in Biostatistics of Real-World Data and Evidence: Causal Inference Frameworks for Study Design and Analysis, Statistics in Biopharmaceutical Research.
16. Ibrahim, J. G., and Chen M.-H. (2000). Power prior distributions for regression models. Statistical Science, 15, 46–60.
17. ICH. (2017) ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials.
18. Inman, H. F., and E. L. Bradley Jr. (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. Communications in Statistics-Theory and Methods 18 (10):3851–3874.
19. Kaplan, E. L.; Meier, P. (1958). Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc. 53 (282): 457–481.
20. Klein J.P., Gerster M., Andersen P.K., Tarima S., Perme, M. P.(2008) SAS and R Functions to Compute Pseudo-values for Censored Data Regression. Comput. methods programs biomed. 89 (3): 289–300.
21. Kopp-Schneider, A., Calderazzo, S., & Wiesenfarth, M. (2020). Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control. Biometrical journal 62(2): 361–374.
22. Levenson, M et al. (2021): Biostatistical Considerations When Using RWD and RWE in Clinical Studies for Regulatory Purposes: A Landscape Assessment, Statistics in Biopharmaceutical Research. DOI: https://doi.org/10.1080/19466315.2021.1883473
23. Lipkovich I et al. (2020) Causal Inference and Estimands in Clinical Trials, Statistics in Biopharmaceutical Research, 12:1, 54–67,
24. Lyddon, S. P., Holmes, C. C., & Walker, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. Biometrika, 106(2), 465–478.
25. Mao, H., Li, L., Yang, W., & Shen, Y. (2018). On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. Statistics in Medicine, 37, 3745–3763.
26. Mi, X., Hammill, B. G., Curtis, L. H., Lai, E. C.-C., and Setoguchi, S. (2016) Use of the landmark method to address immortal person-time bias in comparative effectiveness research: a simulation study. Statist. Med., 35: 4824–4836.
27. Neyman J. (199) On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Statistical Science 5: 465–472.
28. Overgaard, M., Parner, E.T., Pedersen, J., 2017. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. Ann. Statist. 45 (5), 1988–2015.
29. Overgaard, M., Parner, E. T., & Pedersen, J. (2019). Pseudo-observations under covariate-dependent censoring. Journal of Statistical Planning and Inference, 202, 112–122.
30. Phillippo, D. M., Dias, S., Ades, A. E., Belger, M., Brnabic, A., Schacht, A., Saure, D., Kadziola, Z., & Welton, N. J. (2020). Multilevel network meta-regression for population-adjusted treatment comparisons. Journal of the Royal Statistical Society. Series A, 183(3), 1189–1210.
31. Robins, J.M. and Finkelstein, D.M. (2000), Correcting for Noncompliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests. Biometrics, 56: 779–788.
32. Rosenbaum, P. R., & Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55.
33. Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician, 39(1), 33–38.

34. Rosenbaum, P. R. (1987). Model-Based Direct Adjustment. Journal of the American Statistical Association, 82(398), 387–394.
35. Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688–701
36. Saarela, O., Belzile, L. R., and Stephens, D. A. (2016). A Bayesian view of doubly robust causal inference. Biometrika, 103, 667–681.
37. Sachdeva, A., Tiwari, R. C., & Guha, S. (2022). A novel approach to augment single-arm clinical studies with real-world data. Journal of biopharmaceutical statistics, 32(1), 141–157.
38. Schnitzer, M., Steele, R., Bally, M. & Shrier, I. (2016). A Causal Inference Approach to Network Meta-Analysis. Journal of Causal Inference, Vol. 4 (Issue 2), pp. 20160014.
39. Signorovitch, J. E., Sikirica, V., Erder, M. H., Xie, J., Lu, M., Hodgkins, P. S., Betts, K. A., & Wu, E. Q. (2012). Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. Value in health 15(6), 940–947.
40. Sjölander A. (2016). Regression standardization with the R package stdReg. European journal of epidemiology, 31(6), 563–574.
41. Suissa, S. (2008). Immortal time bias in pharmaco-epidemiology. Am J Epidemiol. 167, 492–499.
42. Suissa, S., Moodie, E. E., & Dell'Aniello, S. (2017). Prevalent new-user cohort designs for comparative drug effect studies by time-conditional propensity scores. Pharmacoepidemiology and drug safety, 26(4), 459–468.
43. Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. Statistica Sinica, 21(1), 5–42.
44. Wang, C., Li, H., Chen, W. C., Lu, N., Tiwari, R., Xu, Y., & Yue, L. Q. (2019). Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. Journal of biopharmaceutical statistics, 29(5), 731–748.
45. Wang, C., Lu, N., Chen, W. C., Li, H., Tiwari, R., Xu, Y., & Yue, L. Q. (2020). Propensity score-integrated composite likelihood approach for incorporating real-world evidence in single-arm clinical studies. Journal of biopharmaceutical statistics, 30(3), 495–507.
46. Wang, J. (2018). A simple, doubly robust, efficient estimator for survival functions using pseudo observations. Pharmaceutical Statistics 17 (1), 38–48.
47. Wang, J, & Marion-Gallois, R. (2022) Propensity score matching and stratification using multiparty data without pooling. Pharmaceutical Statistics. 1-16. https://doi.org/10.1002/pst.2250.
48. Wang, J., Zhang, H. & Tiwari, R. (2022) A propensity-score integrated approach to Bayesian dynamic power prior borrowing. Under review. arXiv:2210.01562.
49. Wei, L. J., Lin, D. Y., & Weissfeld, L. (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. Journal of the American Statistical Association, 84(408), 1065–1073.
50. Xie, J., & Liu, C. (2005). Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. Statistics in medicine, 24(20), 3089–3110.
51. Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J., & Hernán, M. A. (2020). A causal framework for classical statistical estimands in failure-time settings with competing events. Statistics in medicine, 39(8), 1199–1236.
52. Zhou, Z., Rahme, E., Abrahamowicz, M., & Pilote, L. (2005). Survival bias associated with time-to-treatment initiation in drug effectiveness evaluation: a comparison of methods. American journal of epidemiology, 162(10), 1016–1023.
53. Zubizarreta, J. R. (2015). Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data. Journal of the American Statistical Association, 110(511), 910–922.

Sensitivity Analyses for Unmeasured Confounding: This Is the Way

Douglas Faries

1 Introduction

Having the spotlight of healthcare decision makers focused on real-world evidence (RWE) has helped raise broader awareness of long understood challenges in the analysis of real-world data (RWD). For instance, the lack of randomization in RWD brings the potential for bias into any comparisons between groups or interventions of interest. In real-world settings, patients are assigned to treatment group based on many factors including patient and physician preferences, access/insurance pricing issues, patient history, baseline patient characteristics, severity of disease, safety profile of treatments under consideration, and concomitant medications. If any of these factors is also related to the outcomes of interest – such as the cost of care, medication persistence, or other clinical outcomes – then such variables are 'confounders'. Any analyses that do not appropriately account for all confounding variables will be biased. Commonly used methods such as propensity score matching can account only for confounding variables that are included in the analysis database, but any confounders not contained in the database are 'unmeasured confounders' and may result in a biased treatment effect estimate.

The lack of randomization is not the only issue healthcare decision makers have with RWE. Data quality issues is another area of concern – as some RWD such as healthcare claims data is collected for reasons other than research and undergoes fewer validity checks and open to measurement biases. For these reasons, much of the focus of the initial wave of FDA guidance on the use of RWE [1–3] has largely focused on ensuring quality of the RWD is sufficient for the decisions being made on the results (fit for purpose).

D. Faries (✉)
Eli Lilly and Company, Indianapolis, IN, USA
e-mail: faries_douglas_e@lilly.com

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_14

Due to the complexity in identifying and accounting for potential biases in RWE, it is easy for healthcare decision makers simply to dichotomize their view of research quality based solely on the presence of randomization (randomized controlled trials, or RCT, vs RWE) and distrust all RWE. The validity of RWE relies on more assumptions than RCT research and without further insights a distrust of RWE is an understandable position. However, as with RCT research, the quality of RWE can vary widely based on study design, data and analytics quality, and the question of interest. This raises the question of how researchers should react to RWE and how can we tell which RWE is actionable and more robust?

Here we focus on the challenging case of comparative analyses based on RWD and the issue of unmeasured confounding. Specifically, we deal with research comparing outcomes between two (or more) groups of patients who initiate different treatment strategies – hoping to make some form of causal inference regarding the potential interventions based on real-world data.

2 Causal Inference and Key Assumptions

To better understand the problem and analytical solutions for unmeasured confounding, a brief refresher on causal inference is needed. We will follow the Rubin Causal model (or Neyman–Rubin model of causal inference) – a framework based on potential outcomes [4] developed across multiple articles by Donald Rubin [5, 6] and summarized by Holland [7]. In short, the potential outcome for each patient (given a particular treatment) is the outcome that they would have had if they had initiated that treatment. Of course, only one potential outcome is observed for each patient and potential outcomes under treatments possibilities they did not take are missing and hence called 'counterfactual'. Formally, each subject i has two (or more) potential outcomes, $Y(T = 0)$ and $Y(T = 1)$, denoting the outcomes the patient would have had given they were assigned to Treatment 0 and Treatment 1, respectively (and ignoring a subscript i to denote the patient). The quantity $Y(T = 1) - Y(T = 0)$ denotes the individual causal treatment effect between two treatment options. Of course, this quantity cannot be observed as a single patient cannot follow two treatment regimens. However, we can estimate, under a set of assumptions discussed below, the quantity $E[Y(T = 1)] - E[Y(T = 0)]$ across a population of patients. A typical goal is to then estimate the average causal effect across an entire population – often referred to as average treatment effect (ATE). Of course, one could be interested in other estimands, such as a treatment effect across a different population such as the treated group, but for the purposes of this chapter, we will not need this detail.

Over the past decades, propensity score-based analyses, such as propensity score matching, have become the gold standard analytical method for causal inference from RWD. For valid causal inference from such analyses, one must make several

assumptions. The first is referred to as the stable unit treatment value assumption (SUTVA). This states that the potential outcomes for any subject do not vary with the treatment assigned to other subjects, and there are no different forms or versions of each treatment level which would lead to different potential outcomes. Second, one must also assume that the data is valid – that there are no systematic errors in variables used to identify patient populations (such as algorithms based on diagnostic codes in claims databases), covariate, and outcomes measured without error. In many research settings, the SUTVA assumption is deemed acceptable as treatments are well-defined interventions and there are limited interactions between patients. Also, let us assume we have collected or identified a quality set of data and assume the models we use in the analyses correctly describe the true data generation process. One still must make additional key analytic assumptions for valid causal inference:

1. *Positivity*: the probability a patient is assigned to either treatment group, given a set of pretreatment covariates, is strictly between 0 and 1 ($0 < P(T|X) < 1$, for all X).
2. *Unconfoundedness*: the assignment to treatment for each subject is independent of the potential outcomes, given a set of pre-treatment covariates. In practice, it means 'no unmeasured confounders'. That is, data on all potential confounders have been collected and used appropriately in the analysis.

If these assumptions hold, even in a non-randomized study, analyses such as propensity score-based methods are able to provide unbiased estimate of the causal effect of the estimand of interest.

Positivity can be assessed by a thorough examination of the baseline characteristics (assuming no unmeasured confounders) and accounted for via trimming (though this will impact the estimand through the target population of inference). The correctness of statistical modeling is challenging, though recent use of machine learning based concepts in causal analyses are providing a greater check of robustness to this assumption [8]. However, producing convincing evidence surrounding the potential impact of unmeasured confounding has proven to be a more complex challenge.

The remainder of this chapter focuses on sensitivity analyses surrounding the assumption of 'no unmeasured confounders'. This is a core assumption and one can never have absolute proof that there is no bias from unmeasured confounding in RWE research. While many analytical challenges remain, growth in research and programs for implementation of unmeasured confounding sensitivity analysis have made this an area where we can now provide decision makers with information on the robustness of any RWE. Such sensitivity analysis is critical for moving us toward optimal use of RWE and better decision making from RWE in the healthcare industry.

3 Current State

Where do we stand today regarding unmeasured confounding analyses? Past work has clearly demonstrated the need for better methods to address the potential for unmeasured confounding. Researcher from the OMOP initiative demonstrated that the operating characteristics of comparative RWE using commonly applied methods may be severely impacted by unmeasured confounding bias. Schuemie and colleagues [9] work estimated that over 50% of statistically significant findings would be non-significant after appropriate calibration and negative controls were found to produce significant findings in 18% of cases [10]. Despite this, the use of quantitative sensitivity for unmeasured confounding has lagged [11], with many publications mentioning this potential bias as a limitation but not providing evidence on the robustness of the finding to the potential bias. A literature survey by Blum et al. [12] found that about 75% of observational research failed to mention the potential for unmeasured confounding.

However, this is slowly changing. Publications such as Zhang et al. [13] demonstrate how use of such methods can change inferences from statistical significance in one direction to statistical significance in the opposite direction when information on a key unmeasured confounder is collected. Other publications have utilized unmeasured confounding sensitivity analysis methods to show both stronger robustness [14] as well as a lack of robustness [15, 16]. Development of new methods such as the E-value, which is broadly applicable, along with a cadre of R-packages have prepared this space for a transformation from 'just state unmeasured confounding as a limitation' to 'apply commonly used of quantitative analyses following best practices in the literature'. We discuss several best practice proposals below.

3.1 Some Notation

While it is likely that some level of unmeasured confounding exists in all non-randomized research, the key challenge for an analyst is to understand whether the unmeasured confounding is of sufficient strength to change the inference that one is concluding from the RW-based comparative analysis. In Fig. 1a, we have a directed acyclic graph (DAG) portraying a simple confounding situation. The variable(s) X is a confounder as it influences treatment selection Z and outcome Y (independent of Z). We assume in this case that X represents a baseline confounder that is measured in the study. Standard bias adjustment methods – such as propensity score matching – will remove a significant portion of any bias caused by X.

Now consider the DAG portrayed in Fig. 1b. This describes the more likely scenario where in addition to the confounder X, we have a variable (or variables) U which denote covariates NOT collected or available for use in the analysis for the study. U is also a confounder as it influences both treatment selection and outcome.

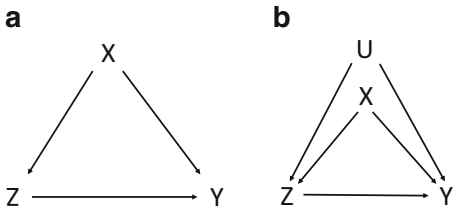

Fig. 1 Directed acyclic graphs portraying standard confounding (graph **a**) and both measured and unmeasured confounding (graph **b**)

In the scenario of Fig. 1b, standard analyses such as propensity score matching will be biased by some unknown amount. How much bias the analysis contains is driven by two factors (denoted by the two arrows extending from U in the DAG): (1) the strength of the influence of U on Z and (2) the strength of influence of U on Y. These two factors are the key quantities in sensitivity analysis methods such as the E-value, rule out, and simulation-based approaches discussed below.

Of course, the situation could be more complex as U and X could be related. However, for most sensitivity analysis approaches, we will make the simplifying assumption of independence of X and U. While technically this is unlikely to be true, we can simply conceptualize U as denoting the remaining impact of the unmeasured confounding not captured through any relationship with measured covariates X.

4 Methods for Unmeasured Confounding Sensitivity Analyses

Three literature reviews [17–19] summarized the various methodology available for sensitivity analysis for unmeasured confounding. One challenge in selecting a method or comparing between methods is due to the simple fact that many methods apply only in specific settings. That is, the applicability of each method depends on the type of information available on the unmeasured confounder. Scenarios include those where there is *no additional information* on unmeasured confounding – one may not even have an idea of what the unmeasured confounder may be. Alternatively, one may be able to identify a specific unmeasured confounder and information about the strength of confounding (relationship to treatment selection and/or relationship to outcome) is available *externally* (patient-level data or summary-level information from patients other than those in the current research dataset). Lastly, a researcher may be able to identify a specific unmeasured confounder and information about the strength of confounding can be obtained *internally* – that is, information on the unmeasured confounder can be obtained from a subset of the patients in the current study via linking, surveys, chart reviews, or some other method. For this reason, there is a noted lack of comparisons of operating characteristics between available methods in the literature.

The Zhang et al. review article [19] provided an overview of available methods stratified by the type of information on the unmeasured confounder necessary for

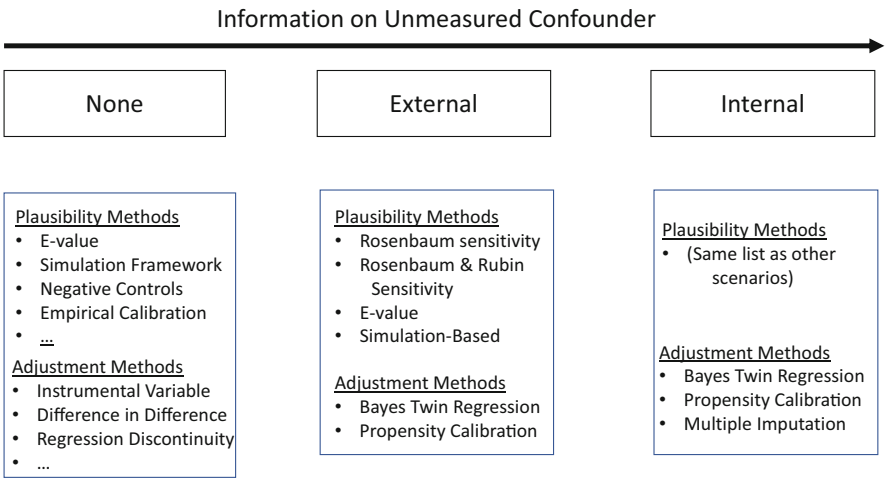

Fig. 2 Overview of methods for sensitivity analysis for unmeasured confounding

implementation. Figure 2 is an adaptation of that work – updated to include select publications since 2018 – and with the three columns representing the types of available information (None, External, Internal).

A second differentiator (incorporated into Fig. 2) among existing methods is that some simply provide bounds for how much of an impact unmeasured confounders could have (*plausibility methods*) and some provide treatment effect estimates adjusted for an unmeasured confounder (*adjustment methods*). In the former case, the goal is often to understand how much confounding would be required to change inferences – such as from an observed statistically significant finding to a non-significant result. This is the case with the Rule Out [20] and E-value [21–23] approaches. Conceptually, if the amount of confounding required to produce the observed effect is very high, then the observed result is considered more robust than in scenarios where it is low.

In the latter case (adjustment methods), as with propensity score calibration or Bayesian twin regression modeling, one uses additional information about the unmeasured confounder to adjust the treatment effect estimate. This additional information makes it possible to directly estimate the impact of the confounding and provide greater amount of information on the robustness of the RWE. For instance, Zhang et al. [13] provided an example of gathering external information to re-analyze a claims base analysis where baseline BMD was a known unmeasured confounder. They utilized data on BMD from a prospective observational study, the literature (information on the relationship between BMD and fracture rates), and survey data all incorporated into Bayesian regression models to produce a treatment effect estimate adjusted for BMD information. Similarly, Faries et al. [14] obtained internal information on HbA1c values – considered to be a potential unmeasured confounder – for a subset of patients from a linked file to supplement a claims-

based comparative analysis of costs of care. They demonstrated the use of Bayesian modeling, multiple imputation, and propensity score calibration to provide adjusted treatment effect estimates.

5 Advances in Broadly Applicable Methods

Since the publication of the review articles [17–19], several additional promising methods have been put forward that are broadly applicable requiring no knowledge of specific unmeasured confounders. First, VanderWeele and colleagues, in a series of publications, put forth the notion of the E-value, or 'evidence for causality' [21–23]. The E-value is defined by VanderWeele as ' ... the minimum strength of association on the risk ratio scale that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment-outcome association, conditional on the measured covariates'. The strength of the impact of the unmeasured confounder is based on the two correlations discussed in Fig. 1: the influence of the unmeasured confounder on the treatment selection and on the outcome of interest. The strengths of this approach include its simplicity and broad applicability. The E-value can be computed even without identifying a specific known unmeasured confounder (thus applies in any of the settings of Fig. 2). The approach makes the simplifying assumption that the relationship of the unmeasured confounder on the treatment and on the outcome is the same. This allows the result to be a single number statistic. Calculations are easily implemented in R-package 'E-value' [24] and the strength of confounding is presented on a risk ratio scale. Without the simplifying assumption of equal relationships mentioned above, one can no longer compute a single summary statistic but the same concepts can be achieved through contour or other plots as outlined later in this work.

For a risk ratio outcome, RR > 1 (take the inverse if RR < 1) the E-value is calculated as $\text{Evalue} = \text{RR} + \sqrt{\text{RR} \cdot (\text{RR} - 1)}$, where RR is the observed RR. Approximations are available if the outcome is based on a continuous, odds ratio, or hazard ratio scale. Typically, researchers will be interested not just in what degree of confounding would account for the observed treatment effect, but what strength of confounding would reverse/render insignificant the inference drawn from the analysis. Thus, it is recommended that one also compute an E-value for the lower confidence limit of the treatment effect (assuming a statistically significant finding where a higher value is greater effect). This would provide the amount of confounding necessary to eliminate the statistical significance of the finding and is helpful for establishing the robustness of the observed effect.

What is a sufficiently high E-value to claim robustness? VanderWeel and Mather [23] note there are no fixed 'cut off' scores for interpretation as what is considered robust and that the decision is context dependent. Of course, the observed effect size drives the size of the E-value. However, in some settings such as analyses from healthcare claims databases, the analysis may be missing

known strong confounders. In other cases, a researcher may have had the ability to do extensive prospective data collection and conducting research in a field with multiple prior studies and well-established knowledge on factors driving the outcome and treatment selection. Thus, the possibility for stronger unmeasured confounders can vary greatly, which affects the interpretation whether an observed E-value is a signal of robustness or not. To help with interpretation of the E-value (e.g. understanding of whether the E-value is signaling a robust finding or not), a couple of recommendations are given. If an unmeasured confounder is identified and external data from the literature or other studies is available (such as the strength of relationship between the confounder and outcome or the confounder and treatment selection), this can greatly help the interpretation of the E-value. While one has to address potential issues with transportability when using information from other data sources, this can provide strong evidence regarding the expected direction and strength of confounding. If an unmeasured confounder has not been identified or no information is available, one can compare the E-value to the strength of confounding from the strongest measured confounding variable. In many research settings, a strong potential confounder is collected and used in the analysis, so one can compare the strength of this confounder to the E-value. If the E-value is larger, then this is evidence for robustness as any unmeasured confounding would need to be stronger than the strongest known confounders in order to change inferences from the results.

A second recent promising and broadly applicable method is the Simulation Framework [25, 26]. Following the notation of Dorie et al. [25], we introduce the following formulas:

$$Y \mid X, U, Z \sim N\left(\beta X^{y} + \zeta^{y} U + \tau Z, \sigma_{Y}^{2}\right)$$

$$Z \mid X, U \sim Bernoulli\left(\Phi\left(\beta X^{Z} + \zeta^{Z} U\right)\right)$$

$$U \sim Bernoulli\left(\pi^{y}\right)$$

where Y is the outcome, X a set of measured covariates, U a single unmeasured confounder, Z denoting a binary treatment option, and with Φ denoting the standard normal cumulative distribution function. Y depends upon the confounders X and U along with treatment Z, while of course the potentially biased analysis has proceeded without information on U. As U is a confounder, treatment also depends upon X and U. For simplification, U is assumed to be a binary variable. The concept behind the simulation-based approach is to first specify a grid of plausible values for ζ^{y} and ζ^{Z} which characterize the level of strength of the unknown confounder (again think back to the drivers of the strength of unmeasured confounding in Fig. 1b). Given each specified pair of values for ζ^{y} and ζ^{Z}, one then can sample values of U from the distribution of U conditional on the observed data and conduct analyses as if U were a measured covariate. They derived the conditional distribution for $U|Z, Y, X$, and thus sampling is consistent with the observed data Z, Y, and X. For each pair of values of ζ^{y} and ζ^{Z} one then generates an average treatment effect adjusting for the assumed level of confounding, using the same scale and models as for the actual

data analyses. Results are presented graphically in a contour plot (as discussed later). This is implemented in the R-package 'treatSens'.

This simulation approach provides several advantages. First, the confounding and results are provided on the same scale as the original analysis (no need to transform to a RR scale). It also provides adjusted treatment effect estimates for any combination of the two driving factors from Fig. 1 (a) strength of relationship between U and Z (b) strength of relationship between U and Y. Of course, the result is presented either in a table or graphical format (contour plot) and is no longer represented by a single summary statistic as is the E-value. Like the E-value this approach can be applied in all scenarios – even if a specific unmeasured confounder is not identified. However, in the case where a specific unmeasured confounder is identified and external information is available, one can also use this approach to generate a treatment effect adjusted for the unmeasured confounder.

A third recent approach of note is the work of Cinelli and Hazlett [11]. They expand on the omitted variables framework (as also in [25]) and utilize the partial R^2 statistic, which is available from standard software output, to describe the strength of evidence. They propose the Robustness Coefficient, RV_q, as an alternative to the E-value. This is defined as the strength of the relationship between the unmeasured confounder and both treatment and outcome needed to reduce the treatment effect by $100 \times q\%$ (to eliminate the treatment effect altogether we would select $q = 1$). As with the E-value, both relationships are assumed to be equal, here based on a percentage of variance explained approach with $R^2_{Y\sim U|X,Z} = R^2_{Z\sim U|X} = RV_q$. The Robustness Coefficient can be calculated as $RV_q = \frac{1}{2}\left\{\sqrt{f_q^4 + f_q^2} - f_q^2\right\}$,where $f_q = q|f_{Y\sim Z\,|\,X}|$ is the partial Cohen's f statistic (Cohen 1988) of the treatment with the outcome times q, which is a proportion of reduction of the treatment effect. Typically one selects $q = 1$ to represent the strength of confounding needed to explain fully the observed treatment effect and the value of q such that the lower confidence limit becomes zero. A high Robustness Coefficient (near 1) means that it would take an unmeasured confounder that could explain almost all of the residual variance in both the outcome and treatment models in order to change the inferences from the study.

To visualize the impact of various levels of confounding on the treatment effect estimate, they utilize contour plots. The X axis describes the strength of relationship between unmeasured confounder and treatment using the partial R^2 of the unmeasured confounder with treatment ($R^2_{Z\sim U|X}$). This is driven by the imbalance between groups in the unmeasured confounder. The Y axis depicts the strength of relationship between unmeasured confounder and outcome using the partial R^2 of the unmeasured confounder with outcome ($R^2_{Y\sim U|Z,X}$). The contour plot (see Fig. 3 for a hypothetical example with an observed treatment difference around 1.5) then displays the treatment effect estimate adjusted for an unmeasured confounder with strength denoted on each axis. For example, an unmeasured confounder would need to explain about 25% of the residual variation in both the outcome and treatment models to move the estimated treatment effect to zero. These plots can also denote the points at which the unmeasured confounding would be

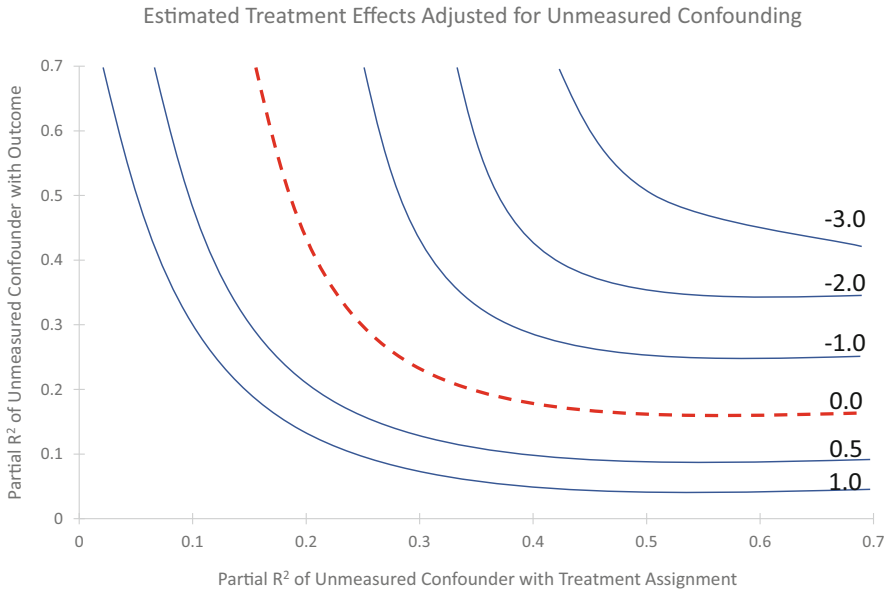

Fig. 3 Example unmeasured confounding sensitivity contour plot

strong enough to change inferences (calculations based on the lower CI instead of the treatment effect), as well as denoting where either external information on specific unmeasured confounders would fall on the plot and where strong measured confounders fall on the plot.

This formulation offers another option of extreme confounding – where researchers can assume that $R^2_{Y\sim U|Z,X} = 1$ (extreme case where the unmeasured confounder accounts for all of the residual variability) – and then assess how strongly would a confounder have to be associated with treatment in order to change inferences. This is helpful when researchers may have some knowledge of the relationship between the unmeasured confounder and treatment but not outcome. Also, to help researchers assess the plausibility of whether confounders could exist, they propose quantifying the strength of evidence in terms of multiples of the strength of an observed strong measured confounder. They note this approach has benefits over bounding based on regression coefficients as in Carnegie et al. [26] by demonstrating a characterization of Z and X that the previous approach would fail to recognize as sufficient confounding to eliminate an observed effect.

6 Proposed Best Practice

More and more authors are urging for a systematic implementation of sensitivity analyses for unmeasured confounding [11, 21, 27–29]. We are in full agreement

with this. In fact, proposals appear to be zeroing in on a consensus of concepts though not yet on specific methodology. VanderWeele and Ding [21], noting the broad applicability and simplicity of the E-value, proposed it as a standard analysis that could be completed alongside of every comparative analysis based on non-randomized data.

Zhang, Stamey, and Mather [28] proposed a best practice flowchart in order to guide researchers on how to perform sensitivity analyses for unmeasured confounding. The proposal recommends that an E-value be the initial sensitivity analyses for all comparative real-world studies as it applies across all research settings. Then, one moves on to methods requiring additional information only if necessary. While the details of the process are contained in Fig. 1 of their paper, at a high level, the steps are as follows:

1. Compute the E-value for both the treatment effect estimate and the key confidence interval limit in order to assess the strength any unmeasured confounder would need to have in order to change inferences being made from the study. If the E-value is larger than any plausible confounder, then stop.
2. If the unmeasured confounder of concern is a binary variable with known very low or very high prevalence, then perform a Rule Out analysis. This is basically a second assessment of plausibility to see if additional more complex work is warranted.
3. If plausibility analyses fail to demonstrate robustness, then additional information regarding the confounder will likely be necessary. Follow the guidance of Zhang et al. [19] to select from among potential methods given the type of information that can be gathered. For instance, if the strength of the relationship of a known confounder with either the outcome or treatment choice is available in the literature, then Bayesian Twin Regression modeling [14] could incorporate the information and re-estimate the treatment effect.

Cinelli and Hazlett [11] propose a three-step process:

1. Compute the strength an unmeasured confounder would need to have (in terms of influence on both treatment and outcome) to change inferences regarding the causal effect estimate. They argue for the use of the partial R^2 scale for the strength of the confounding – as opposed the E-value or parameter estimates as in other approaches – but the overarching concept is similar to the E-value.
2. Complete a worst-case scenario analysis to see if assuming unmeasured confounding accounted for all of the remaining variability in the outcome would change inferences. The concept here is that in some cases, a researcher may be able to demonstrate strong evidence for robustness if the above is true.
3. Help researchers assess whether the amount of confounding needed to alter inferences is plausible in their setting. Cinelli and Hazlett [11] propose to assess this relative to multiples of the strength of an existing strong measured confounder.

While these approaches differ in specifics, the high-level concepts are similar. Start with a broadly applicable summary statistic and graphical approach such as a

contour plot. Apply plausibility techniques, extreme case examination, and draw on external data if available. This conceptual approach should serve as a foundation for standard best practices.

One aspect that has not received enough attention is the importance of quantitative assessment of the potential for unmeasured confounding at the design stage of the study. This was discussed by Girman et al. [29] and tools such as DAGs and the E-value are beneficial at this stage. It is common that researchers think about unmeasured confounding when they assess the appropriateness of the database (if retrospective study) or variables to collect (if prospective study). However, the DAG is a well-established tool that pictorially depicts your assumptions on the relationships between factors – independent of the current data collection or database at hand. This will help the process of identification of potential unmeasured confounders.

Step 2 is the discussion of the expected direction and strength of unmeasured confounders relative to the expected treatment effect. The E-value – based on an expected or minimally clinically relevant treatment effect (as opposed to the observed effect after conducting the analysis) – is a valuable tool at this point. If the E-value is low relative to expected strength of unmeasured confounders, then it is likely that the study will not lead to robust causal inference. This clearly points to whether gathering additional information on unmeasured confounding – either internally in the study design or externally through other studies – is likely to be necessary for actionable RWE from the study. Fang et al. [30] proposed assessing the potential size of bias from unmeasured confounding via the E-value prior to the study. They then demonstrated the value of computing the sample size based on not only the expected effect size but adjusting for potential uncertainty from unmeasured confounding.

For critical decision making from RWE such as regulatory decisions, the step of additional data gathering to ensure minimal impact of unmeasured confounding may be necessary in some settings. In addition, quantitative bias analysis techniques [31] are also valuable to assess concerns in the lack of quality of the data at this point, demonstrating whether robust conclusions from such data are even possible.

It is also important to have carefully planned sensitivity analyses plan described in the study protocol. This demonstrates that thought was put into accounting for as much bias before data collection (e.g. through a DAG and plans for any necessary internal or external data gathering) and assessing the potential for uncontrolled bias after data collection. The plans laid out in both Zhang et al. [28] and in Cinelli and Hazlett [11] are excellent approaches that should serve as the foundation for standard practice moving forward. These begin with broadly applicable quantitative measures that are directed at the core interpretation of the effect (the objective of the study!) and how confident we are that the result should be incorporated into medical decision making. Analysis for this has been made practical through the publishing of R-packages: ‘E-value’, ‘treatSens’, and ‘Sensmaker’. It is less clear regarding best practices when external information is necessary. The Bayesian framework is well suited for incorporating information from multiple sources and could be

extremely beneficial here. More work towards standardizing and implementation tools for those approaches would benefit the field.

In addition to E-value, we believe the simulation based and partial R^2 approaches are potentially useful as initial sensitivity analyses. These methods apply broadly (even if unmeasured confounder not identified), there is no need to transform confounders to a risk ratio scale, and single graphic (contour plot) provides an adjusted treatment effect for an array of possible confounders, not just assuming the impact of unmeasured confounder on treatment and outcome is the same.

At its core, unmeasured confounding is a problem of missing data. There is no better solution for missing data than to gather the data! While overly simplistic, this does emphasize the value of gathering additional data which may be necessary to provide higher level confidence in a RWE finding. When an unmeasured confounder can be identified, a chart review, data linkage, survey, or any other means of obtaining information on the confounder may be possible at least in a subset of data. Using techniques such as multiple imputation or Bayesian regression modeling, this data can provide strong evidence that the full sample findings are robust [14]. In cases where an unmeasured confounder cannot be identified, the challenge of generating convincing additional data is tougher. However, data on other outcomes and controls (empirical calibration [9, 10]) or prior data (prior rate ratio [32]) or instrumental variables [15] are other approaches to providing additional evidence when the standard E-value or RV or Simulation framework are insufficient.

7 Conclusions

No unmeasured confounding is a critical assumption for causal inference. For reliable decision making based on RWE, researchers must address potential bias resulting from the violation of this assumption in order to generate credible RWE. New and broadly applicable quantitative approaches are available along with freely available software for implementation. It is no longer acceptable to just state in the discussion section of a publication that unmeasured confounding is a potential limitation of the work. Prior surveys showing the lack of attention paid to this potential bias are alarming. The field is demanding more and analysts now can efficiently provide guidance on the robustness of any findings.

While we emphasize the availability of broadly applicable methods at this point, it is acknowledged that gaps in research and uncertainty in best practices remain. For instance, research and tools for implementation are needed for the scenarios with longitudinal bias adjustment such as g-methods. Similarly, focusing on best practices for personalized medicine causal inference analyses or the application of external controls are also less developed.

We support recent efforts from researchers to establish structured best practices for quantitatively evaluating the potential impact of unmeasured confounding. Such efforts, combined with better planning at the design stage, will provide meaningful information that will guide consumers of RWE toward better decision making. In

the Disney Mandolorian series, the characters use the phrase 'This is the way' – a saying that refers to agreement that their way of living is the right way. For comparative analyses from non-randomized studies, 'the way' needs to include addressing unmeasured confounding both at the design stage and with quantitative sensitivity analysis during the analysis stage of the research. While research gaps remain and best practices do change, statisticians now have sufficient tools and understanding of assumptions to provide critical insight to help improve appropriate use and decision making from RWE. 'The way' forward – to establishing better quality RWE – includes a transformation where quantitative sensitivity analyses for unmeasured confounding are a regular part of all comparative analysis based on RWE. This will enhance decision making based on RWE and thus improve patient outcomes.

References

1. U.S. Food and Drug Administration Guidance Document (2018). Framework for FDA's Real-World Evidence Program. https://www.fda.gov/media/120060/download.
2. U.S. Food and Drug Administration Guidance Document (2021). Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision Making for Drug and Biologic Products. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-real-world-data-and-real-world-evidence-support-regulatory-decision-making-drug.
3. U.S. Food and Drug Administration Guidance Document (2021). Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biologic Products. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory
4. Neyman, J. On the Application of Probability Theory to Agricultural Experiments: Essay on Principles. Chapter 9, (1923). Translated in *Statistical Science*, 1990: 5(4), 465–472.(1990).
5. Rubin DB. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *J Educ Psychol*. 66(5):688–701 (1974).
6. Rubin DB. Assignment of Treatment Group on the Basis of Covariates. *J Educational Statistics* 2:1–26 (1977).
7. Holland PW. Statistics and Causal Inference. *J American Statistical Association* 81(396):945–960 (1986).
8. Zagar A, Kadziola Z, Lipkovich I, Madigan D, Faries D. Evaluating Bias Control Strategies in Observational Studies Using Frequentist Model Averaging *J Biopharm Stat 27(3):535–553* (2022).
9. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting Observational Studies: Why Empirical Calibration is Needed to Correct p-values. *Statist. Med.,* 33:209–218 (2014).
10. Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, and Hartzema AG. Empirical Assessment of Methods for Risk Identification in Healthcare Data: Results from the Experiments of the Observational Medical Outcomes Partnership. *Statist. Med.,* 31 4401–4415 (2012).
11. Cinelli C and Hazlett C. Making Sense of Sensitivity: Extending Omitted Variable Bias. *J.R. Statst. Soc. B* 82:, Part 1, 39–67 (2020).

12. Blum MR, Tan YJ, Ioannidis JPA. Use of E-values for Addressing Confounding in Observational Studies—an Empirical Assessment of the Literature. *Int J Epidemiol* 49:1482–94 (2020).
13. Zhang X, Faries DE, Boytsov N, Stamey JD, Seaman JWA Jr. Bayesian Sensitivity Analysis to Evaluate the Impact of Unmeasured Confounding with External Data: a Real World Comparative Effectiveness Study in Osteoporosis. *Pharmacoepidemiol Drug Saf*., 25(9):982–992 (2016).
14. Faries D, Peng X, Pawaskar M, Price K, Stamey JD, Seaman JW Jr. Evaluating the Impact of Unmeasured Confounding with Internal Validation Data: An Example Cost Evaluation in Type 2 Diabetes. *Value in Health* 16:259–266 (2013).
15. Federspiel JJ, Anstrom KJ, Xian Y, McCoy LA, Effron MB, Faries DE, Zettler M, Mauri L, Yeh RW, Peterson ED, Wang TY for the Treatment With Adenosine Diphosphate Receptor Inhibitors–Longitudinal Assessment of Treatment Patterns and Events After Acute Coronary Syndrome (TRANSLATE-ACS) Investigators. Comparing Inverse Probability of Treatment Weighting and Instrumental Variable Methods for the Evaluation of Adenosine Diphosphate Receptor Inhibitors After Percutaneous Coronary Intervention. *JAMA Cardiol*., 1(6):655–665 (2016).
16. Choong CK, Belger M, Koch AE, Meyers KJ, Marconi VC, Abedtash H, Faries D, Krishnan V. Comparative Effectiveness of Dexamethasone in Hospitalized COVID-19 Patients in the United States. To appear in *Advances in Therapy* (2022).
17. Uddin MJ, Groenwold RHH, Ali MS, de Boer A, Roes KCB, Chowdhury AB, Klungel OH. Methods to Control for Unmeasured Confounding in Pharmacoepidemiology: an Overview. *International Journal of Clinical Pharmacy* 38(3):1–10 (2016).
18. Streeter AJ, Lin NX, Crathorne L, Haasova M, et al. Adjusting for Unmeasured Confounding in Non-randomised Longitudinal Studies: a Methodological Review. *Journal of Clinical Epidemiology* 87:23–34 (2017).
19. Zhang X, Faries DE, Li H, Stamey JD, Imbens GW. Addressing Unmeasured Confounding in Comparative Observational Research. *Pharmacoepidemiol Drug Saf*. 27:373–382 (2018).
20. Schneeweiss S. Sensitivity Analysis and External Adjustment for Unmeasured Confounders in Epidemiologic Database Studies of Therapeutics. *Pharmacoepidemiology and drug safety* 15(5):291–303 (2006). .
21. VanderWeele TJ and Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine*https://doi.org/10.7326/m16-2607 (2017).
22. VanderWeele TJ, Ding P, Mathur M. Technical Considerations in the Use of the E-Value. *J. Causal Infer.* 2019:1–11 (2019).
23. VanderWeele TJ and Mathur MB. Commentary: Developing Best-practice Guidelines for the Reporting of E-values. *International Journal of Epidemiology*, 49(5): 1495–1497 (2020).
24. Mathur MB, Ding P, Riddell CA, and VanderWeele TJ. Website and R Package for Computing E-Values. *Epidemiology* 29(5): e45–e47 (2018).
25. Dorie V, Harada M, Carnegie NB, Hill J. A Flexible, Interpretable Framework for Assessing Sensitivity to Unmeasured Confounding. *Stat in Med* 35:3453–3470 (2016).
26. Carnegie NB, Harada M, Hill JL. Assessing Sensitivity to Unmeasured Confounding Using a Simulated Potential Confounder. *J. Res. Educational Effectiveness* 9(3):395–420 (2016).
27. Faries D, Zhang X, Kadziola Z, Siebert U, Kuehne F, Obenchain RL, and Haro JM. *Real World Health Care Data Analysis: Causal Methods and Implementation Using SAS®*. Cary, NC: SAS Institute Inc. 2020.
28. Zhang X, Stamey J, Mather MB. Assessing the Impact of Unmeasured Confounders for Credible and Reliable Real-world Evidence. *Pharmacoepi and Drug Safety* 29:1219–1227, 2020.
29. Girman CJ, Faries D, Ryan P, Rotelli M, Belger M, Binkowitz B, O'Neill R, for the Drug Information Association CER Working Group. Pre-Study Feasibility and Identifying Sensitivity Analyses for Protocol Pre-Specification in Comparative Effectiveness Research. *Journal of Comparative Effectiveness* 3(3): 259–270 (2014).

30. Fang Y, He W, Hu X, Wang H. A Method for Sample Size Calculation via E-value in the Planning of Observational Studies. *Pharmaceutical Statistics* 20:163–174 (2021).
31. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good Practices for Quantitative Bias Analyses. *Int J Epidemiol* 43(6):1969–85 (2014).
32. Tannen RL, Weiner MG, Xie D. Use of Primary Care Electronic Medical Record Database in Drug Efficacy Research on Cardiovascular Outcomes: Comparison of Database and Randomised Controlled Trial Findings. *British Medical Journal* 338:b81 (2009).

Sensitivity Analysis in the Analysis of Real-World Data

Yixin Fang and Weili He

1 Introduction

ICH E9(R1) [1] defines estimand as "a precise description of the treatment effect reflecting the clinical question posed by the trial objective. It summarises at a population-level what the outcomes would be in the same patients under different treatment conditions being compared." In chapter "Key Considerations in Forming Research Questions and Conducting Research in Real-World Setting", we discuss key considerations for forming sound research questions in real-world setting and suggest that we can use estimand as the "touchstone" to test whether or not a research question can be answered. For a research question, if an estimand reflecting the question can be defined, then the question is answerable. If not, we may need to enhance the causal identifiability assumptions or revise the PROTECT elements to make the question answerable.

Figure 1 shows the flowchart of forming, revising, and answering a research question in real-world setting. Figure 1 also shows the three sets of assumptions behind the scene. These three sets of assumptions are (1) the set of causal identifiability assumptions behind the causal model, (2) the set of intercurrent events (ICE) assumptions behind the strategies of ICE handling, and (3) the set of statistical assumptions behind the estimation process of the estimand.

As defined in ICH E9(R1), sensitivity analysis is "a series of analyses conducted with the intent to explore the robustness of inferences from the main estimator to deviations from its underlying modeling assumptions and limitations in the data." Figure 1 shows the anatomy of the "underlying modeling assumptions and limitations," where the underlying modeling assumptions are mainly coming from

Y. Fang (✉) · W. He
Data and Statistical Sciences, AbbVie, North Chicago, IL, USA
e-mail: yixin.fang@abbvie.com

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_15

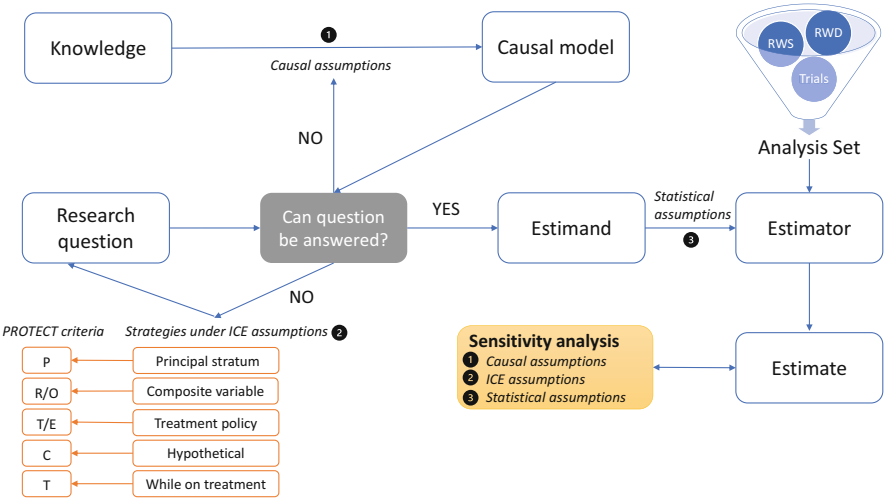

Fig. 1 The flowchart of forming, revising, and answering a research question in real-world setting, and the three sets of assumptions in the defining and estimating an estimand

the three places numbered in the figure and the limitations are mainly coming from the data quality and the external validity between the analysis set and the study population. In this chapter, we discuss how to conduct sensitivity analysis of the robustness of inferences from the main estimator to deviations from those three sets of assumptions. Such sensitivity analysis findings reflect the internal validity of the inferences from the main estimator.

The remaining of the chapter is organized as follows. In Sects. 2–4, we discuss how to conduct sensitivity analysis of the robustness of inferences from the main estimator to deviations from those three sets of assumptions, respectively. In Sect. 5, we conclude with some discussion and an emphasis on the difference between sensitivity analysis and supplemental analysis.

2 Sensitivity Analysis of Identifiability Assumptions

Different answerable research question is in need of different set of identifiability assumptions. As pointed out in ICH E9(R1), "central questions for drug development and licensing are to establish the existence, and to estimate the magnitude, of treatment effects: how the outcome of treatment compares to what would have happened to the same subjects under alternative treatment (i.e., had they not received the treatment, or had they received a different treatment)."

We start with a simple but generic example with the following research question in terms of the average treatment effect (ATE) of treatment $A = 1$ compared to the standard of care (SOC) $A = 0$: how the outcome Y (continuous or binary) at

time T after treatment initiation compares to what would have happened to the same subjects of a defined population under $A = 0$? This research question has all the five PROTECT elements discussed in chapter "Key Considerations in Forming Research Questions and Conducting Research in Real-World Setting": (1) Population of the subjects to whom the treatments would be applied; (2) Response/Outcome is Y; (3) Treatment/Exposure is A, (4) the abstract meaning of "C" element is counterfactual thinking and the tangible meaning is that a covariate vector W will be used to adjust for confounding, and (5) W and A are measured at baseline (treatment initiation) and Y is measured at time T after baseline.

The counterfactual thinking plays an important role in forming the research and defining the estimand of interest [2]. Let Y^a be the potential outcome if the subject is treated by $A = a, a = 0, 1$. For each subject, only one of the two potential outcomes is observed, and the other is counterfactual (i.e., what would have happened if the same subject under alternative treatment). Under the following three identifiability assumptions [3], the consistency assumption,

$$Y = AY^1 + (1 - A)Y^0, \tag{1}$$

the no-unmeasured confounder (NUC) assumption,

$$Y^a \perp\!\!\!\perp A|W, \text{ for } a = 1, 2, \tag{2}$$

and the positivity assumption,

$$P(A = a|W = w) > 0, \text{ for } a = 1, 2, \text{ and } w \text{ with } P_W(w) > 0, \tag{3}$$

where P_W is the probability distribution of W, we can define the following estimand reflecting the research question,

$$E(Y^1 - Y^0) = E\{E(Y|A = 1, W) - E(Y|A = 0, W)\} \triangleq \theta. \tag{4}$$

Assume that we construct an estimator $\widehat{\Theta}$ by some method discussed in chapters "Clinical Studies Leveraging Real-World Data Using Propensity Score-based Methods" and "Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods" of this book, obtain an analysis set $\mathcal{D} = \{(W_i, A_i, Y_i), i = 1, \cdots, n\}$, and produce an estimate $\widehat{\theta} = \widehat{\Theta}(\mathcal{D})$. Now we want to explore the robustness of inferences from $\widehat{\theta}$ to deviations from the three identifiability assumptions.

2.1 Sensitivity Analysis of the Consistency Assumption

The consistency assumption assumes that: if $A = 1$, then $Y = Y^1$; if $A = 0$, then $Y = Y^0$. If the consistency assumption is violated, then for some subjects with $A = 1, Y \neq Y^1$, while for some subjects $A = 0, Y \neq Y^0$. This violation may come

from the discrepancy between the observed A and the true treatment A^* received by some subjects.

In clinical trials, the intervention variable is explicitly defined in a pre-specified protocol, data on the variable are well collected according to the protocol, and any deviations from the protocol are clearly documented. However, in non-interventional real-world setting, the data on the treatment/exposure variable is usually captured from real-world data. Therefore, there are several ambiguities, including potential ambiguities in codes that allow identification of the specific medical products, about whether or not the duration can be ascertained, and about whether or not the data on compliance and concomitant medication. These ambiguities cause the discrepancy between A and A^*.

For example, there are six Pancreatic Enzyme Replacement Therapy (PERT) medications that have been approved by the Food and Drug Administration (FDA) to treat exocrine pancreatic insufficiency (EPI). Therefore, one may define A as taking one given PERT medication (i.e., A is 7-level categorical variables), while the other may define A as taking any of six PERT medications (i.e., A is a binary variable). In another example, claims data are used to capture data on one treatment for hepatitis C virus (HCV) and a pre-specified compliance threshold is used to distinguish $A = 1$ vs. $A = 0$. Therefore, different compliance thresholds (say, at least one week treatment, or at least two weeks treatment) lead to different versions of A.

We can conduct sensitivity analysis to explore the robustness of the inference due to these potential ambiguities in defining A. For the selected definition of A, we obtain estimate $\widehat{\theta}$ along with its statistical inferences (say, 95% confidence interval estimate and p-value). Meanwhile, we provide some alternative definitions of A, denoted as $A^{(1)}, \cdots, A^{(k)}$, producing the corresponding estimates, $\widehat{\theta}^{(1)}, \cdots, \widehat{\theta}^{(k)}$, along with their statistical inferences.

2.2 Sensitivity Analysis of the NUC Assumption

In randomized controlled clinical trials (RCTs), the randomization technique ensures the NUC assumption is automatically satisfied, where W is empty for unstratified RCTs and contains stratification factors for stratified RCTs.

However, in non-randomized real-world setting, we need to check the validity of the NUC assumption. As defined in FDA guidance document [4], a confounder is a "variable that can be used to decrease confounding bias when properly adjusted for in an analysis." We can apply the backdoor criterion [5] to identify a confounder vector W such that the NUC assumption is satisfied. In a directed acyclic graph (DAG), there may be several versions of W such that the NUC assumption is satisfied. We attempt to select a version of W such that all the confounders in W are measured.

If every version of W includes some unmeasured confounder(s), there are two ways to proceed. One way is to look for other data sources to collect some of these unmeasured confounders such that we are able to identify a version of W

consisting of only measured confounders. Chapter "Key Variables Ascertainment and Validation in RW Setting" of this book discusses the key variables ascertainment and validation in real-world setting, including ascertainment of covariates (confounders and effect modifiers). The other way is to select one version of W, exclude those unmeasured confounders from W, and then conduct sensitivity analysis to explore the robustness of the inference if there is unmeasured confounding. Chapter "Sensitivity Analyses for Unmeasured Confounding: This Is the Way" of this book focuses on conducting sensitivity analysis for unmeasured confounding.

For the purpose of completeness, we briefly review one recent method for conducting sensitivity analysis of unmeasured confounding, the E-value [6]. Let $\widehat{\theta}$ be the estimate of the estimand θ under the NUC assumption, U be an unmeasured confounder, and $\widehat{\theta}(U)$ be the estimate with the presence of U. As the association of U with both A and Y increases, the estimated treatment effect $\widehat{\theta}(U)$ becomes weaker. The E-value is the minimum strength of association that an unmeasured confounder U would need to have with both the exposure A and the outcome Y, conditional on the measured confounders W, to fully explain away a specific exposure-outcome association $\widehat{\theta}(U)$. The bigger E-value the more robust of the inference based on $\widehat{\theta}$. Formulas of E-value for binary, continuous, and time-to-event outcomes have been developed along with R package "EValue" [6].

2.3 *Sensitivity Analysis of the Positivity Assumption*

The positivity assumption requires sufficient variability in treatment or exposure assignment within strata of confounders. Positivity violations can arise for two reasons [7]. First, it may be impossible in the population level for subjects with certain covariate values to receive a given exposure of interest. For example, the investigative treatment $A = 1$ is only available to female and the SOC $A = 0$ is available to both male and female, leading to $P(A = 1|\text{male}) = 0$, which violates the positivity assumption. If the positivity violation is due to this reason, we may redefine the estimand by restricting the population to some subpopulation or by restricting the treatment levels that do not result in positivity violation. This is one topic that has been discussed in chapter "Key Considerations in Forming Research Questions and Conducting Research in Real-World Setting", how to make a research question answerable.

Second, violations or near violations of the positivity assumption can arise in finite samples due to chance, in particular, in the settings where the sample size is small or the set of covariates is large. In this section, we discuss how to conduct sensitivity analysis to explore the robustness of the inference when there is violation or near violation due to this reason. As in [7], here we use the term "sparsity" to refer to positivity violations or near violations.

Continue the example with data $\mathcal{D} = \{(W_i, A_i, Y_i), i = 1, \cdots, n\}$. Let

$$g(a|w) = P(A = a|W = w), \tag{5}$$

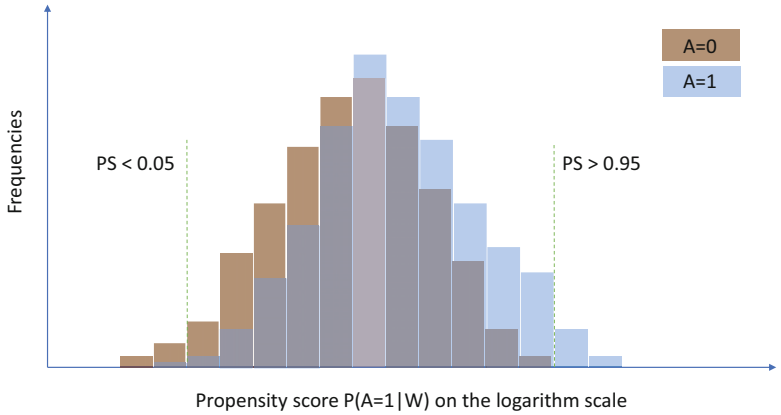

Fig. 2 Positivity violations happen in the left-most bar and in the two bars on the right-most, while one near violation happens on the second left-most bar

where $a = 1, 0$, be the propensity score function, which is estimated by $\widehat{g}(a|w)$ based on the data. Let

$$Q(a, w) = E(Y|A = a, W = w), \tag{6}$$

be the regression function for either binary or continuous outcome, which is estimated by $\widehat{Q}(a, w)$ based on the data.

Figure 2 shows an example of sparsity. Let $\mathcal{S} = \{(W_{si}, A_{si}, Y_{si}), i = 1, \cdots, m\}$ consist of those of subjects with propensity scores at two extreme ends, say, with $\widehat{g}(1|W_i) > 1 - 0.05$ or $\widehat{g}(1|W_i) < 0.05$, where 0.05 is some pre-specified threshold to be used to truncate the propensity scores. Since the robustness to the positivity violations is estimator-specific, let us discuss the sparsity-triggered behaviors of two basic classes estimators reviewed in [8]: one class of estimators constructed using the standardization approach and the other class of estimators constructed using the weighting approach.

One estimator constructed by the standardization approach is maximum likelihood estimator (MLE),

$$\widehat{\theta}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \widehat{Q}(1, W_i) - \widehat{Q}(0, W_i) \right\}. \tag{7}$$

For each subject in $\mathcal{S}$, one of the two estimates in the summand of the above formula, $\widehat{Q}(1, W_i)$ and $\widehat{Q}(0, W_i)$, is not supported by data due to sparsity and therefore requires extrapolation, resulting in potential bias in $\widehat{\theta}_{MLE}$.

One estimator constructed by the weighting approach is inverse probability of treatment-weighted estimator (IPTW),

$$\widehat{\theta}_{IPTW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_i}{\widehat{g}(1|W_i)}Y_i - \frac{1-A_i}{\widehat{g}(0|W_i)}Y_i\right\}. \tag{8}$$

For each subject in $\mathcal{S}$, one of the two weights in the summand of the above formula, $1/\widehat{g}(0|W_i)$ and $1/\widehat{g}(1|W_i)$, is very large due to sparsity, resulting in large variance of $\widehat{\theta}_{IPTW}$. Using some threshold, such as 0.05 displayed in Fig. 2, to truncate the denominators of the weights can reduce the variance but can increase the bias in $\widehat{\theta}_{IPTW}$.

Doubly robust estimators including Augmented IPTW (AIPTW, [3]) and Targeted MLE (TMLE, [7]) combine the strength of the above two approaches. Doubly robust estimators are consistent if either (1) $\widehat{g}$ is a consistent estimator of g and g satisfies positivity or (2) $\widehat{Q}$ is a consistent estimator of Q and a truncated version of $\widehat{g}$ converges to a distribution g^* that satisfies positivity. However, the potential bias due to sparsity still remains due to the implicit extrapolation in $\widehat{Q}$ if some propensity truncation is applied.

After we think through the source of potential bias due to sparsity, we can add some disturbance to explore the robustness. If the bigger value in continuous Y means better performance or 1 in binary Y means success, then a positive estimate $\widehat{\theta}$ indicates a favorable treatment effect. Hence, we subtract a positive disturbance δ from $\widehat{Q}(1, W_i) - \widehat{Q}(0, W_i)$ for each subject in $\mathcal{S}$ to manifest the unfavorable extrapolation bias, leading to a disturbed estimate

$$\widehat{\theta}(\delta) = \widehat{\theta} - m\delta/n. \tag{9}$$

We may solve δ_0 such that $\widehat{\theta}(\delta_0) = 0$, which is similar to E-value or the tipping-point method in the literature of missing data. The bigger δ_0 the more robust to sparsity. If δ_0 is unrealistically large, we may conclude the finding is trustful even under positivity violations.

3 Sensitivity Analysis of ICE Assumptions

We refer to the assumptions behind the ICH E9(R1) strategies for ICE handling as the ICE assumptions. The literature on sensitivity analysis in missing data handling is rich, but the literature on sensitivity analysis in ICE handling is lacking. The ICH E9(R1) covers both estimand and sensitivity analysis. Although the ICH E9(R1) provides the definition and emphasizes the importance of sensitivity analysis, it does not explicitly address how to conduct sensitivity analysis to the ways of ICE handling. As far as we know, Chapter 16 of book "*Estimands, Estimators, and Sensitivity Analysis in Clinical Trials*" [9] is the only literature on how to conduct sensitivity analysis to the ways of ICE handling. In this section, we follow their thoughts to discuss the ICE assumptions and how to conduct sensitivity analysis of the ICE assumptions.

We start with a brief review of two basic classes of methods for sensitivity analysis in missing data handling, which we can adopt for sensitivity analysis in ICE handling. The first class includes the delta adjustment, along with the tipping-point method [10]. The second class includes pattern-mixture modeling, in particular, the reference-based imputation methods [10]. In practice, we often consider the analysis under the missing at random (MAR) assumption as the main analysis, and then select a sensitivity analysis method to explore the robustness of the inference when the MAR assumption is violated, that is, under the missing not at random (MNAR) assumption.

The multiple imputation procedure (MI) [11] is often used for analysis under the MAR assumption. Consider traditional MI using a regression on residuals. Each subject's deviation from their respective arm (e.g., the mean of their respective arm) prior to the missing data occurrence is used to impute the missing residuals. Under MAR, imputed values are computed based on the imputed residuals plus the mean of the arm to which the subject is initially in. Both the delta adjustment and the reference-based imputation can be implemented by SAS procedure "PROC MI."

In delta adjustment, a sensitivity parameter called delta (δ) is added to the imputation model, with $\delta = 0$ associated with the MAR assumption. We increase δ from zero across a range of values that progressively deviates further from the MAR assumption to explore the robustness of the inference. If the treatment effect remains significant across plausible deviations from the MAR assumption, we can conclude that the inference from the primary analysis is robust. This comes with the tipping-point method, which determines the tipping-point δ_0 when the inference turns into non-significance from significance. The bigger δ_0 means the more robust inference based on the primary analysis. The use of delta adjustment is not limited to assessing deviations from the MAR assumption. It can also be used in combination with MI to assess departures from any specific assumption, including any certain MNAR assumption or any certain ICE assumption.

Reference-based imputation is a specific version of pattern-mixture modeling, in which the control or reference arm (the arm with $A = 0$, the SOC arm, the external control arm, etc.) is used to envision some hypothetical scenarios or patterns under the MNAR assumption. Two variants of reference-based imputation are jump-to-reference (J2R) imputation and copy-reference (CR) imputation. In J2R imputation, the imputed values for patients in the investigative treatment arm (the arm with $A = 1$) takes on the attributes of the reference arm immediately after missing data occur. In CR imputation, the investigative treatment effect that has been obtained up to the time when missing data occur gradually diminishes after missing data occur, in accordance with the correlation structure implied by the imputation model. Recently, a wide class of imputation models including CR and J2R as two extreme ends has been proposed [12].

Let Y_t be outcome variable measured at visit t, where $t = 1, \cdots, T$, and $t = 0$ be the baseline. Assume that the outcome at the final visit T, Y_T, is defined as the primary endpoint. Define $\overline{Y}_t = (Y_1, \cdots, Y_t)'$, the vector of all outcomes up to visit t, $t = 1, \cdots, T$, and $\overline{Y} = (Y_1, \cdots, Y_T)'$, the vector of the outcomes at all visits. Let L be the random variable denoting the last visit prior to treatment

Table 1 Some examples of treatment sequences and ICEs

Sequence	Explanation
$\overline{1} = \text{rep}(1, T)$	Initially treated by 1 and through
$\overline{0} = \text{rep}(0, T)$	Initially treated by 0 and through
$\overline{1}_{L,0} = (\text{rep}(1, L+1), \text{rep}(0, T-L-1))$	Initially 1, change to 0 after L
$\overline{1}_{L,\texttt{NA}} = (\text{rep}(1, L+1), \text{rep}(\texttt{NA}, T-L-1))$	Initially 1, change to NA after L
$\overline{0}_{L,\texttt{NA}} = (\text{rep}(0, L+1), \text{rep}(\texttt{NA}, T-L-1))$	Initially 0, change to NA after L
$\overline{1}_{L,2} = (\text{rep}(1, L+1), \text{rep}(2, T-L-1))$	Initially 1, change to 2 after L
$\overline{0}_{L,2} = (\text{rep}(0, L+1), \text{rep}(2, T-L-1))$	Initially 0, change to 2 after L

1 stands for the investigative treatment
0 stands for the reference treatment
NA stands for no treatment
2 stands for other treatment such as rescue medication

discontinuing (e.g., withdrawal, loss of follow up) or treatment changing (e.g., change to rescue treatment, add an alternative treatment). Let $\overline{A} = (A_0, \cdots, A_{T-1})$ be the actually received treatment sequence, and $\overline{A}_t = (A_0, \cdots, A_t)$ be the treatment up to t, $t = 0, \cdots, T-1$. Let W_0 be baseline covariates, and let $\overline{W}_t = (W_0, \cdots, W_t)$ be the vector consisting of all the observed history up to time t including baseline covariates, time-dependent covariates, and intermediate outcomes, $t = 0, \cdots, T-1$.

Let $\overline{a} = (a_0, \cdots, a_{T-1})$ be a given treatment sequence and $\text{rep}(a, p)$ be a p-dim vector of all a's. Let $Y_T^{\overline{a}}$ be the potential outcome at time T for any given treatment sequence $\overline{a}$. Some examples of treatment sequence $\overline{a}$ are in Table 1. In $\overline{1}$ and $\overline{0}$, there is no ICE. In all the other sequences in the table, there is one ICE occurring between visit $L+1$ and L after visit L.

3.1 Sensitivity Analysis for the Hypothetical Strategy

According to ICH E9(R1), if the hypothetical strategy is applied to handle ICEs, "a scenario is envisaged in which the intercurrent event would not occur." In this hypothetical scenario in which the ICE would not occur, if $A_0 = 1$, then $\overline{A} = \overline{1}$, and if $A_0 = 0$, then $\overline{A} = \overline{0}$. Therefore, the average treat effect comparing two treatment sequences, $\overline{a} = \overline{1}$ and $\overline{a} = \overline{0}$, is expressed in terms of potential outcomes as $E\{Y_T^{\overline{a}=\overline{1}}\} - E\{Y_T^{\overline{a}=\overline{0}}\}$. Under the consistency assumption, the positivity assumption, and the following static sequential exchangeability assumption [3]:

$$Y_T^{\overline{a}} \perp\!\!\!\perp A_0 | \overline{W}_0, \text{ for any } \overline{a}, \tag{10}$$

$$Y_T^{\overline{a}} \perp\!\!\!\perp A_t | \overline{A}_{t-1}, \overline{W}_t, \text{ for any } \overline{a} \text{ and } t = 1, \cdots, T-1, \tag{11}$$

we can show that via the g-formula [3],

$$E\{Y_T^{\overline{a}}\} = \sum_{\overline{w}} \Bigg[E(Y_T|\overline{W} = \overline{w}, \overline{A} = \overline{a}) \times \tag{12}$$
$$\Bigg\{ \prod_{t=1}^{T-1} P\Big(W_t = w_t \Big| \overline{W}_{t-1} = \overline{w}_{t-1}, \overline{A}_t = \overline{a}_t)\Big)\Bigg\} P(W_0 = w_0)\Bigg],$$

where the summation is over the sample space of $\overline{W}$. Here for demonstration purpose we assume $\overline{W}$ is discrete, and for general case we can replace summation by integration and probability by density. Hence, under the above three assumptions, $E\{Y_T^{\overline{a}=\overline{1}}\} - E\{Y_T^{\overline{a}=\overline{0}}\}$ is equal to the following estimand,

$$\theta_h = \sum_{\overline{w}} \left[E(Y_T|\overline{W} = \overline{w}, \overline{A} = \overline{1}) \times \prod_{t=1}^{T-1} P(w_t|\overline{w}_{t-1}, \overline{1}_t)) P(W_0 = w_0) \right] -$$
$$\sum_{\overline{w}} \left[E(Y_T|\overline{W} = \overline{w}, \overline{A} = \overline{0}) \times \prod_{t=1}^{T-1} P(w_t|\overline{w}_{t-1}, \overline{0}_t)) P(W_0 = w_0) \right]. \tag{13}$$

Assume that we construct an estimator $\widehat{\theta}_h$ to estimate θ_h. Note the estimand (13) does not depend on data after the occurrence of an ICE, regardless of being collected or not. Therefore, if the hypothetical strategy is applied to handle the ICE, all the data after the occurrence of an ICE are considered as "missing data." Hence, the sensitivity analysis for the hypothetical strategy becomes the sensitivity analysis for missing data handling. This can also be seen by breaking the static sequential exchangeability assumption into two parts, the first part (10) at $t = 0$ that is the same as the NUC assumption, and the second part (11) at $t > 0$ that is the same as the MAR assumption, equalizing the occurrence of ICE to the occurrence of "missing data." The sensitivity analyses of the consistency, positivity, and NUC assumptions are discussed in Sect. 2. And we can adopt the missing data sensitivity methods for conducting sensitivity analysis of the MAR assumption (11).

Both the delta adjustment and reference-based imputation can be applied in combination with the multiple imputation procedure. Let $\mathcal{D}_h$ be the data excluding all the data after ICE occurrence even if they are collected. For each given delta value in the adjustment or each given scenario in the reference-based imputation (e.g., CR, J2R), let $\mathcal{D}_h^{(k)}$, $k = 1, \cdots, I$, be multiple versions of completed dataset with "missing data" imputed by the multiple imputation procedure. For each version $\mathcal{D}_h^{(k)}$, we use some method (e.g., MLE, IPTW, AIPTW, TMLE) to construct an estimate $\widehat{\theta}_h^{(k)}$ of θ. Then we apply Rubin's rule to combine I versions of estimate and produce an estimate $\widehat{\theta}_h^{(\cdot)}$. When we vary the delta values or the reference-based scenarios, we obtain a series of $\widehat{\theta}_h^{(\cdot)}$ estimates to explore the robustness of the inference based on the primary estimate $\widehat{\theta}_h$.

3.2 Sensitivity Analysis for the Treatment Policy Strategy

According to ICH E9(R1), if the treatment policy strategy is applied, "the intercurrent event is considered to be part of the treatments being compared." Therefore, by the treatment policy strategy, the treatment variable of interest becomes a set of treatment regimes that include more than just the initial treatment. Table 1 shows some examples of treatment regimes, where $\overline{1}$ and $\overline{0}$ stand for regimes of adhering to the initially received, $\overline{1}_{L,0}$ stands for regime of switching from the investigative treatment to the reference treatment (e.g., the SOC), $\overline{1}_{L,\mathtt{NA}}$ and $\overline{0}_{L,\mathtt{NA}}$ stand for regimes of treatment discontinuation to no treatment, and $\overline{1}_{L,2}$ and $\overline{0}_{L,2}$ stand for regimes of treatment changing to an alternative treatment (e.g., rescue medication or added-on treatment).

To apply the treatment policy strategy, we first identify a set of treatment regimes of interest through clinical judgement, or through empirical evidence on the treatment patterns, such that they are relevant to the clinical practice in real-world setting, and the data after the ICE occurrence are still collected with high percentage. For example, if data after rescue medication and switching to the SOC are collected and the SOC and rescue medication are relevant to the clinical practice real-world setting, we can consider the comparison between the investigative treatment policy $\mathcal{A}_1 = \{\overline{1}, \overline{1}_{L,0}, \overline{1}_{L,2}\}$ and the reference treatment policy $\mathcal{A}_0 = \{\overline{0}, \overline{0}_{L,2}\}$.

The treatment policy strategy is often applied along with the hypothetical strategy, because usually there are certain treatment regimes that are not relevant to the clinical practice and there are missing data due to loss of follow up. For the ICEs that are not included in the treatment policies $\mathcal{A}_1$ and $\mathcal{A}_0$ and the ICEs that lead to missing data, we apply the hypothetical strategy to handle them, which means we need to envisage a hypothetical scenario in which these other intercurrent events would not occur.

First, if all types of ICEs are incorporated in either the investigative treatment policy $\mathcal{A}_1$ or the reference treatment policy $\mathcal{A}_0$ and the data after ICE occurrences are collected, then by the treatment policy strategy, under the consistency, positivity, and NUC assumption, the estimand of interest is

$$\begin{aligned}\theta_{tp} &= E\left\{E(Y_T|\overline{A} \in \mathcal{A}_1, W_0) - E(Y_T|\overline{A} \in \mathcal{A}_0, W_0)\right\} \\ &= E\left\{E(Y_T|A_0 = 1, W_0) - E(Y_T|A_0 = 0, W_0)\right\}. \end{aligned} \tag{14}$$

Assume that we construct an estimator $\widehat{\theta}_{tp}$ to estimate θ_{tp}. Therefore, we can conduct similar sensitivity analysis discussed in Sect. 2 to explore the robustness of the inference based on $\widehat{\theta}_{tp}$ to the deviations from the consistency assumption, the positivity assumption, or the NUC assumption.

Second, if, besides those types of ICEs that are incorporated in policy $\mathcal{A}_1$ or $\mathcal{A}_0$, there are other types of ICEs and/or missing data due to loss of follow up, then we may apply the hypothetical strategy to handle these other ICEs and missing data, in combination with the treatment policy strategy. By the hypothetical strategy, all the

data collected after the occurrence of these other ICEs are treated as extra "missing data." Therefore, besides the sensitivity analysis for the treatment policy strategy, we should conduct additional sensitivity analysis to explore the robustness of the inference to the deviations from the MAR assumption (11), using methods such as the delta adjustment and the reference-based imputation.

3.3 Sensitivity Analysis for the Composite Variable Strategy

According to ICH E9(R1), if the composite variable strategy is applied, "an intercurrent event is considered in itself to be informative about the patient's outcome and is therefore incorporated into the definition of the variable." That means, we need to revise the definition of response or outcome variable from Y_T to Y_T^* to include the intercurrent event as part of it. For example, if the outcome variable is already success or failure, discontinuation of treatment or rescue medication would simply be considered another mode of failure. If the outcome variable is an order categorical variable, we may assign the least favorable category as the outcome for the subject with an ICE. If the outcome variable is a continuous variable, we may assign the least favorable value as the outcome to the subject with an ICE. Assume the new outcome variable Y_T^* is a composite variable of Y_T and ICEs. Then the estimand reflecting the treatment effect can be defined as

$$\theta_{cv} = E\left\{E(Y_T^*|A_0 = 1, W_0) - E(Y_T^*|A_0 = 0, W_0)\right\}. \tag{15}$$

In real-world setting, we should distinguish among "good" ICEs, "bad" ICEs, and "neutral" ICEs. For example, if treatment discontinuation is an ICE, there may be treatment discontinuation due to early response ("good" ICE), due to lack of effectiveness ("bad" ICE), or due to insurance change ("neutral" ICE). If we are able to distinguish these types of ICEs, we may consider "good" ICE as success, "bad" ICE as failure, and "neutral" ICE as censoring.

Assume that we construct an estimator $\widehat{\theta}_{cp}$ to estimate θ_{tp}. We need to conduct sensitivity analysis for the identifiability assumptions discussed in Sect. 2, along with sensitivity analysis for the validity of the composite outcome Y_T^*, exploring the robustness of the inference to different ways of defining Y_T^*.

If we assign all the ICEs as another mode of failure in Y_T^* for the primary analysis, in sensitivity analysis we may only assign some ICEs (e.g., rescue medication, discontinuation due to lack of effectiveness) as another mode of failure, but assign with some rate less than one for other ICEs (e.g., loss of follow up, treatment switch). Consequently, we may adjust the rate across a range of values to explore the robustness of the inference. If the outcome variable is continuous, we may adjust the assigned least favorable value across a range of values to explore the robustness of the inference.

3.4 Sensitivity Analysis for the While on Treatment Strategy

According to ICH E9(R1), if the while on treatment strategy is applied, "response to treatment prior to the occurrence of the intercurrent event is of interest." Therefore, the estimator is constructed using data at $t = 1, \cdots, L$, ignore the data at $t = L + 1, \cdots, T$ although they may be collected. An underlying assumption behind this strategy is that the treatment effect is following some fixed trend. Here are some examples of treatment trend over time: (i) the treatment effect is only temporary (e.g., treating symptoms), (ii) the effect of one-time treatment is permanent (e.g., surgery), (iii) the treatment effect is cumulative with a constant rate. Different assumption on the treatment trend leads to different way to define an estimand of interest and construct an estimator to estimate it, along with corresponding sensitivity analysis. For (iii), we should define the treatment duration carefully and, if necessary, conduct sensitivity analysis for the consistency assumption as discussed in Sect. 2.1.

Consider Assumption (i), which assumes that the treatment effect is only temporary. We can consider the rate of binary outcome per unit time or the average of continuous outcome as the new primary endpoint, that is, $Y^* = \sum_{t=1}^{\min(T,L)} Y_t / \min(T, L)$, where L is the last visit prior to the ICE occurrence and is equal to infinity if there is no ICE. Then the estimand reflecting the treatment effect can be defined as

$$\theta_{wot} = E\left\{E(Y^*|A_0 = 1, W_0) - E(Y^*|A_0 = 0, W_0)\right\}. \tag{16}$$

Assume that we construct an estimator $\widehat{\theta}_{wot}$ to estimate θ_{wot}. We need to conduct sensitivity analysis for the identifiability assumptions discussed in Sect. 2, along with sensitivity analysis for the validity of Assumption (i). First, we treat all the data after the ICE occurrence as "missing data." Then we apply the delta adjustment method in combination with multiple imputation to add a delta to impute the "missing data." For each delta, multiple complete datasets (in which $\min(T, L) = T$ for each subject) are generated, we obtain an estimate of $\widehat{\theta}_{wot}$ using the Rubin's rule. After we repeat this process for a range of delta values, we explore the robustness of the inference.

Consider Assumption (ii), which assumes that the effect of an one-time treatment is permanent. We can consider the outcome at the last visit prior to the ICE occurrence as the new primary endpoint, that is, $Y^* = Y_{\min(T,L)}$. Then the estimand reflecting the treatment effect can be defined similarly as in (16), using the newly defined Y^*. Assume that we construct an estimator $\widehat{\theta}_{wot}$ to estimate θ_{wot}. We can apply the same sensitivity analysis method described in the previous paragraph, except that in this case Y^* is the outcome at the last visit instead of the average as in the previous case.

Consider Assumption (iii), which assumes that the treatment effect is cumulative with a constant rate. We can consider the rate of change in the outcome variable over time as the new primary endpoint; e.g., $Y^* = [Y_{\min(T,L)} - Y_0]/\min(T, L)$,

or the average of rates of change over $\min(T, L)$ time intervals, or the subject-specific slope estimated from the subject-wise linear regression or the mixed effect model using the whole dataset. Then the estimand reflecting the treatment effect can be defined similarly as in (16), using the newly defined Y^*. Assume that we construct an estimator $\widehat{\theta}_{wot}$ to estimate θ_{wot}. We can apply the same sensitivity analysis method described in the previous two paragraphs, except that in this case Y^* is the rate of change.

3.5 Sensitivity Analysis for the Principal Stratum Strategy

According to ICH E9(R1), if the principal stratum strategy is applied, "the target population might be taken to be the principal stratum in which an intercurrent event would occur. Alternatively, the target population might be taken to be the principal stratum in which an intercurrent event would not occur." The principal stratification approaches rely on covariates to predict to which stratum each subject belongs when in real-world setting the strata are not observable.

Define two potential outcomes for ICE occurrence, $C^{a=1}$ and $C^{a=0}$, where $C^{a=1} = 1$ is the indicator that an ICE would occur if the subject was treated by $a = 1$ and $C^{a=0} = 1$ is the indicator that an ICE would occur if the subject was treated by $a = 0$. Therefore, the principal stratum in which an ICE would occur (PS_1) and the principal stratum in which an ICE would not occur (PS_2) can be defined as

$$PS_1 = \{C^{a=1} = 1, C^{a=0} = 1\},$$
$$PS_2 = \{C^{a=1} = 0, C^{a=0} = 0\}.$$

Using the data, we can train a ICE predictive model $\widehat{C}(w, a)$, based on which we can predict the corresponding principal strata,

$$\widehat{PS}_1 = \{1 \le i \le n | \widehat{C}(W_{0,i}, 1) = 1, \widehat{C}(W_{0,i}, 0) = 1\},$$
$$\widehat{PS}_2 = \{1 \le i \le n | \widehat{C}(W_{0,i}, 1) = 0, \widehat{C}(W_{0,i}, 0) = 0\},$$

where $W_{0,i}$ is the baseline covariate vector for subject i. For a given principal stratum that is identified as the target population, assume that an estimand reflecting the treatment effect is defined, and an estimator is constructed aligned with the estimand.

Therefore, an sensitivity analysis for the principal stratum strategy is to explore the robustness of the inference to deviations between the predicted stratum memberships and the true stratum memberships. For this aim, we may consider a series of predictive models, such as logistic regression, logistic regression with LASSO, random forest, boosting, support vector machine, and other predictive models in the statistical learning literature [13]. If the estimates based on the predicted principal

strata using different predictive models are similar, then we may conclude that the finding is robust.

4 Sensitivity Analysis of Statistical Assumptions

Under the identifiability assumptions and the ICE assumptions, an estimand θ reflecting the research question is defined. Then the causal inference problem becomes a statistical inference problem. Under some statistical assumptions, we construct an estimator $\widehat{\theta}$. This section is focused on sensitivity analysis of statistical assumptions.

Although the identifiability and ICE assumptions are untestable, statistical assumptions are often testable. For example, if the analysis of variance (ANOVA) depends on the normality assumption, we can conduct a test (e.g., Kolmogorov–Smirnov test) to test the normality, and then we should conduct sensitivity analysis to explore the robustness of the results if the normality assumption is violated. On the other hand, we can consider a non-parametric version of ANOVA, say Kruskal–Wallis one-way analysis of variance, without making the normality assumption, avoiding the need of conduct sensitivity analysis of the normality assumption.

Continue the example in Sect. 2 with data $\mathcal{D} = \{(W, A, Y)\}$. If we apply linear regression analysis to estimate the regression function $Q(a, w)$ and logistic regression analysis to estimate the propensity score $g(a|w)$, then behind the scene we assume that the regression function $Q(a, w)$ is linear and the logarithm of the propensity odds $\log\{g(1|w)/g(0|w)\}$ is also linear. Hence we should conduct sensitivity analysis to explore the robustness of the results if the linearity assumptions are violated. On the other hand, if we apply some non-parametric method (say, super learner [14]), then there is no need to conduct sensitivity analysis to explore the robustness of the results for the linearity assumptions.

Semi-parametric theory plays a critical role in reducing statistical assumptions [15]. The estimand θ defined in (4) can be written as

$$\theta = \int [Q(1, w) - Q(0, w)] d P_W(w), \tag{17}$$

which is a function of regression $Q(a, w)$ and marginal distribution of W, $P_W(w)$. If we make any parametric assumptions on $Q(a, w)$ and $P_W(w)$, we need to conduct sensitivity analysis for them. Therefore, in order to lighten the burden of sensitivity analysis, we do not want to make any parametric assumptions on $Q(a, w)$ and $P_W(w)$. Then the estimation problem becomes a semi-parametric problem, with θ being a parameter of interest and $Q(a, w)$ and $P_W(w)$ being two non-parametric functions. The targeted learning framework reviewed in chapter "Causal Inference with Targeted Learning for Producing and Evaluating Real-World Evidence" of this book has been developed based upon the semi-parametric theory to construct an efficient and semi-parametric estimator of estimand θ.

5 Discussion

In order to make a real-world research question answerable, we often make a set of identifiability assumptions and a set of ICE assumptions. An estimand is defined to reflect the answerable question. In order to estimate the estimand, we sometimes make a set of statistical assumptions. This chapter discusses how to conduct sensitivity analysis to explore the robustness of the inference based on the primary analysis to the violations of these assumptions.

Here is some discussion on the novelty of the contents of this chapter. The literature on the sensitivity analysis of the NUC assumption is rich. The literature on the sensitivity analysis of the consistency and positivity assumptions is lacking. Chapter 9 of [7] discusses the role of the positivity assumption and proposes a method to detect the potential bias due to the positivity violation, but without discussion on sensitivity analysis. The sensitivity analysis of positivity violation discussed in Sect. 2 is novel. The sensitivity analysis methods of the ICE assumptions discussed in Sect. 3 are motivated by Chapter 16 of [9], but Sect. 3 contains much more practical details. The discussion in Sect. 4 on the role of the semi-parametric theory in lightening the burden of sensitivity analysis for statistical assumptions is also novel.

We conclude this chapter by emphasizing the difference between sensitivity analysis and supplementary analysis, which is defined in ICH E9(R1) as "a general description for analyses that are conducted in addition to the main and sensitivity analysis with the intent to provide additional insights into the understanding of the treatment effect." Assume that there is a primary estimand that is aligned with the research question and objective. Also assume that the main analysis and sensitivity analysis associated with the main analysis are pre-specified for the primary estimand. Then the analysis other than the main analysis (e.g., using a different statistical method, under a different set of assumptions, using a different dataset) is considered as supplemental analysis. In addition, if there are other estimands defined differently but for the same research question and objective, all the analyses corresponding to these other estimands are also considered as supplemental analysis. Therefore, if a series of estimands are defined using different ICE strategies, except that the analysis for the primary estimand is the main analysis, the analyses associated with the other estimands are supplement analysis rather than sensitivity analysis.

References

1. FDA Guidance Document (2021): E9(R1) Statistical Principles for Clinical Trials; Addendum: Estimand and Sensitivity Analysis in Clinical Trials, https://www.fda.gov/media/148473/download
2. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic books, New York (2018)

3. Hernan, M., Robins, J: Causal Inference: What If. Boca Raton: Chapman & Hall/CRC (2020)
4. FDA Guidance Document (2021): Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products, https://www.fda.gov/media/152503/download
5. Pearl, J.: Causality. Cambridge university press (2009)
6. VanderWeele, T., Ding, P.: Sensitivity analysis in observational research: introducing the E-value. *Annals Of Internal Medicine*. **167**, 268–274 (2017)
7. van der Laan, M., Rose, S.: Targeted Learning: Causal Inference for Observational and Experimental Data. Springer (2011)
8. Fang, Y.: Two basic statistical strategies of conducting causal inference in real-world studies. *Contemporary Clinical Trials*. **99** pp. 106193 (2020)
9. Mallinckrodt, C., Molenberghs, G., Lipkovich, I., Ratitch, B.: Estimands, Estimators and Sensitivity Analysis in Clinical Trials. CRC Press (2020)
10. O'Kelly, M., Ratitch, B.: Clinical Trials with Missing Data: A Guide for Practitioners. John Wiley & Sons (2014)
11. Rubin, D.: Multiple imputation after 18+ years. *Journal Of The American Statistical Association*. **91**, 473–489 (1996)
12. Fang, Y., Jin, M.: Sequential modeling for a class of reference-based imputation methods in clinical trials with quantitative or binary outcomes. *Statistics In Medicine, Accepted*. **41**, 1525–1540 (2022)
13. Hastie, T., Tibshirani, R., Friedman, J., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer (2009)
14. van der Laan, M., Polley, Hubbard, A.: Super learner. *Statistical Applications In Genetics And Molecular Biology*. **6** (2007)
15. Bickel, P., Klaassen, C., Ritov, Y., Wellner, J.: Efficient and Adaptive Estimation for Semiparametric models. Springer (1993)

Personalized Medicine with Advanced Analytics

Hongwei Wang, Dai Feng, and Yingyi Liu

1 Background

1.1 What Is Personalized Medicine

Personalized medicine has been an integral part of modern time medicine. From treatment choice at daily patient–physician interaction in routine clinical practice to major regulatory and reimbursement decision by health authority, from bench work at basic research lab to large-scale real-world study quantifying benefit-risk profiles of different interventions, PM supported by evidence plays a central role in drug development, improving patients' quality of life and increasing the productivity of healthcare system. There is no universal definition of PM, and the Horizon 2020 Advisory Group of the EU defines personalized medicine as "a medical model using characterization of individuals' phenotypes and genotypes (e.g., molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention" [1].

An example of PM is the biomarker-driven treatment choice in breast cancer. Three molecular biomarkers, namely estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), provide prognostic information and guide the choice of therapies [2]. This is enabled by breakthrough in sciences which shed light on the mechanism of disease. Drug development in this area centers around the patient segmentation defined by these biomarkers, and

H. Wang (✉) · D. Feng · Y. Liu
Medical Affairs and Health Technology Assessment Statistics, Data and Statistical Sciences, AbbVie, North Chicago, IL, USA
e-mail: hongwei.wang@abbvie.com

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_16

biomarker screening has also been part of treatment guidelines and routine clinical practice.

Another example of PM is the more recent introduction of CAR-T (chimeric antigen receptors T-cell) therapies for the management of specific blood cancers. This is a highly individualized intervention where a sample of a patient's T cells are collected and specifically modified, then reinfused into the patient to kill the tumor cells [3]. Many of these therapies demonstrate impressive efficacy effect size and durability in hematological malignancies [4]. The complexity nature of such interventions also means they are resource intensive and can be costly. How to balance PM with generalizability and resource constrain remains a challenge here, e.g., "off-the-shelf" allogeneic CAR-T.

1.2 Why Personalized Medicine

As demonstrated by the examples in Sect. 1.1, evolvement of sciences and technologies are bringing more and more treatment options for patients in needs. However, not all patients are the same, and there is usually no one-size-fits-all solution. The patients' demographics, lifestyle, social-economic status, geographic environment, comorbidities, concurrent medications, genetics, and healthcare system are part of many which may predispose them to different courses of disease progression and varying level of response to a given treatment. To achieve best outcomes, the treatment choice, including initiation, optimization, adjustment, and sequencing, will need to be customized for each individual patient.

PM is also a must for an efficient healthcare system. With an aging society and financial constrain facing all the payers worldwide, avoiding treatments with low probability of success and going with the interventions associated with highest expected outcomes will save resource, improve productivity, and serve patients better.

1.3 How to Practice Personalized Medicine

PM requires systematic collection of wide range of data for a heterogeneous population. This has not been practical in the era of pencil-and-paper for recording medical records. With wide adoption of electronic health records (EHRs) in the current digital age, interaction between tens of millions of patients and healthcare systems are being accumulated on a daily basis. In countries with a single payer system, almost each of the citizen is contributing data continuously from birth to advanced age. This provides a unique opportunity for researchers to tap into the rich real-world data (RWD) sources to conduct research for the purpose of PM.

In the meantime, the pure volume of RWD, the complexity of RWD types (e.g., structured and unstructured data elements), the higher proportion of missingness of

certain variables than the randomized clinical trials also pose challenges for robust PM research. To address these road blockers, on the one hand, cloud computing or parallel distributed computing infrastructure enables researchers to analyze big data in real-time. On the other hand, the active research in machine learning, the advancement beyond the classical statistics to effectively analyze unstructured data such as free text, imaging, voice provide the tools for researchers to gain further insights despite the challenges.

This chapter is organized as follows. Section 2 introduces the important role of causal inference framework in personalized medicine and advanced analytics to check its underlying assumption. Also included are advanced analytic methodologies to address linked data sources and unstructured data. Section 3 reports subgroup identification and individualized treatment regimens at a single time point under causal inference framework with advanced analytics to implement specific steps. Section 4 extends it to multiple time points, i.e., dynamic treatment regimens (DTR). Concluding remarks are summarized in Sect. 5.

2 Role of Causal Inference and Advanced Analytics

2.1 Conditional Average Treatment Effects

As stated in ICH E9(R1) [5], the fundamental research question in drug development is to establish causal relationship between medical intervention and outcomes. The causal inference framework based on counterfactual outcome plays a central role in effectiveness assessment. Chapters "Causal Inference with Targeted Learning for Producing and Evaluating Real-World Evidence", "Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods", and "Sensitivity Analysis in the Analysis of Real-World Data" of this book describe the causal framework and different causal inference methodologies in more detail. One important and commonly used measure for causal inference is the average treatment effect (ATE), which measures the difference in mean (average) outcomes between the treatment and control groups. The potential outcome framework provides a useful way of identifying ATEs in observational studies. The conditional ATE (CATE) which follows the same concept as ATE but conditioned on patient's characteristics is of special relevance for personalized medicine and serves as the criteria for treatment decision making specific to that patient.

Let $\{(X_i, A_i, Y_i), i = 1, \cdots, n\}$ denote the data, where X_i is the covariate vector for subject i, $A_i = a \in \{0, 1\}$ is the treatment assignment, and Y_i is the observed outcome. Using the notation of potential outcome under causal inference framework [6], let $Y_i(1)$ and $Y_i(0)$ denote the potential outcome for subject i receiving the new treatment $A_i = 1$ and standard treatment $A_i = 0$. CATE is defined as the conditional mean difference between potential outcomes given $X = x$:

$$\Delta(x) = E\left(Y(1)|X = x\right) - E\left(Y(0)|X = x\right)$$

Under the strongly ignorable treatment assignment assumption ($A \perp \{Y(0), Y(1)\}|X$), consistency (the potential outcome is $Y(a) = Y$ for every individual with $A = a$), and positivity ($P(A = a|X = x) > 0$) assumptions,

$$\begin{aligned}\Delta(x) &= E(Y(1)|X = x) - E(Y(0)|X = x)\\ &= E(Y(1)|A = 1, X = x) - E(Y(0)|A = 0, X = x)\\ &= E(Y|A = 1, X = x) - E(Y|A = 0, X = x.)\end{aligned}$$

Note that the last expression only involves conditional expectations from observed values.

For causal inference to be valid, the underlying assumptions described above need to hold. Sensitivity analysis plays an important role in assessing the impact of violating such assumptions (see chapter "Sensitivity Analysis in the Analysis of Real-World Data" for a detailed account of this topic). Here, we focus on some recent developments on quantifying and addressing two of the assumptions, positivity and strongly ignorable treatment assignment assumption (also known as unconfoundness).

Positivity requires each patient to have a non-zero probability of receiving each treatment. If certain patients always receive a specific treatment, we would not be able to quantify the treatment difference for them without additional assumptions. This can be especially challenging for high dimensional patients' characteristics. To this end, several machine-learning methods have been developed to identify overlapping regions where different treatment groups are well represented. As recommended in the review paper of Bica [7], the algorithm OverRule [8] outputs rule-based characterization of overlap where positivity assumption is valid. The authors formalize the problem as finding minimum volume sets subject to coverage constraints and reduce it to a binary classification. They further demonstrated that the algorithm leads to intuitive and informative explanations via application to several real-world case studies. The open source code built upon Python is available at https://github.com/clinicalml/overlap-code. In practice, this step can be conducted before the formal inference, either to verify if there is sufficient overlap in the overall region or identify specific regions where positive assumption holds to allow valid causal inference.

Unconfoundness refers to no unobserved confounders. This is a strong assumption and usually cannot be tested in practice. The domain knowledge is critical in assessing this, e.g., opinions from subject matter experts on completeness and relevance of the data, established causal links from previous research. As an effort to weaken the requirements of no unobserved confounder, Wang and Blei [9] proposed the deconfounder which employs unsupervised machine learning to derive latent variable that will substitute the unobserved confounders in multiple-cause setting. The idea is to fit a good latent variable model of the treatment assignment mechanism to capture the joint distribution of all causes. A probabilistic factor model such as mixture models, mixed-membership models, and deep generative models are good candidates. Assume the fitted factor model works well, all causes will be conditionally independent given the local latent factors. This is the same

idea as generalized propensity score where the latent variables are used in place of confounders. This approach has two advantages: (1) Instead of the nontestable unconfoundness assumption, the performance of the latent variable model can be quantitatively assessed and (2) it requires weaker assumption than classical causal inference. In practice, this approach represents an attractive alternative to the propensity score based or other causal inference methods.

2.2 Data Source for Personalized Medicine

Randomized clinical trials (RCTs) remain the gold standard to establish the efficacy and safety of a medical intervention. It continues to be the foundation for vast majority of regulatory approvals due to its high internal validity and strong control of Type I error rate. However, two main limitations of RCT are the external generalizability to a more diverse, more heterogeneous patient population and the controlled experimental environment which is systematically different from the routine clinical practice. For these reasons, real-world data provides unique value in the era of personalized medicine. For optimal treatment recommendation, there is the distinction of static or longitudinal setting. In the static setting, the objective is to make a one-time treatment decision (see Sect. 3 for more details). In the longitudinal setting, the aim is to choose a sequence of treatments and associated timing (covered in Sect. 4 of dynamic treatment regimes, or DTR).

The data for constructing optimal DTR usually comes from either sequentially randomized trials or longitudinal observational studies. A special class of sequentially randomized trials called Sequentially Multiple Assignment Randomized Trial (SMART) is designed with a goal to inform development of optimal DTRs [10–12]. In SMART design, same subjects proceed through multiple treatment stages, and at each stage, subjects may be randomized to one of the available treatment options based on information collected after previous treatments, but prior to assigning the new treatment. SMART data provides high-quality evidence free from confounding bias through randomization. But it is limited to reflect real-world circumstances and usually requires substantial resource to conduct a high-quality trial with adequate power. This is where longitudinal observational studies have several advantages over sequentially randomized trials as it is less costly, more feasible, better reflects the heterogeneity among patients and treatment options in real-world.

Challenges remain when leveraging RWD for personalized medicine. As many of the RWDs are repurposed for research purposes, one single data source may not contain all the necessary data elements. Linking different data sources can substantially increase their breadth and depth to enable more robust RWE generation. Chapter "Privacy-Preserving Record Linkage for Real-World Data" details the methodology and practical consideration for data linkage. While integrating, synthesizing different data sources, data disparity, which refers to the fact that not all data sources contain comparable information, needs to be carefully addressed. One example is that a national health insurance claims database may contain information

on millions of patients' encounter with the healthcare system while a regional electronic health records database captures more detailed clinical info for a subset of patients. There are rich literatures existing to address data disparity involving multiple data sources and we report two below.

Chatterjee et al. [13] built regression models using individual-level data from one source ("internal" study) and aggregated data information from another ("external" study). The "internal" study contains the outcome and all covariates of interest while the "external" study includes the outcome and a subset of these covariates. They adopted a semiparametric framework which allows the distributions of all covariates to be unspecified. By assuming the distribution functions of all covariates to be the same for the "external" and "internal" studies, the external information is converted into a set of constrains when the full model is correctly specified. This allows the improvement of efficiency of parameter estimates and generalizability of models via Lagrange multipliers for model calibration. The assumption that the underlying populations for the internal and external studies are identical may not be practical. Deviation from this assumption can lead to severe bias for any type of calibration method. The performance of the model can be improved by availability of an external reference sample that can estimate the covariate distribution for the external population unbiasedly.

A more recent literature [14] considered the estimation of causal effects to combine a big main data and a subset of the main as validation data. The main data is large but has unmeasured confounders while the validation data is carefully designed to provide supplementary information of unmeasured confounders. Causal inference solely based on the large main data leads to error-prone estimator. The estimators solely based on validation dataset are valid but may not be efficient. The authors applied the same error-prone procedure to both the main and validation data, i.e., leaving out the supplemental information of unmeasured confounders in the validation data. The only requirement for the two error-prone estimators is they are consistent for the same parameter such that their difference is a consistent estimator of 0. Furthermore, this difference is associated with the average causal effect estimated from the full validation data. Therefore, the difference of two error-prone estimators can be leveraged to improve the efficiency of initial estimator solely based on validation data. This framework can be applied to the commonly used causal inference estimators such as regression imputation, inverse probability weighting, augmented inverse probability weighting, and matching. And it does not require the patients in the validation study to be a random sample of those from the large main study. This framework can also cover the setting with multiple data sources.

Another key feature of RWD is the availability of free text data that can substantially augment the structured data elements. For example, physician notes captured in the electronic health records can reflect the underlying reasoning, context of the clinical decision making, such as rationale for initiating, switching, discontinuing a medical intervention. They can also capture more detailed information of patients' characteristics and outcomes, e.g., the pain level as measured by a score, emerging, worsening, resolving of an adverse event, environmental or genetic risk factors.

Natural language processing (NLP) is developed to more effectively process free text data, and great progress has been made in this area. In November 2018, Google researchers published the source code of a new language representation model called Bidirectional Encoder Representations from Transformers (BERT) where the algorithm exceeded human performance in the very competitive Stanford Question Answering Dataset [15]. This is a pre-trained model that can be fine-tuned with just one additional output layer for customization toward specific task. Therefore, it enables effective transfer learning.

In a pilot project we conducted [16], the outcome was generated from a logit model where the linear predictor includes 10 correlated biomarkers (structured data) and one text predictor describing loneliness. The text predictor includes one to three sentences with noise added. The training dataset consists of 300 records and the testing dataset includes another 300 records. Without labeling the text predictor, the customization of BERT extracted a total of 1024 features using the training dataset. These features were then combined with the 10 biomarkers to assess their performance in the testing dataset. Compared with the model solely based on structured data, the area under curve (AUC) increases from 0.765 to 0.885 after the incorporation of the unstructured data. Although each of the 1024 features extracted from text predictor does not allow a clinically relevant interpretation, the text score calculated from a generalized linear model has clear interpretation with the high score corresponding to a feeling of loneliness.

3 Subgroup Analysis

One aspect of personalized medicine is to understand heterogeneity in patient populations and hence to develop new treatments that target a subgroup of patients with enhanced risk-benefit profile or tailor the currently available therapies to a given patient. The question of identifying the right patient for a given treatment is fundamentally different from asking which treatment performs the best in the overall population [17].

Wijn et al. [18] reviewed guidance from industry, health technology assessment (HTA) organizations, academic/non-profit research organizations, and regulatory bodies on subgroup analysis. They found that statistical recommendations were less common and often limited to a formal test of interaction. Detection of interaction between treatment and covariates is essential for subgroup analysis. However, such an approach has faced several challenges. First, the predefined subgroups may not identify all true heterogeneity structure of the effect of the investigational treatment. Second, significance testing of many subgroups leads to the challenge of multiplicity control and a potential lack of power. Third, it might be hard to clearly define subgroups a priori, even for the field experts. Fourth, the form of interactions can be linear vs. non-linear, order can be first- second- vs. high-order. Fifth, the interaction on one scale can disappear when the data have been transformed to another scale [19–21]. In the RW setting, the multiplicity issue could be of less concern for a

study that is exploratory in nature. For a RW study, however, the handling of casual inference and potentially high-dimensional complex data introduces an additional layer of complexity.

In the remainder of this section, we first review three different classes of methods for subgroup identification for RWD in Sect. 3.1, and then provide some discussions in Sect. 3.2.

3.1 Methods for Subgroup Identification

We witness the recent development of a variety of approaches for subgroup identification. In this subsection, we review three different classes of methods for RWD. The promising methods feature a duet between causal inference and statistical/machine learning. To address causal effect inference for observational studies, G-formula [22], weighting [23], and doubly robust methods [24] are widely used. To analyze RWD of larger size and more complexity, regression with regularization and supervised learning [25] are broadly adopted. In the following, we introduce different methods for subgroup identification addressing causal inference for observational studies. Furthermore, we summarize statistical/machine-learning methods for implementation of different proposals.

3.1.1 Identify a Subgroup by Thresholding Treatment Effect

We can identify a subgroup by specifying a predetermined threshold of clinical significance, and then searching for subjects whose treatment effect satisfy the prespecified criterion. Following the notation in Sect. 2, we can obtain a subgroup S(x) based on the value of estimate of CATE: $\hat{\Delta}(x)$, we can obtain a subgroup S(x). For example, $S(x) = \{x : \hat{\Delta}(x) > \delta\}$, a subgroup in which every subject has a conditional CATE larger than δ. Note that a subgroup can also be obtained based on the estimate of $E(Y|A = 1, X = x)/E(Y|A = 0, X = x)$ or $E(U(Y)|A = 1, X = x) - E(U(Y)|A = 0, X = x)$, where $U(\cdot)$ is a monotone transformation.

To estimate a treatment effect, we can first fit respective response surface for different treatment groups. We can use one model/algorithm $m(x, a)$ to estimate the conditional mean $E(Y|A = a, X = x)$. For example, we can fit a regression model to both treatment groups. Alternatively, we can adopt different models/algorithms for different treatment groups: $E(Y|A = a, X = x) = m_a(x)$, where $a \in \{0, 1\}$ and $m_1 \neq m_0$. Furthermore, we can select different covariates for m_1 and m_0. Following [26], we hereafter refer to the approach using a single model/algorithm as "S-learner" (with "S" being short for single), and the approach fitting different groups separately as "T-learner" (with "T" being short for two). Künze et al. [26] recently proposed a "X-learner," which builds on the T-learner and uses each observation in the training set in an "X"-like shape. There are three stages in X-learner. First, obtain the estimate of response for new and standard treatment $\hat{m}_1$ and $\hat{m}_0$. Second,

obtain the imputed individual treatment effect for a subject i in new treatment group as $Y_i - \hat{m}_0(X_i)$ and a subject j in standard treatment group as $\hat{m}_1(X_j) - Y_j$. Using the imputed treatment effects as the response variable in the new treatment group to obtain $\hat{\Delta}_1(x)$ and similarly in the standard treatment group to obtain $\hat{\Delta}_0(x)$. Finally, a weighted average of the two estimates in stage 2 is the final estimate: $\hat{\Delta}(x) = w(x)\hat{\Delta}_0(x) + (1 - w(x))\ \hat{\Delta}_1(x)$, where $w(x) \in [0, 1]$ is a weight function. We can use propensity score as an estimate of $w(x)$.

For covariates, we can distinguish between prognostic and predictive ones X_{prog} and X_{pred}, where X_{prog} are the main terms in the model and X_{pred} are incorporated into the model as interaction terms with treatment. There can be overlap between X_{prog} and X_{pred}, and the number of X_{pred} is typically less than the number of X_{prog}.

To mitigate the confounding, the counterfactual framework or potential outcomes model using g-formula was studied, for examples, in [21, 27–30]. Modeling a response surface that depends on estimates of the propensity score as a covariate was proposed in [31].

Motived by estimating parametric components in partially linear models (Robinson's transformation) [32], an example of a doubly robust estimator, Nie and Wager [33] recently developed a general two-step algorithms for treatment effect estimation in observational studies. This approach was referred to as "R-learner" in recognition of Robison's work.

Let $m^*(x) = E(Y|X = x)$ and $\pi(\mathrm{x}) = P(A = 1|X = x)$, by [32],

$$Y_i - m^*(x_i) = (A_i - \pi(x_i))\,\Delta(x_i) + \varepsilon_i.$$

The estimate $\hat{\Delta}(\cdot)$ can be obtained by empirical loss minimization,

$$\hat{\Delta}(\cdot) = \mathrm{argmin}_{\Delta}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\left(Y_i - m^*(x_i)\right) - (A_i - \pi(x_i))\,\Delta(x_i)\right]^2 + \Lambda_n[\Delta(\cdot)]\right\} \tag{1}$$

where $\Lambda_n[\Delta(\cdot)]$ is a regularization term on the complexity of the $\Delta(\cdot)$ function.

A two-step estimation was proposed. In the first step, obtain estimate $\hat{m}^*$ and $\hat{\pi}$ with cross-validation for optimal predictive accuracy. In the second step, with a plug-in estimate $\hat{m}^*$ and $\hat{\pi}$, solve the optimization with respect to Δ in Eq. (1).

Motived by the R-leaner, a forest-based method named causal forest was implemented in an R package 'grf' [34, 35]. Causal forest starts by obtaining out-of-bag estimate $\hat{m}^*$ and $\hat{\pi}$, and then grow a causal forest via:

$$\hat{\Delta}(\mathrm{x}) = \frac{\sum_{i=1}^{n}\alpha_i(x)\left(Y_i - \hat{m}^*(x_i)\right)\left(A_i - \hat{\pi}(x_i)\right)}{\sum_{i=1}^{n}\alpha_i(x)\left(A_i - \hat{\pi}(x_i)\right)^2},$$

where $\alpha_i(x)$ measures how often the i-th training example falls in the same leaf as x. Causal forests with different values of the tuning parameters are trained to choose the ones that minimize the loss in Eq. (1).

Instead of first estimating the outcome from different treatment groups respectively, we can focus on the treatment effect $\Delta(x)$ directly. A transformed/modified outcome method used to estimate the treatment effect directly by modeling the interaction effect without modeling of the main effects for continuous endpoints was proposed in [36]. To conduct causal inference, a transformed/modified outcome was defined using weighting method as follows [37, 38]:

$$Y_i^* = A_i \frac{Y_i(1)}{\pi\,(X_i)} - (1 - A_i)\,\frac{Y_i(0)}{1 - \pi\,(X_i)}.$$

Note that $E(Y^* | X = x) = \Delta(x)$ and transformed outcome only depends on the potential outcome corresponding to the realized treatment level, and hence can be calculated from the observed outcome of Y_i. With the transformed outcome

$$Y_i^* = Y_i \cdot A_i^*,$$

where $A_i^* = \frac{A_i}{\pi(X_i)} - \frac{1-A_i}{1-\pi(X_i)}$, we can implement any existing methods to directly estimate the treatment effect.

To handle binary and time-to-event outcome, Tian et al. [36] proposed a modified covariate estimator. However, as a limitation, the modified covariate method is primarily meant for analyzing the data from randomized clinical trials.

For implementation of different approaches, regression models were widely used. Different models and parameter estimation approaches were adopted. Parametric (linear, generalized linear), semi-parametric (generalized additive models), and non-parametric (kernel regression) models can be adopted to model outcome [27, 39]. Various regulation methods including L2, Lasso penalty and different Lasso penalties for prognostic and predictive variables were utilized. Furthermore, to fit model with regularization, constrained cross-validation and generalized cross-validation were proposed. In the Bayesian framework, a Bayesian two-step Lasso method was proposed in [40] for variable selection. The first step used a group Lasso to screen out unimportant variables and a more efficient variable selection could be achieved in the second step using an adaptive Lasso. In addition to accommodating variable selection, to handle multiplicity issue, Schnell et al. [41] developed Bayesian simultaneous credible bands for continuous endpoints. The method was extended to survival and count data in [42, 43].

In addition to regression modeling, tree ensemble methods were proposed in the literature to model outcomes, which can provide better predictive performance for high dimensional data. A random forest (RF) was used to estimate the outcome in [21]. Instead of using the ordinary RF, the counterfactual synthetic RF provided even more promising results [28]. In [30], gradient boosting trees (GBT) was proposed to estimate the treatment effect for continuous, binary, and time-to-event endpoints. The Bayesian adaptive regression trees (BART) (which can be viewed as a Bayesian regularized tree boosting method) [44] has been demonstrated as a promising method for causal inference [31]. Estimate of treatment effect using the

counterfactual approach for continuous and time-to-event outcomes was studied in [28, 29, 45].

3.1.2 Identify a Subgroup by Maximizing Difference of Treatment Effect Using Tree-Based Method

Another class of methods directly search subgroups using tree-based methods. They focus on maximization of the difference of treatment effect between two resultant child nodes during construction of a tree. In other words, the best split demonstrates the greatest interaction with the treatment.

The interaction trees (IT) using the classification and regression trees was proposed to conduct subgroup analysis. In [46], the best split s^* when constructing a tree is the one that maximizes the following G statistic among all permissible splits:

$$G(s) = \left(\frac{\left(\overline{y}_1^L - \overline{y}_0^L\right) - \left(\overline{y}_1^R - \overline{y}_0^R\right)}{\hat{\sigma}\sqrt{\sum 1/n_A^C}} \right)^2,$$

where $\overline{y}_A^C$ is the sample mean in left ($C = L$) or right ($C = R$) child node for treatment group A, n_A^C is the corresponding sample size, and $\hat{\sigma}^2$is the estimate of variance.

An initial large tree was constructed given the constraint of purity of a node, size of a node, and depth of the tree. A pruning and selection of the best subtree after generating a nested sequence of subtrees was proposed based on the trade-off between interaction and complexity of a tree (in the same spirit as model section criteria AIC, BIC, etc.). Furthermore, they proposed to have an "honest" estimate of the goodness-of-split using an independent subset of the data or cross-validation or bootstrapping method when finding the best subtree. IT method for subgroup analysis of survival, continuous, and longitudinal data were studied in [47, 46, 19], respectively.

For observational studies, a causal inference tree (CIT) which splits data in a way that both the propensity and the treatment effect become more homogeneous within each resultant partition was proposed in [20].

There are other tree-based methods for subgroup identification for randomized studies. Focusing on the identification of "interesting" areas in the covariate space, instead of the whole covariate space, a subgroup identification based on differential effect search (SIDES) approach was proposed in [48]. Only variables that have not been previously chosen are considered for splitting each node. Methods proposed in [49] extends the generalized unbiased interaction detection and estimation (GUIDE) methods [50] to overcome selection bias. Furthermore, a sequential bootstrapping and aggregating of thresholds from trees (BATTing) was proposed in [51] to address the potential issue of unstableness of tree to small perturbations in the data and proneness to over-fitting. The extension of these methods to conduction of causal inference for observational data, however, warrants further research.

3.1.3 Identify a Subgroup by Selection of an Optimal Treatment Regime

Different patients may benefit from different available treatments. A patient can be assigned to a subgroup by identification of an optimal individualized treatment rule/regime (ITR). Patients of the same optimal ITR form a subgroup. Let $d(X)$ be a decision rule (map/scoring system) guiding a treatment assignment decision. An optimal ITR, d^{opt}, is a rule that maximizes the value function: an expectation of weighted outcome as follows:

$$E\left[\frac{I\left(T=d(X)\right)}{T\pi(X)+(1-T)/2}Y\right] \tag{2}$$

where I is the indicator function, π is the propensity score, and $T = 2A - 1$. Note that the ITR focuses on the decision at a single timepoint, which is a special case of a sequence of decision rules at different timepoints as discussed in dynamic treatment regime in Sect. 4.

Maximizing the value function in Eq. (2) is equal to minimizing:

$$E\left[\frac{Y}{T\pi(X)+(1-T)/2}I\left(T\neq d(X)\right)\right]$$

which can be viewed as a weighted classification error, i.e., we classify T (assign treatment) using covariates X with weight of each misclassification equal to $Y/(T\pi + (1 - T)/2)$. This was referred to as the outcome weighted learning (OWL) in [39].

Furthermore, since $d(X)$ can always be represented as $\text{sign}(g(X))$, for some scoring function $g(\cdot)$, given observed data, the optimal ITR can be obtained by finding a solution to minimize the following weighted classification error:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{T_i\pi\left(X_i\right)+(1-T_i)/2}\mathrm{I}\left(T_i\neq \text{sign}\left(g\left(X_i\right)\right)\right) \tag{3}$$

The d^{opt} can be obtained by finding the solution $g^*(X_i)$ to the above optimization and then setting $d^{\text{opt}} = \text{sign}\left(g^*(X_i)\right)$. For a subject with large outcome, this rule is more likely to recommend the same treatment assignment that the subject has received. However, for a subject with small outcome, the rule tends to assign the opposite treatment assignment to what has been received. See [39] for details.

Note that the above-described method solves the ITR question by optimization of an inverse probability weighted estimator (IPWE). A doubly robust augmented inverse probability weighted estimator (AIPWE) was proposed in [52]. An optimal ITR, d^{opt}, is a rule that maximizes the following function:

$$E\left[\frac{I\left(T=D(X)\right)}{T\pi(X)+(1-T)/2}\,Y\right.$$
$$\left.-m\Big(X,D(x)\Big)\left\{\frac{I\left(T=D(X)\right)-\{T\pi(X)+(1-T)/2\}}{T\pi(X)+(1-T)/2}\right\}\right]$$

The estimate is consistent if either propensity score or the response model, but not both, is mis-specified.

A general framework using either propensity weighting or advantage learning (A-learning) was proposed in [53]. Note that the A-learning was originally proposed for estimating optimal dynamic treatment regime. It focuses on modeling the treatment effect and has double robustness property. See Sect. 4 for more details.

For the weighting method, we need to find a score function $g(X)$ that minimizes the following quantity

$$L_W\left(g,X\right)=E\left[\frac{M\left\{Y,Tg(X)\right\}}{T\pi(X)+(1-T)/2}|X=x\right] \tag{4}$$

For the A-learning method, we need to find a score function $g(X)$ that minimizes the following quantity

$$L_A\left(g,X\right)=E\left[M\Big\{Y,\left((T+1)/2-\pi(X)\right)g(X)\Big\}|X=x\right] \tag{5}$$

where $M\{Y,v\}$ is a loss function. Note that let $g_W^*=\text{argmin}_g L_W(g)$ and $g_A^*=\text{argmin}_g L_A(g)$, the d^{opt} equals sign $\left\{g_W^*(X)\right\}$ and sign $\left\{g_A^*(X)\right\}$ for weighting and A-learning, respectively. Furthermore, the magnitude of treatment effect can be estimated.

Different loss functions were described for continuous, binary, and survival outcomes in [53]. For examples, quadratic loss for continuous, logistic loss for binary outcome, and negative log partial likelihood for the time-to-event outcome. In addition, a doubly robust AIPWE estimator proposed in [52] can be obtained using a generalized augmented loss.

Minimization of Eq. (3) can be solved in the context of solving a classification problem. A support vector machine (SVM) with a hinge loss as a convex surrogate loss to 0–1 loss was proposed in [39]. For model assumptions of $g(X)$ in solving minimization of Eqs. (4) and (5), a linear model with regularization term, such as lasso-type of penalty, can be used when the number of covariates is large. When there are functional baseline covariates in addition to scalar ones, we can represent the functional covariates and their corresponding coefficient functions in terms of some suitably chosen set of basic functions, then we can view the loss function as wholly consisting of scalar quantities. Refer to [53, 54] for details. A boosting approach can be implemented as well.

3.2 Discussion

As we described in Sect. 3.1, different subgroup identification approaches fit for different purposes. For example, approaches introduced in Sects. 3.1.1 and 3.1.2 identify subgroups of patients achieving a target treatment effect and maximized treatment effect difference, respectively.

For implementation of different subgroup identification methods, there is no universally optimal learner or model/algorithm. A learner and model/algorithm will be competitive when they approximate the data-generation process sufficiently. For example, among learners introduced in Sect. 3.1.1, the R-learner stands out in certain simulation setups but is not as good as the T-learner when the two treatment groups are unrelated [33]. If the response surfaces of the outcomes under new and standard treatment are very different, fitting different treatment group separately using a T-learner will outperform a S-learner, which pools the data. The X-learner performs well especially when one of the treatment groups is much larger than the other or when the separate parts of it are able to capture the properties of the response and treatment effect functions [26].

With respect to model/algorithm, we need to continue tapping into the development of statistical/machine learning, a field of rapid growth. For examples, a sparse Gaussian process regression model, using recursive partitioning in Bayesian additive frame work, was proposed in [55]. The new proposal can capture both global trends and local refinements. By using the pseudo values, a deep neural network method that reduces a complex survival analysis to a standard regression problem was proposed in [56]. The method greatly simplified the neural network construction and provided promising results. These new approaches can potentially improve results when integrated into the subgroup analysis. Furthermore, when there are multiple models/algorithms (base learners) to estimate outcome or propensity score, a super learner can be built to ensemble a group of base learners [57].

When there are different approaches to estimate treatment effect of a subgroup, one may choose the one which gives us the largest subset of patients; the largest area under the treatment effect curve (AUC) or consider the area between treatment effect curves (which measures the relative improvement from one method to another). See reference [58] for details.

We need to synergize successful estimation using highly adaptive and more accurate methodologies with ease of interpretability. To identify key covariates for subgroup identification, the variable importance can be measured in different ways, e.g., by permutation of each covariate and assessing the loss in performance. An example of identification of important predictive biomarkers using RF and boosting methods can be found in [59]. A general discussion on measuring predictor importance with different outcomes and models/algorithms can be found in [60]. An advantage of tree-based methods is that we can directly obtain the information on subgroups as space partitioning in the form of $X_j \leq c_j$ or $X_j > c_j$. To facilitate interpretation, in [21] a tree was used after obtaining estimate of treatment effect using RF.

The missing data can be handled as an integrated part of algorithms/methods described in previous subsections, in addition to pre-imputation. Missing data can be simultaneously imputed while using tree-ensemble methods including RF, BART, and GBT [61–63].

The approaches described in this section focus on two treatment groups. The generalization to multiple treatments can be naturally accommodated for response surface fitting methods by either pooling all treatment groups together or fitting individual group separately. An extension was proposed with an additional assumption on the loss $M\{Y, v\}$ using the weighting method in [53]. A multivariate version of Robinson's transformation with regularization term reflecting relationships between the treatment effects of different arms was described in [33].

After identification of a promising subgroup (or multiple subgroups), we need to assess how good a subgroup really is. The usual statistical inference on treatment effect, assuming that the subgroup is chosen independent of the data, may lead to an overly optimistic evaluation due to selection bias (see [64, 65] and references therein for details). Cross-validation or repeated cross-validation were proposed in [21, 51, 58] to assess the treatment effect in subgroup. Parametric bootstrap and bias-corrected bootstrap methods were studied in [21]. Bootstrap method was also proposed in [49] to obtain confidence interval of treatment effect. Asymptotic confidence intervals for the treatment effect were derived using honesty tree, where an observation is used to estimate treatment effect or to construct a tree, but not both [66]. The posterior intervals can be used to quantify uncertainty for Bayesian methods such as the BART.

BioPharmNet has a website dedicated to subgroup analysis software: https://biopharmnet.com/subgroup-analysis-software/. The website collects a variety of tools (mostly R packages and code) for execution of subgroup analysis.

4 Dynamic Treatment Regime

Personalized medicine is a medical paradigm offering data-driven decision support for treating patients in the presence of heterogeneity. The goal of personalized medicine is to enhance patient outcomes by tailoring treatment based on patient characteristics. A dynamic treatment regime provides a framework for formalizing personalized medicine. A dynamic treatment regime is a set of sequential decision rules, one per stage of treatment. Each decision rule takes input information on the patient (such as demographics, prior medical history, genetic information, evolving physiological and clinical variables, results of diagnostic tests, genetic information, etc.) and returns a recommended treatment option. Dynamic treatment regimes have also been referred to as *adaptive treatment strategies* [10, 12, 67]. Identifying optimal DTRs offers an effective tool for personalized management of diseases and helps physicians tailor the treatment strategies dynamically and individually based on clinical evidence, which provides a key foundation for enhanced care of chronic disease [67, 68].

In the remainder of this section, we will first introduce the statistical framework of dynamic treatment regime in Sect. 4.1, then review different classes of methods for estimation optimal DTRs in Sect. 4.2. Finally, we will provide some discussion in Sect. 4.3.

4.1 Basic Framework

To formalize the concept of DTR, we introduce the basic definitions of a regime involving $K \geq 2$ stages defined in the book [69]. At each stage, a treatment must be selected from a set of available treatment options. We can index the stages by j where $j = 1, \ldots, K$.

For $j = 1, \ldots, K$, let $\mathcal{A}_j$ be the set of treatment options available to decision point j, where a_j denotes an option in $\mathcal{A}_j$. For simplicity, we consider the case where the number of options in $\mathcal{A}_j$ is finite. It is possible for $\mathcal{A}_j$ to be an infinite set, i.e., when the treatment options are drug doses in a continuous range of possible doses, which we will discuss later in this section.

Let X_1 denote baseline information, X_j denote intermediate information collected between decision $j-1$ and j, $j = 2, \ldots, K$. In general, let $\mathcal{X}_j$ denote the support of X_j, $j = 1, \ldots, K$. Let Y denote the final outcome and let Y_j denote stage-specific outcome following decision point j. For simplicity, only continuous outcome and binary outcome are considered here assuming larger outcome is preferred. Let H_j denote the accrued information or history at decision point j. At decision 1, the accrued information or history is simply the baseline information, $H_1 = X_1$. At subsequent decision points, $H_j = \{X_1, A_1, \ldots, X_{j-1}, A_{j-1}, X_j\}$ for $j = 2, \ldots, K$. In general, let $\mathcal{H}_j$ denote support of H_j.

For $j = 1, \ldots, K$, let $\overline{X}_j = \left(X_1, \ldots, X_j\right) \in \overline{\mathcal{X}}_j = \mathcal{X}_1 \times \cdots \times \mathcal{X}_j$ and $\overline{A}_j = \left(A_1, \ldots, A_j\right) \in \overline{\mathcal{A}}_j = \mathcal{A}_1 \times \cdots \times \mathcal{A}_j$. It follows that $\mathcal{H}_1 = \mathcal{X}_1$, and $H_j = \overline{\mathcal{X}}_j \times \overline{\mathcal{A}}_{j-1}, j = 2, \ldots, K$.

At decision j, a decision rule $d_j(h_j)$ is a function that maps an individual's history to a treatment option in $\mathcal{A}_j$, that is, $d_j : \mathcal{H}_j \rightarrow \mathcal{A}_j, j = 1, \ldots, K$. This means that at decision j, a decision rule is a function that takes the patient history as input and returns a treatment option from the available options. Then a dynamic treatment regime d under this setting is defined as a sequence of such rules; that is

$$d = \{d_1(h_1), d_2(h_2), \ldots, d_K(h_K)\}$$

For simplicity, we express it as $d = \{d_1, d_2, \ldots, d_K\}$. Similarly, we can refer to the subset of the first j rules in a K-decision regime d as

$$\overline{d}_j = \left\{d_1, d_2, \ldots, d_j\right\}, \quad j = 1, \ldots, K$$

$$d = \overline{d}_K = \{d_1, d_2, \ldots, d_K\}$$

Let $\mathcal{D}$ denote the class of all possible dynamic treatment regimes d. The performance of regime $d \in \mathcal{D}$ is represented by the expected outcome that would be achieved if all patients in the population were to receive treatment according to regime d; that is, the value of d. Assuming larger outcomes are preferred, the optimal treatment regime $d^{opt} \in \mathcal{D}$ is the one that maximizes the expectation of cumulative outcome among all $d \in \mathcal{D}$.

To formalize this, we state it under potential outcomes framework [6] following the book [69]. Consider a randomly chosen patient with baseline information X_1. At decision 1, $H_1 = X_1$. Suppose the patient receives treatment option $a_1 \in \mathcal{A}_1$, then the information that would accrue on this individual between decision 1 and 2 depends on the treatment option the patient received at decision 1. Let a random variable $X_2^*(a_1)$ represent the potential intervening information between decision 1 and 2 that would occur if a randomly chosen individual were to receive option $a_1 \in \mathcal{A}_1$ at decision 1. Continuing in this way, for the sequence $\overline{a}_{j-1}$ at decision 1 to $j-1$, the potential intervening information that would occur between decision $j-1$ and j is represented by the random variable $X_j^*\left(\overline{a}_{j-1}\right), j = 2, \ldots, K$. If options $a_1, a_2, \ldots, a_K$ were given at all K decisions, the potential outcome that would be achieved if a randomly chosen individual were to receive treatment sequence $\overline{a} = \overline{a}_K = (a_1, \ldots, a_K)$ across all K decision points is

$$Y^*(\overline{a}_K) = Y^*(\overline{a})$$

The potential information arising between decision points throughout the K decisions and the potential outcome arising if an individual were to receive the K-stage sequence of treatments $\overline{a} = \overline{a}_K = (a_1, \ldots, a_K)$ is

$$\left\{X_1, X_2^*(a_1), X_3^*(\overline{a}_2), \ldots, X_K^*(\overline{a}_{K-1}), Y^*(\overline{a})\right\}$$

Therefore, for a given regime $d \in \mathcal{D}$, the potential outcome that would be achieved if an individual were to receive treatment according to the K rules in d is denoted by

$$Y^*(d) = Y^*\left(\overline{d}_K\right)$$

We can also write the potential outcomes associated with regime $d \in \mathcal{D}$ as

$$\left\{X_1, X_2^*(d_1), X_3^*\left(\overline{d}_2\right), \ldots, X_K^*\left(\overline{d}_{K-1}\right), Y^*(d)\right\}$$

We define the *value* of regime $d \in \mathcal{D}$ as the expected outcome that would be achieved if all K rules in d were followed to select treatment.

$$V(d) = EY^*(d)$$

Therefore, an optimal regime, denoted $d^{opt} \in \mathcal{D}$ is the one that maximizes the value among all $d \in \mathcal{D}$. That is

$$EY^*(d^{opt}) \geq EY^*(d) \text{ for all } d \in \mathcal{D}$$

Or, equivalently

$$d^{opt} = \underset{d \in \mathcal{D}}{\text{argmax}} EY^*(d) = \underset{d \in \mathcal{D}}{\text{argmax}} V(d)$$

In order to estimate a DTR from either randomized or observational data, we need to make the following identifiability assumptions [69]. Detailed description of these assumptions can also be found in Sect. 2.

1. Stable unit treatment value assumption (SUTVA): A subject's outcome is not influenced by other subjects' treatment allocation. This assumption has also been referred as consistency assumption.
2. Positivity: every subject follows a specific DTR with a non-negative probability, which is bounded away from 0. We can state it as $P(A_j = a_j | H_j = h_j) > 0$ for options a_j that are feasible for history h_j and for all possible histories h_j that satisfy $P(H_j = h_j) > 0$.
3. Sequential Randomization Assumption (SRA): this assumption is a generalization of the no unmeasured confounders assumption to the multiple decision case. It states that treatment selection at decision j depends only on an individual's observed history H_j and not additionally on potential outcomes.

4.2 *Methods for Estimating Optimal Dynamic Treatment Regimes*

An optimal DTR is the one that optimizes the expected cumulative clinical outcome. The optimal DTRs can provide evidence-driven precision medicine and are especially valuable in chronic disease management. Various methods have been proposed to estimate the optimal DTRs, which can be generally classified as either indirect or direct estimation methods [70–72]. Indirect estimation methods use approximate dynamic programming with parametric or semiparametric methods to first estimate models for the conditional means or contrasts of conditional mean outcomes and then from these models infer the optimal DTR [70]. Methods under this class include g-estimation in structural nested models [73–75] and its variations, Q-learning [72, 76–78], A-learning [68, 79], and regret regression [80]. Direct estimation methods, also known as *value maximization methods* [81] or *policy search methods* [72, 81], directly estimate the value or marginal mean for all DTRs in a pre-specified class and then select the regime that maximizes the estimated value. A variety of methods, such as marginal structural models [82, 83] and inverse probability weighting [84], fall into this class. We will not give a comprehensive

review of all the methods, instead we will review some influential and commonly used methods. A more thorough review could be found in these books [69, 72].

4.2.1 Indirect Estimation Methods

One of the most commonly used indirect estimation methods for estimating optimal DTR is Q-learning, where "Q" denotes "quality" [70, 85]. This method models the conditional expectation of the outcome given history and action. Q-learning is an analog of dynamic programming. It moves backward and recursively estimate the conditional mean, which is known as Q-function. The Q-function is defined as:

$$Q_j\left(H_j, A_j\right) = E\left[\mathrm{Y}_j + \max_{a_{j+1}} Q_{j+1}\left(H_{j+1}, a_{j+1}\right) | H_j = h_j, A_j = a_j\right]$$

where $Q_{K+1} \equiv 0$, Y_j is the reward observed at the end of each stage, $j = 1, \ldots, K$.

The estimated optimal DTR is given by $\left(\hat{d}_1, \ldots, \hat{d}_j\right)$:

$$\hat{d}_j\left(h_j\right) = \underset{a_j}{\operatorname{argmax}} \hat{Q}_j\left(H_j, a_j\right)$$

In practice, the true Q-functions are unknown and must be estimated from the data. Since Q-functions are conditional expectation, a typical approach to model them is through linear regression models. However, one can use more flexible models (e.g., splines, neural networks, etc.) for the Q-functions. Q-learning is appealing because of computational and conceptual simplicity. Regression models are easy to implement and allowing the use of standard diagnostic tools. In addition, it can be performed in most statistical software. However, it suffers the problem of model misspecification as linear models are rarely correctly specified for Q-function.

Other indirect methods estimate optimal DTRs by modeling contrasts of conditional mean outcomes, rather than modeling conditional means themselves (e.g., Q-learning). Popular methods include G-estimation, A-learning, and regret regression. G-estimation relies on structural nested mean models (SNMM) [86, 87]. Typically, the SNMM parameterizes the difference between the conditional expectation of the outcome following observed treatment and the conditional expectation of the counterfactual outcome under potentially unobserved treatment regime. Here, we give a brief introduction of this method following the book [72]. The optimal blip-to-zero function $\gamma_j(h_j, a_j)$ at any stage j is defined as the expected difference in outcome when using a "zero" treatment (refers to placebo or standard of care) instead of a_j at stage j, in persons with treatment and history h_j who subsequently receive the optimal regime $\underline{d}_{j+1}^{opt}$:

$$\gamma_j\left(h_j, a_j\right) = \mathrm{E}\left[\mathrm{Y}\left(\overline{a}_j, \underline{d}_{j+1}^{opt}\right) - \mathrm{Y}\left(\overline{a}_{j-1}, 0, \underline{d}_{j+1}^{opt}\right) | \mathrm{H}_j = \mathrm{h}_j\right]$$

Let ψ be parameters of the optimal blip function. If the true form of the optimal blip function and the true value of ψ were known, then the optimal regime is

$$d_j^{opt} = \arg\max_{a_j} \gamma_j \left(h_j, a_j; \psi\right), \quad j = 1, \ldots, K$$

Robins [79] proposed a method for finding the parameters ψ of the optimal blip function through G-estimation. The expected counterfactual outcome $G_j(\psi)$ is defined as

$$\begin{aligned} G_j(\psi) &= \mathrm{Y} + \sum_{k=j}^{K} \left[\gamma_k \left(h_k, d_k^{opt}; \psi\right) - \gamma_k \left(h_k, a_k; \psi\right)\right] \\ &= Y + \sum_{k=j}^{K} E\left[\mathrm{Y}\left(\bar{\mathrm{a}}_{k-1}, \underline{d}_k^{opt}\right) - \mathrm{Y}\left(\bar{\mathrm{a}}_k, \underline{d}_{k+1}^{opt}\right) | \mathrm{H}_k = \mathrm{h}_k\right] \end{aligned}$$

$G_j(\psi)$ then can be interpreted as subject's outcome adjusted by the expected difference between the average outcome for individual who received a_j and individual who was given the optimal treatment at stage j, where both had the same treatment and history to the start of stage $j-1$ and were subsequently treated optimal.

Robins [79] has proposed the following estimating equation:

$$U(\psi, \alpha) = \sum_{j=1}^{K} G_j(\psi) \left\{S_j\left(A_j\right) - E\left[S_j\left(A_j\right) | H_j; \alpha_j\right]\right\}$$

where $S_j(A_j)$ is a vector-valued function that contains variables thought to interact with treatment to affect a difference in expected outcome, here let $S_j(A_j) = \frac{\partial \gamma_j}{\partial \psi_j} = H_j^{\psi} A_j$. Then $E[S_j(A_j)|H_j; \alpha_j]$ is a function of treatment probability $p_j(A_j = 1|H_j; \alpha_j)$, and α_j is usually estimated from the data using logistic regression. Then G-estimation algorithm proceeds in a recursive manner. It begins with estimating ψ_j by solving the equation system $U_j(\psi_j) = 0$, then moving backward. Assuming the blip function is always correctly specified, the estimators would have the double robustness property.

A-learning is proposed by Murphy [68], where "A" stands for "Advantage". Similar to Q-learning, it also involves recursive backward induction algorithm to find the optimal DTR. The main distinctive feature is the form of the underlying models. While Q-learning models the conditional mean outcomes, A-learning models the contrast function, or equivalently the regret function, which represents the loss incurred by not following the optimal treatment regime. Minimizing the regret function leads to the optimal decision rule at each stage.

For simplicity, we consider the case of two treatment options denoted as $A_j \in \{0, 1\}$ at each stage $j = 1, \ldots, K$, where option 0 is the control or reference treatment. The contrast function is given by

$$C_j\left(h_j\right) = Q_j\left(H_j, 1\right) - Q_j\left(H_j, 0\right)$$

The optimal treatment regime is defined as

$$d_j^{opt}\left(h_j\right) = I\left\{C_j\left(h_j\right) > 0\right\}$$

Since a_j is a binary indicator, $Q_j(H_j, a_j)$ can be written as

$$Q_j\left(h_j, a_j\right) = v_j\left(h_j\right) + a_j C_j\left(h_j\right)$$

where $v_j(h_j) = Q_j(h_j, 0)$. In [79], $Q_j(h_j, a_j) - Q_j(h_j, 0) = a_j C_j(h_j)$ is referred as the optimal blip-to-zero function, which compares the expected difference in outcome between using the reference treatment 0 and using $a_j$among patients with history h_j. A-learning is based on postulating a model $C_j(h_j; \psi_j)$ for the contrast function or blip-to-zero function, depending on a parameter vector ψ_j. Once the estimator of ψ_j is constructed, the estimated optimal regime is obtained by maximizing the estimated optimal blip-to-zero function.

We can see that A-learning does not require full knowledge of Q-function to characterize and estimate an optimal regime. It only requires part of the Q-function representing contrasts among treatments. By reducing the dependence of the estimation procedure on the full data distribution, A-learning is more robust to model misspecification than Q-learning for consistent estimation of the optimal DTR. Schulte et al. [78] have examined the performance of A-learning and Q-learning to identify regions in which one method is superior to the other. The simulation studies suggested that A-learning may be inefficient relative to Q-learning in estimating parameters when having correctly specified models, but A-learning produced more accurate estimator if the Q-function was mis-specified. A-learning may be preferred in settings where it is expected that the form of the decision rules defining the optimal regime is not overly complex. However, A-learning increases in complexity with more than two treatment options at each stage, which may limit its attractiveness.

Another commonly used method in this class is regret regression method, which is similar to G-estimation. Henderson et al. [80] and Almirall et al. [88] proposed two similar methods to model blip or regret function parameters using regret regression. Henderson et al. [80] proposed a method, namely regret-regression, incorporates the regret function of Murphy [68] with regression model for observed response. Almirall et al. [88] proposed a similar methods in a two-stage setting. Review of these methods could be found in the book [72]. Compared with G-estimation, the regression-based estimators have lower variability with correct model specification. In addition, regret regression is appealing as it can make use of the standard regression functions in statistical software, and we can apply the common diagnostic techniques used in linear regression for the choice of the regret function.

In this subsection, we have reviewed indirect methods to estimate the optimal DTR via modeling conditional mean outcomes or contrasts of conditional mean outcomes. All of these methods discussed above can be applied for observational data, and several of them have double-robust property. One advantage of the indirect methods is that the outcome models can be developed using standard statistical

models and be assessed for goodness of fit. However, a major drawback is that the estimator of the optimal DTR requires correct model specification. It is possible that either the Q-functions or the contrast functions are poorly fitted, and thus the derived DTRs may be far from optimal.

4.2.2 Direct Estimation Methods

The methods described above indirectly estimate the optimal DTRs. Another class of methods on the contrary directly estimates the value of each subject in a pre-specified class of regimes $\mathcal{D}$, and then select the regime that maximizes the estimated value. The estimated optimal regime is given by

$$\hat{d}^{opt} = \arg\max_{d \in \mathcal{D}} \hat{V}(d)$$

where $\hat{V}(d)$ is the estimated value function of the regime d.

The most essential part of the above procedures is the estimation of the value function for regime d. A variety of methods have been proposed for estimating $V(d)$. One well-known value estimator $\hat{V}(d)$ from the literature is the IPWE [84]. This approach is motivated by a representation of $EY^*(d)$ in terms of the observed data depending on the conditional probability of receiving treatment given past history. Following the notations in the book [72], let $\pi_j(a_j|H_j)$ denote the probability of taking treatment a_j given history H_j, $j = 1, \ldots, K$. We assume that $\pi_j(a_j|h_j) > 0$ for each action $a_j \in \mathcal{A}_j$ and for each possible value h_j. This is the positivity assumption we discussed before. Let Y be the final outcome. Let $\mathbb{I}\left[A_j = d_j\left(H_j\right)\right]$ be the indication function of whether or not the treatment option an individual actually received is the same as the one selected by d, $j = 1, \ldots, K$. Then the inverse probability weighted estimator of the value $V(d) = EY^*(d)$ is given by

$$\hat{V}_{IPWE}(d) = \mathbb{P}_n\left[\prod_{j=1}^{K} \frac{\mathbb{I}\left[A_j = d_j\left(H_j\right)\right] Y}{\pi_j\left(A_j|H_j\right)}\right]$$

where $\mathbb{P}_n$ is the empirical average over a sample of size n. In the case of a SMART, $\pi_j(A_j|H_j)$ is the randomization probabilities and is known by design, while for an observational study, this can be estimated by the propensity score.

Under the SUTVA, the SRA and the positivity assumption, the IPWE is asymptotically consistent. However, the IPWE estimator is known to be unstable under certain generative models [52]. Zhang et al. [52] proposed a doubly robust estimator of the value function for a single-stage treatment regime, namely AIPWE. In addition to being more robust to model misspecification, double robust estimators tend to be more efficient than IPWE [79]. However, even the principle is straightforward, the implementation is challenging as the AIPWE is a discontinuous function

of the observed data, which makes the optimization computationally demanding even in moderate-sized problems. Details on AIPWE can be found in Sect. 3.1.3.

An alternative approach for estimation of the value of a fixed regime is Marginal Structural Models (MSMs). Marginal structural models are well-established methods to address the problem of time-varying confounding. These models estimate the marginal expectation of a counterfactual outcome under a specific treatment regime. The introduction of MSMs can be found in chapter "Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods". Marginal structural models were initially introduced to estimate the value for static treatment regime [84, 89, 90]. Now these models are becoming increasingly popular for the estimation of optimal DTRs. Orellana et al. [82] proposed method based on an extension of the marginal structural mean model of Robins [89, 91], termed dynamic regime marginal structural mean model, is suitable for estimating the optimal treatment regime with low-dimensional histories and a small class of potential regimes *D*. They introduced both parametric and semiparametric dynamic regime marginal structural models and proposed double robust AIPWE and non-augmented IPWE for the parameters of the dynamic regime MSMs. The attractiveness of MSMs is the simplicity of implementation. A comprehensive review of this method can also be found in these books [69, 72].

A change in perspective for estimating optimal dynamic treatment regime is to connect the problem of maximizing inverse probability-weighted estimator and augmented inverse probability-weighted estimators of *V(d)* with weighted classification problems. Within this framework, the class of treatment regime does not need to be pre-specified and can instead be identified in a data-driven way by minimizing an expected weighted misclassification error and were thereby able to leverage existing classification algorithm to approximately compute $\arg\max_{d\in\mathcal{D}} \hat{V}(d)$. Both Zhao et al. [39] and Zhang et al. [52] have transformed the problem of estimating an optimal treatment regime into weighted classification problem for a single stage setting. Zhang et al. [52] focused on single decision case and estimated the optimal DTR by maximizing across all regimes in the class a suitable doubly robust AIPWE. They showed that the classification-based estimator of the optimal DTR using the AIPWE of the contrast is robust to misspecification of either the propensity score model or the outcome regression model. Zhao et al. [39] developed outcome-weighted learning (OWL) based on the IPWE to identify the optimal regime in the single decision case. The optimal DTR can be obtained by solving a minimization problem with weighted classification error. They employed the hinge loss function that is used in the field of support vector machine (SVM) for solving the classification problem. More details can be found in Sect. 3.1.3.

Zhao et al. [92] further introduced two novel methods under multi-stage setting, named backward outcome-weighted learning (BOWL) and simultaneous outcome-weighted learning (SOWL). These approaches formalize the problem of estimating an optimal DTR as either sequential or simultaneous classification problem. BOWL is a backward formulation of the classification-based approach. The estimation proceeds backward to find the optimal decision rule at each stage to maximize

the cumulative rewards over the subsequent time. It estimates the optimal decision rule at future stage first, and then estimates the optimal decision rule at current stage by restricting the analysis to the subjects who have followed the estimated optimal decision rules thereafter. The benefit of this method is that it is highly flexible to mitigate the risk of model misspecification. SOWL is built on a simultaneous optimization method, which utilizes the whole dataset for estimating each treatment assignment. SOWL converts estimation of an optimal DTR into a single classification problem. This is the first time that learning multi-stage decision rules is performed simultaneously and integrated into a single algorithm. Current algorithms from support vector machines are adjusted and further developed for SOWL.

Compared with indirect methods, direct estimation methods usually employ non-parametric or semi-parametric estimators of $EY^*(d)$, requiring only mild assumptions about the data distribution and as a result are more robust to model misspecification. However, one potential drawback is the high variability of the value function estimates, which result in higher variance than indirect estimation methods.

Estimating optimal DTRs is an extremely active area of research. In addition to theoretical development discussed above, more tools for estimation and inference are becoming available and are continually being improved. Q-learning is accessible via the R package qLearn [93]. More recently, the alternative "interactive Q-learning" approach has become available via iqLearn package [94]. Two regression-based approaches: G-estimation can be implemented through R package DTRreg. Some more complex approaches (e.g. BOWL, value search methods based on IPTW and AIPTW) can be implemented with R package DynTxRegime [95].

The increasing availability of longitudinal observational studies has also provided new opportunities for the estimation of optimal DTRs. A lot of researchers have utilized real-world data such as cohort studies, EHRs, and clinical registries to estimate optimal DTRs. Examples include using the Center for International Blood and Marrow Transplant Research (CIBMTR) registry database for sequential prevention and treatment of acute graft-versus-host disease (GVHD) [96, 97], and using electronic medical record data such as MIMIC-III (Medical Information Mart for Intensive Care version III), MIMIC-IV (version IV) Clinical Database [98, 99]. MIMIC are openly available datasets on PhysioNet, comprising of detailed information regarding the care of real patients [100, 101]. Statistical methods adopted included Q-learning [97, 102], regret regression [80, 103], G-formula [104], marginal structural models [82], and other machine learning methods (e.g., adaptive contrast-weighted learning [105], deep reinforcement learning [96], stochastic tree-based reinforcement learning (ST-RL) method [106]). The application have covered different medical areas, including but not limited to HIV/AIDS, cancer, diabetes, and psychiatric disorders [107].

4.3 Discussion

The area of DTRs is a growing field, and there are many topics which are only beginning to be explored. In this section, we raise a few topics for discussion.

4.3.1 Alternative Outcome Types

Most DTR application has focused on continuous outcomes (e.g., symptom scores, CD4 count), however, research and analyses have been conducted for more complicated outcome types. In this subsection, we will briefly discuss the development of methods for alternative outcome types, including time-to-event outcomes and discrete outcomes.

In the cancer and other chronic disease research, the outcome of interest is usually time-to-event, for example, overall survival time or disease-free or progression-free survival time. Censoring of the time-to-event outcomes raises challenges to the estimation of optimal DTRs using the standard methods for non-censored outcomes. Several methods have been proposed regarding the estimation of optimal dynamic treatment regimes for survival outcomes. In the context of Q-learning, Huang et al. [108] used linear regression to fit accelerated failure time (AFT) model to estimate the optimal DRT in a time-to-event setting for a two-stage problem, with censoring handled by inverse probability weighting. Simoneau et al. [109] proposed a doubly robust method, named dynamic-weighted survival modeling, for estimating optimal DTRs for survival outcomes with right-censoring. They extended the dynamic weighted ordinary least square regression in [110] for non-censored outcomes. However, this method has some limitation as it is restricted to binary treatment. Recently, Cho et al. [111] proposed a reinforcement learning method to address the limitation of existing methods for survival outcomes, which allows a flexible number of treatment stages and arms. The estimator maximizes either the mean survival time or the survival probability using a generalized random survival forest-based algorithm. Simoneau et al. [109] could be implemented using R package DTRreg, and Cho et al. [111] could be implemented using R package dtrSurv.

The application of optimal DTR also involves discrete outcomes. The discrete outcome could be, for example, an indicator of no myocardial infarction within 30 days (a binary outcome) or the number of emergency room visits in a given period (a count, possibly Poisson-distributed). When the outcome is discrete, the estimation should take into account the constraints of the outcomes. Q-learning is appealing in terms of computational and conceptual simplicity, but it is mainly considered for continuous outcomes. Moodie et al. [112] proposed a Q-learning framework using generalized additive model (GAM) with penalized regression splines selected via generalized cross-validation (GCV) for developing optimal DTRs when outcomes are discrete (Bernoulli outcome and Poisson outcome, respectively). This approach added flexibility to the Q-learning procedure. Based on the simulation results,

GAM adapted Q-learning have superior performance compared with Q-learning with linear models and other methods based on propensity scores in terms of bias, MSE, and coverage.

4.3.2 Personalized Dose Finding

Dose finding is an important topic for personalized medicine. How to estimate the optimal dosage regime within multi-stage is a challenging problem. Extending the existing methods for handling finite treatment options to personalized dose finding with continuous doses might be infeasible as there could be an infinite number of treatment options for a given dose interval. Unlike the finite multiple-treatment DTR problem, only a few patients may be observed using the given dose level as the dose level follows a continuous distribution, so that the probability of observing a dose equal to the rule-specified dose is zero [113].

Chen et al. [113] proposed a robust outcome-weighted learning method based on a nonconvex loss function to find the optimal individualized dose rule. This is an extension of OWL method proposed by Zhao et al. [39]. With this approach, the dose finding problem is converted to a weighted regression with individual rewards as weights. This method has advantages over regression-based methods through the direct estimation of the optimal dose. However, like its originated method, this approach is prone to retain the actual observed dose, because only an observation in which the observed dose is close to the estimated optimal dose can contribute to the loss function.

5 Conclusions

Modern healthcare systems and drug development demand personalized medicine to improve patients' care and productivity within financial constrains. Personalized medicine has been more and more utilized to guide the drug development. Building upon the evolution of sciences and better understanding of disease, mechanism of action, appropriate targeted population can be hypothesized earlier in drug development. PM as a data-driven approach can also generate unique insights to quantify benefit-risk profiles across a heterogeneous patient population. For example, a subset of enrolled patients in Phase II studies with promising product profile can be carried into Phase III for confirmation.

The counterfactual outcome under the causal inference framework, specifically the conditional average treatment effect, is a central measure in recommending personalized treatment option. In practice, assessment of underlying assumptions of causal inference needs to follow the recommendation detailed in other chapters of this book. In this chapter, we also introduce two advanced analytic methods around the underlying assumptions of causal inference, specifically the algorithm OverRule

on identifying where the positivity assumption holds and the deconfounder method that relaxes the no unobserved confounder assumption.

This chapter focused on leveraging the widely available real-world data sources for PM research. Such data sources typically capture large number of patients over an extensive period of time. Specifically, the longitudinal cohort design is most valuable. It records the sequences of treatments for a heterogeneous population in routine clinical practice and provides unique opportunity for conducting PM research. To better analyze linked data sources where different data elements are captured, we report a few advanced analytic methods from literature developed to address such challenges. The importance and value of natural language processing to incorporate unstructured data are also highlighted to augment structured data.

Focusing on the treatment recommendation at one static time or longitudinally, Sects. 3 and 4 review the latest development in these areas. The existing methods are classified into different groups with a focus on their strengths, limitations, and general usage. Together with the availability of associated tools and examples of their applications, we believe this will further facilitate their adoption in conducting PM research.

In the meantime, no one methodology outperforms alternatives across all scenarios. Each individual personalized medicine research project warrants careful assessment of fit-for-purpose data source, study design, and analytic framework. As Type I error control has not traditionally been the focus of personalized medicine, the robustness of findings is of paramount importance. Some good practices include carefully assessing assumptions underlying different methodologies, employing alternative analysis under different set of assumptions, incorporating clinical judgments, and verifying the conclusions in independent data sources. Overall, personalized medicine is a fast-growing research field where great advancements have been made. With better awareness and further methodology research, personalized medicine will see more applications and have bigger impact in the clinical practice.

References

1. Council E. Council conclusions on personalised medicine for patients. Off J Eur Union [Internet] 2015;431:1–4.
2. Gamble P, Jaroensri R, Wang H, Tan F, Moran M, Brown T, et al. Determining breast cancer biomarker status and associated morphological features using deep learning. Communications Medicine 2021;1:1–12.
3. Cruz-Ramos M, García-Foncillas J. CAR-T cell and Personalized Medicine. Translational Research and Onco-Omics Applications in the Era of Cancer Personal Genomics 2019:131–45.
4. Srivastava S, Riddell SR. Chimeric antigen receptor T cell therapy: challenges to bench-to-bedside efficacy. The Journal of Immunology 2018;200:459–68.
5. ICH E9 (R1) 2021 Statistical Principles for Clinical Trials: Addendum: Es-timands and Sensitivity Analysis in Clinical Trials.

6. Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association 2005;100:322–31.
7. Bica I, Alaa AM, Lambert C, Van Der Schaar M. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. Clinical Pharmacology & Therapeutics 2021;109:87–100.
8. Oberst M, Johansson F, Wei D, Gao T, Brat G, Sontag D, et al. Characterization of overlap in observational studies. International Conference on Artificial Intelligence and Statistics, PMLR; 2020, p. 788–98.
9. Wang Y, Blei DM. The blessings of multiple causes. Journal of the American Statistical Association 2019;114:1574–96.
10. Lavori PW, Dawson R. A design for testing clinical strategies: biased adaptive within-subject randomization. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2000;163:29–38.
11. Lavori PW, Dawson R. Dynamic treatment regimes: practical design considerations. Clinical Trials 2004;1:9–20.
12. Murphy SA. An experimental design for the development of adaptive treatment strategies. Statistics in Medicine 2005;24:1455–81.
13. Chatterjee N, Chen Y-H, Maas P, Carroll RJ. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. Journal of the American Statistical Association 2016;111:107–17.
14. Yang S, Ding P. Combining multiple observational data sources to estimate causal effects. Journal of the American Statistical Association 2019.
15. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv Preprint ArXiv:181004805 2018.
16. Sun Y, Jang J, Huang X, Wang H, and He W. Leveraging Free Text Data for Decision Making in Drug Development. JSM 2019 Online Program https://ww2.amstat.org/meetings/jsm/2019/onlineprogram/AbstractDetails.cfm?abstractid=305116.
17. Woodcock J. The prospects for "personalized medicine" in drug development and drug therapy. Clinical Pharmacology & Therapeutics 2007;81:164–9.
18. Wijn SR, Rovers MM, Le LH, Belias M, Hoogland J, IntHout J, et al. Guidance from key organisations on exploring, confirming and interpreting subgroup effects of medical treatments: a scoping review. BMJ Open 2019;9:e028751.
19. Su X, Meneses K, McNees P, Johnson WO. Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. Journal of the Royal Statistical Society: Series C (Applied Statistics) 2011;60:457–74.
20. Su X, Kang J, Fan J, Levine RA, Yan X. Facilitating score and causal inference trees for large observational studies. Journal of Machine Learning Research 2012;13:2955.
21. Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. Statistics in Medicine 2011;30:2867–80.
22. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling 1986;7:1393–512.
23. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research 2011;46:399–424.
24. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. American Journal of Epidemiology 2011;173:761–7.
25. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. 2nd ed. Springer; 2021.
26. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences 2019;116:4156–65.
27. Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. The Annals of Applied Statistics 2013;7:443–70.

28. Lu M, Sadiq S, Feaster DJ, Ishwaran H. Estimating individual treatment effect in observational data using random forest methods. Journal of Computational and Graphical Statistics 2018;27:209–19.
29. Hill JL. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics 2011;20:217–40.
30. Sugasawa S, Noma H. Estimating individual treatment effects by gradient boosting trees. Statistics in Medicine 2019;38:5146–59.
31. Hahn PR, Murray JS, Carvalho CM. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). Bayesian Analysis 2020;15:965–1056.
32. Robinson PM. Root-N-consistent semiparametric regression. Econometrica: Journal of the Econometric Society 1988:931–54.
33. Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. Biometrika 2021;108:299–319.
34. Tibshirani J, Athey S, Friedberg R, Hadad V, Hirshberg D, Miner L, et al. Package 'grf' 2022.
35. Athey S, Wager S. Estimating treatment effects with causal forests: An application. Observational Studies 2019;5:37–51.
36. Tian L, Alizadeh AA, Gentles AJ, Tibshirani R. A simple method for estimating interactions between a treatment and a large number of covariates. Journal of the American Statistical Association 2014;109:1517–32.
37. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 2016;113:7353–60.
38. Hitsch GJ, Misra S. Heterogeneous treatment effects and optimal targeting policy evaluation. Available at SSRN 3111957 2018.
39. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association 2012;107:1106–18.
40. Gu X, Yin G, Lee JJ. Bayesian two-step Lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. Contemporary Clinical Trials 2013;36:642–50.
41. Schnell PM, Tang Q, Offen WW, Carlin BP. A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. Biometrics 2016;72:1026–36.
42. Ngo D, Baumgartner R, Mt-Isa S, Feng D, Chen J, Schnell P. Bayesian credible subgroup identification for treatment effectiveness in time-to-event data. Plos One 2020;15:e0229336.
43. Quartey, Daniel, Schnell, Patrick, Baumgartner R, Mt-Isa S, Feng D, Chen J, et al. Bayesian credible subgroup for count data with excess zeroes. Under Review 2022.
44. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. The Annals of Applied Statistics 2010;4:266–98.
45. Henderson NC, Louis TA, Rosner GL, Varadhan R. Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. Biostatistics 2020;21:50–68.
46. Su X, Tsai C-L, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. Journal of Machine Learning Research 2009;10.
47. Su X, Zhou T, Yan X, Fan J, Yang S. Interaction trees with censored survival data. The International Journal of Biostatistics 2008;4.
48. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. Statistics in Medicine 2011;30:2601–21.
49. Loh W-Y, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. Statistics in Medicine 2015;34:1818–33.
50. Loh W-Y. Regression tress with unbiased variable selection and interaction detection. Statistica Sinica 2002:361–86.
51. Huang X, Sun Y, Trow P, Chatterjee S, Chakravartty A, Tian L, et al. Patient subgroup identification for clinical drug development. Statistics in Medicine 2017;36:1414–28.

52. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics 2012;68:1010–8.
53. Chen S, Tian L, Cai T, Yu M. A general statistical framework for subgroup identification and comparative treatment scoring. Biometrics 2017;73:1199–209.
54. Ciarleglio A, Petkova E, Ogden RT, Tarpey T. Treatment decisions based on scalar and functional baseline covariates. Biometrics 2015;71:884–94.
55. Luo H, Nattino G, Pratola MT. Sparse Additive Gaussian Process Regression. Journal of Machine Learning Research 2022;23:1–34.
56. Zhao L, Feng D. Deep neural networks for survival analysis using pseudo values. IEEE Journal of Biomedical and Health Informatics 2020;24:3308–14.
57. Van der Laan MJ, Polley EC, Hubbard AE. Super learner. Statistical Applications in Genetics and Molecular Biology 2007;6.
58. Zhao L, Tian L, Cai T, Claggett B, Wei L-J. Effectively selecting a target population for a future comparative study. Journal of the American Statistical Association 2013;108:527–39.
59. Huang X, Li H, Gu Y, Chan IS. Predictive Biomarker Identification for Biopharmaceutical Development. Statistics in Biopharmaceutical Research 2021;13:239–47.
60. Kuhn M, Johnson K. Applied predictive modeling. vol. 26. Springer; 2013.
61. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The Annals of Applied Statistics 2008;2:841–60.
62. Kapelner A, Bleich J. Prediction with missing data via Bayesian additive regression trees. Canadian Journal of Statistics 2015;43:224–39.
63. Chen T, Guestrin C. Xgboost: A scalable tree boosting system, 2016, p. 785–94.
64. Guo X, He X. Inference on selected subgroups in clinical trials. Journal of the American Statistical Association 2021;116:1498–506.
65. Bornkamp B, Ohlssen D, Magnusson BP, Schmidli H. Model averaging for treatment effect estimation in subgroups. Pharmaceutical Statistics 2017;16:133–42.
66. Athey S, Tibshirani J, Wager S. Generalized random forests. The Annals of Statistics 2019;47:1148–78.
67. Chakraborty B, Murphy SA. Dynamic treatment regimes. Annual Review of Statistics and Its Application 2014;1:447–64.
68. Murphy SA. Optimal dynamic treatment regimes. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2003;65:331–55.
69. Tsiatis AA, Davidian M, Holloway ST, Laber EB. Dynamic treatment regimes: Statistical methods for precision medicine. Chapman and Hall/CRC; 2019.
70. Laber EB, Lizotte DJ, Qian M, Pelham WE, Murphy SA. Dynamic treatment regimes: Technical challenges and applications. Electronic Journal of Statistics 2014;8:1225.
71. Barto AG. 2 Reinforcement Learning and Its. Handbook of Learning and Approximate Dynamic Programming 2004;2:47.
72. Chakraborty B, Moodie EE. Statistical methods for dynamic treatment regimes. Springer-Verlag 2013;10:978–1.
73. Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. Health Service Research Methodology: A Focus on AIDS 1989:113–59.
74. Robins JM, Berkane M. Latent variable modeling and applications to causality. Causal Inference from Complex Longitudinal Data 1997:69–117.
75. Robins JM. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. Proceedings of the Biopharmaceutical Section, American Statistical Association, vol. 24, San Francisco CA; 1993, p. 3.
76. Murphy SA. A generalization error for Q-learning 2005.
77. Laber EB, Linn KA, Stefanski LA. Interactive model building for Q-learning. Biometrika 2014;101:831–47.
78. Schulte PJ, Tsiatis AA, Laber EB, Davidian M. Q-and A-learning methods for estimating optimal dynamic treatment regimes. Statistical Science: A Review Journal of the Institute of Mathematical Statistics 2014;29:640.

79. Robins JM. Optimal structural nested models for optimal sequential decisions. Proceedings of the second seattle Symposium in Biostatistics, Springer; 2004, p. 189–326.
80. Henderson R, Ansell P, Alshibani D. Regret-regression for optimal dynamic treatment regimes. Biometrics 2010;66:1192–201.
81. Zhao Y-Q, Laber EB. Estimation of optimal dynamic treatment regimes. Clinical Trials 2014;11:400–7.
82. Orellana L, Rotnitzky A, Robins JM. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part I: main content. The International Journal of Biostatistics 2010;6.
83. Robins J, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. Statistics in Medicine 2008;27:4678–721.
84. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. vol. 11. Lww; 2000.
85. Nahum-Shani I, Qian M, Almirall D, Pelham WE, Gnagy B, Fabiano GA, et al. Q-learning: a data analysis method for constructing adaptive interventions. Psychological Methods 2012;17:478.
86. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. Communications in Statistics-Theory and Methods 1994;23:2379–412.
87. Vansteelandt S, Goetghebeur E. Causal inference with generalized structural mean models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2003;65:817–35.
88. Almirall D, Ten Have T, Murphy SA. Structural nested mean models for assessing time-varying effect moderation. Biometrics 2010;66:131–9.
89. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. Statistical models in epidemiology, the environment, and clinical trials, Springer; 2000, p. 95–133.
92. Hernán MÁ, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 2000:561–70.
91. Robins JM. Marginal structural models. 1997 proceedings of the American Statistical Association, section on Bayesian statistical science (pp. 1–10). Retrieved From 1998.
92. Zhao Y-Q, Zeng D, Laber EB, Kosorok MR. New statistical learning methods for estimating optimal dynamic treatment regimes. Journal of the American Statistical Association 2015;110:583–98.
93. Xin J, Chakraborty B, Laber EB. qLearn: Estimation and inference for Q-learning. R Package Version 2012;1:87.
94. Linn KA, Laber EB, Stefanski LA. iqLearn: Interactive Q-learning in R. Journal of Statistical Software 2015;64.
95. Holloway ST, Laber EB, Linn KA, Zhang B, Davidian M, Tsiatis AA. Dyn-TxRegime: methods for estimating optimal dynamic treatment regimes, 2019. R Package Version;4.
96. Liu N, Liu Y, Logan B, Xu Z, Tang J, Wang Y. Learning the dynamic treatment regimes from medical registry data through deep Q-network. Scientific Reports 2019;9:1–10.
97. Krakow EF, Hemmer M, Wang T, Logan B, Arora M, Spellman S, et al. Tools for the precision medicine era: how to develop highly personalized treatment recommendations from cohort and registry data using Q-learning. American Journal of Epidemiology 2017;186:160–72.
98. Raghu A, Komorowski M, Celi LA, Szolovits P, Ghassemi M. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. Machine Learning for Healthcare Conference, PMLR; 2017, p. 147–63.
99. Laha N, Sonabend-W A, Mukherjee R, Cai T. Finding the Optimal Dynamic Treatment Regime Using Smooth Fisher Consistent Surrogate Loss. ArXiv Preprint ArXiv:211102826 2021.
100. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data 2016;3:1–9.
101. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2021). MIMIC-IV (version 1.0). PhysioNet. https://doi.org/10.13026/s6n6-xd98.

102. Moodie EE, Chakraborty B, Kramer MS. Q-learning for estimating optimal dynamic treatment rules from observational data. Canadian Journal of Statistics 2012;40:629–45.
103. Rosthøj S, Fullwood C, Henderson R, Stewart S. Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach. Statistics in Medicine 2006;25:4197–215.
104. Young JG, Cain LE, Robins JM, O'Reilly EJ, Hernán MA. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. Statistics in Biosciences 2011;3:119–43.
105. Tao Y, Wang L. Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. Biometrics 2017;73:145–55.
106. Sun Y, Wang L. Stochastic tree search for estimating optimal dynamic treatment regimes. Journal of the American Statistical Association 2021;116:421–32.
107. Mahar RK, McGuinness MB, Chakraborty B, Carlin JB, IJzerman MJ, Simpson JA. A scoping review of studies using observational data to optimise dynamic treatment regimens. BMC Medical Research Methodology 2021;21:1–13.
108. Huang X, Ning J, Wahed AS. Optimization of individualized dynamic treatment regimes for recurrent diseases. Statistics in Medicine 2014;33:2363–78.
109. Simoneau G, Moodie EE, Nijjar JS, Platt RW, Investigators SERAIC. Estimating optimal dynamic treatment regimes with survival outcomes. Journal of the American Statistical Association 2020;115:1531–9.
110. Wallace MP, Moodie EE. Doubly-robust dynamic treatment regimen estimation via weighted least squares. Biometrics 2015;71:636–44.
111. Cho H, Holloway ST, Kosorok MR. Multi-stage optimal dynamic treatment regimes for survival outcomes with dependent censoring. ArXiv Preprint ArXiv:201203294 2020.
112. Moodie EE, Dean N, Sun YR. Q-learning: Flexible learning about useful utilities. Statistics in Biosciences 2014;6:223–43.
113. Chen G, Zeng D, Kosorok MR. Personalized dose finding using outcome weighted learning. Journal of the American Statistical Association 2016;111:1509–21.

Use of Real-World Evidence in Health Technology Assessment Submissions

Yingyi Liu and Julia Ma

1 Introduction

Health technology assessment (HTA) is a process of investigating the medical, economic, organizational, social, and ethical impacts of health technologies. The evaluation can help determine the value of a health technology or medical intervention and whether it should be made available in a health system. Guidance on how the technology can be used in health systems will be developed and published following the decision-making. HTA is a systematic and multidisciplinary process as it requires a thorough assessment of every aspect of new technologies and their impact on a healthcare system. Professionals and researchers from a range of disciplines work together using explicit analytical frameworks drawn from a variety of methodologies.

Technology assessment should be done in a transparent and unbiased manner, with the major purpose to provide the best available research-based evidence and to inform technology-related healthcare policy decision-making. HTA seeks to provide health policymakers with accessible, useable, and evidence-based information to guide their decisions about the appropriate use of new and existing technologies and efficient allocation of resources. Therefore, HTA is often described as "the bridge between evidence and policy making" [1].

HTA has a strong foundation in research on the health effects and broad implications of the use of technology in health care. Its potential for contributing to safer and more effective health care is widely acknowledged. Most countries have a formal process for collecting and reporting scientific evidence to support healthcare policy decision-making. Formal HTA has become a prerequisite for

Y. Liu · J. Ma (✉)
Medical Affairs and Health Technology Assessment Statistics, Data and Statistical Sciences, AbbVie, North Chicago, IL, USA
e-mail: julia.y.ma@abbvie.com

W. He et al. (eds.), *Real-World Evidence in Medical Product Development* ,
https://doi.org/10.1007/978-3-031-26328-6_17

access to reimbursement in key markets such as the United Kingdom (UK), Canada, Australia, and France. HTA is playing an increasingly important role in informing reimbursement and pricing decisions and providing clinical guidance on the use of medical technologies around the world.

A wide range of use cases of RWE can be found along different health technology lifecycles. RWE from sources such as observational studies, registry, health surveys, and electric health records (EHRs) have been submitted to HTA bodies to identify treatment patterns, to characterize patients, and to provide supportive evidence for the economic evaluations. Such evidence offers broad and rich information about what really happens in routine clinical practice. Depending on the market prototype, HTA appraisals appear to focus mainly on safety and clinical effectiveness, followed by the economic and budgetary impacts of health technology. Evidence collected from real-world clinical practice provides valuable information to fill the efficacy–effectiveness gap which is the limitation of evidence from randomized clinical trials (RCTs) [2].

The remainder of the chapter is organized as follows. Types of RWE included in technology assessment, the role of RWE in market access and reimbursement, and the impact of RWE on decision-making are discussed in Sect. 2. The strength of RWE in addressing the specific needs of key healthcare stakeholders are recognized and the value of RWE in HTA decision-making is surveyed in Sect. 3, together with some recent RWE usage examples for HTA purpose. While more and more HTA bodies are becoming more open to and more comfortable with RWE, practice varies among the agencies and guidance development is also at different stages. An overview of the most recent trends in the use of RWE for technology assessments, focusing on several influential HTA bodies, such as the National Institute for Health and Care Excellence (NICE) in the UK is provided in Sect. 4 prior to the conclusions concerning the future of RWE in HTA submissions and decision-making.

2 Role of RWE in HTA Submissions

2.1 Data Sources and Types of RWE

Based on the definition of the Food and Drug Administration (FDA), real-world data (RWD) are defined as data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources. RWE is the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from the analysis of RWD [3].

The most common sources of RWD include EHRs, administrative data, claims data, patient registries, patient-generated health data, chart reviews, patient survey data, healthcare provider survey data, and other observational data [4]. The most frequently utilized RWE study designs include retrospective study, prospective study, cross-sectional study, and pragmatic clinical trials.

2.2 *Acceptability of RWE Across HTA Agencies*

The growing interest in incorporating RWE into the HTA process to assess the clinical relative effectiveness and economic value of drugs is mainly due to the limited availability of evidence from RCTs and the desirable property of RWE in reflecting outcomes in real-world settings. HTA agencies worldwide are increasingly accepting RWE as part of the information they use to guide their decisions. In this subsection, we review the acceptability and requirements for the use of RWE by several influential HTA agencies around the world.

Makady et al. [5] investigated policies on the use of RWE for initial drug submissions to six European HTA agencies, namely AIFA, HAS, IQWiG, NICE, TLV, and ZIN (the full names of HTA agencies are provided in Table 1). All HTA agencies accept all available evidence for initial drug assessments, including RWE. Some agencies (NICE, IQWiG, ZIN) also provided advice for specific RWD sources as well as guidance on the suitability of these sources [5]. RWD may be used to demonstrate treatment effects only under specific circumstances. For example, RWD may be used when lack of RCT data on drug efficacy (NICE, IQWiG, ZIN); RWD may be utilized to inform indirect treatment comparisons (NICE, ZIN) when head-to-head RCTs are not available; or RWD can be used to supplement RCT data when there is no available data on specific subpopulations or long-term follow-up (NICE, ZIN) [5]. Under these scenarios, the agencies required an explicit justification of why RWD were used and a clear discussion of the biases associated with the RWD and their consequences on treatment effect estimates (TLV, NICE, IQWiG, HAS, ZIN). Makady et al. [5] found that all agencies adopted similar

Table 1 List of selected HTA agencies in the world

Country	HTA agency	HTA agency full name
Australia	PBAC	Pharmaceutical Benefits Advisory Committee
Canada	CADTH	Canadian Agency for Drugs and Technologies in Health
Canada	INESSS	Institut National d'Excellence en santé et en Services Sociaux
Europe	EUnetHTA	European Network for Health Technology Assessment
Finland	PPB	Pharmaceuticals Pricing Board
France	HAS	Haute Autorité de Santé
Germany	IQWiG	Institute for Quality and Efficiency in Healthcare
Germany	G-BA	The Federal Joint Committee
Italy	AIFA	Italian Medicines Agency
Netherlands	ZIN	Zorginstituut Nederland
New Zealand	PHARMAC	Pharmaceutical Management Agency
Scotland	SMC	Scottish Medicines Consortium
Sweden	TLV	Dental and Pharmaceutical Benefits Agency
UK	NICE	National Institute for Health and Care Excellence
United States (US)	ICER	Institute for Clinical & Economic Review

evidence hierarchies that placed RWD on a lower level in terms of quality and reliability. RWD can be used to confirm or supplement RCTs, but not to replace them. Therefore, conclusions on treatment effects that were based on RWE would be considered with greater caution than those based on evidence from RCTs.

Another study published by CADTH in 2018 [6] reviewed the acceptability and requirement of use of RWE by key HTA agencies around the world, namely CADTH, EUnetHTA, HAS, INESSS, IQWiG, TLV, NICE, PBAC, PPB, PHARMAC, SMC, and ZIN. Generally, the agencies accepted both randomized and non-randomized clinical data as part of the initial drug submission, but the requirements vary substantially. IQWiG in Germany had strict evidence hierarchies that value RCT data over RWE. IQWiG required that conclusions for benefit assessments are usually inferred only from the results of direct comparative studies. RCTs are required to demonstrate causality; other study designs mostly cannot answer required questions with sufficient certainty due to potential biases. The use of non-randomized data for benefit assessment requires particular justification or specific preconditions and special demands on quality [6]. UK NICE preferred head-to-head RCTs, but also accepts non-RCT studies as supplementary when head-to-head RCT data is not available or insufficient. CADTH accepted all study types, while data from one or more RCTs are preferred. Non-RCTs may be particularly useful when long-term follow-up evaluation is required, if RCT is impractical due to limited number of patients or for ethical reasons, if RCT data lack relevant comparators or when RCTs have limited external validity [6]. HAS accepted studies according to the evidence hierarchy; meta-analysis of good methodological quality; clinical trial, or observational study design; and implementation according to current methodological requirements. Resubmissions are the same as initial submissions or extension of indications [6]. Regarding additional evidence for safety, most agencies requested additional non-RCT safety data, such as periodic safety reports (PSURs) or other pharmacovigilance data (EUnetHTA, HAS, IQWiG, PBAC, and PPB) [6]. A comprehensive summary can be found in this document [6].

With regard to rare diseases, the policies on the use of RWE may vary. For rare diseases, evidence available at the time of reimbursement decision-making may have a higher degree of uncertainty. Studies may be too short for assessing long-term outcomes, or may lack important patient-relevant outcomes, or have no standard of care comparison [7]. In early 2020, G-BA passed a law in Germany to mandate the collection of post-launch RWE for Advanced Therapy Medicinal Products (ATMPs). IQWiG followed up with a G-BA commissioned report that high-quality registries can be used to conduct added benefit assessments for new drugs, especially in scenarios where there is limited evidence available at the time of market authorization. However, IQWiG considered other RWE sources such as EHRs and claims data far less promising due to concerns with data quality and completeness [7]. On February 2021, the G-BA announced the plan for the first mandatory collection of RWE for Zolgensma, a gene therapy approved for the treatment of spinal muscular atrophy (SMA). Novartis will run a registry study to collect RWE [8]. While Germany authorities are well-known as less willing to use RWE to inform HTA decision-making and have very high standard for RWE, these

recent advancements have demonstrated a need for RWE to inform reimbursement decision when there is unmet need and lack of RCTs data.

Currently, the acceptability and requirement of RWE among HTA bodies varies greatly. Generally, there has been a trend toward more openness to RWE for most HTA agencies in the last few years.

2.3 Role of RWE in Market Access and Reimbursement

In recent years, HTA agencies are increasingly turning to RWE to enrich their evidence base for decision-making. Several studies have summarized the use of RWE in HTA decisions across different HTA agencies.

In a subsequent study, Markady et al. [9] reviewed appraisals by five HTA agencies in Europe (NICE, SMC, HAS, IQWiG, and ZIN) within the context of treatments for patients with melanoma (ipilimumab, vemurafenib, dabrafenib, cobimetinib, trametinib, nivolumab, pembrolizumab). This review included 52 HTA reports published on agency website between 2011 and 2016 (a full list of report can be found in supplementary material Appendix 4 of Markady's paper [9]), of which 28 (54%) included RWD. The majority of appraisals included in this study were reassessment reports. For relative effectiveness assessment (REA), RWD were primarily used to estimate the prevalence of melanoma in all 28 reports, and the majority of RWD used to estimate melanoma prevalence came from registries. In addition, RWD were used to estimate effectiveness, and the RWD included for effectiveness were mainly derived from observational studies and non-randomized phase I/II studies [9]. For cost-effectiveness assessment (CEA), RWD were mainly used to extrapolate long-term effectiveness, identify drug-related costs and medical offset, and estimate utilities using quality-of-life information. Long-term effectiveness data usually came from registries and national statistics databases. Costs were estimated using data from claims databases, observational studies, or cost-of-illness studies.

Markady et al. [9] noticed differences in the use of RWE across all HTA agencies. Among all the REAs, all NICE and ZIN reports included RWD, whereas RWD were included in 23% of SMC reports, 62% of HAS reports, and 53% of IQWiG. IQWiG and ZIN did not use RWD to inform safety or efficacy in any report. SMC used RWD for safety in 6% of cases and in effectiveness for 12% of cases. HAS used RWD for safety and effectiveness in 9% of cases, respectively. NICE used RWD for safety and effectiveness in 22% of cases, respectively.

The economic modeling for HTA submission usually includes CEA and budget impact analysis (BIA). CEA is a type of economic evaluation that compares the costs and outcomes of two or more interventions with a common health outcome but different effectiveness [10]. For the cost-effectiveness models, the key components include population characteristics, relative efficacy and safety, disease model for long-term outcome, utility, and costs. BIA is an economic assessment that estimates the expected extra budget and cost-offsets following the introduction of a new

Table 2 Frequency and distribution of types of inputs informed by RWE

Input	Category	Number of RWE uses	Percentage (%)
Non-drug-specific clinical inputs	Disease progression/mortality rates	117	28.7
	Prevalence	44	10.8
	Patient characteristics	23	5.7
	Other clinical inputs	14	3.4
	Incidence	10	2.5
	Non-drug-specific discontinuation rate	2	0.5
Drug-specific clinical inputs	Effectiveness	6	1.5
	Drug-specific discontinuation rate	5	1.2
	Adverse drug event rates	2	0.5
Utility inputs	Non-drug-specific utility	38	9.3
	Drug-specific utility	1	0.2
Economic inputs	Healthcare costs	86	21.1
	Non-healthcare costs	27	6.6
	Treatment pattern/market share	13	3.2
Assumptions	RWE used to support assumptions	19	4.7

Source : Lee et al. [13]

healthcare technology [11]. The key components for BIA include size of patient population, market share, usage pattern, and drug costs. RWE can be used to validate assumptions in these models and be a valuable source for inputs such as event rate, population characteristics, utility, and treatment costs.

Lee et al. [12] reviewed CEA and BIA in 33 pharmaceuticals reports published by ICER between January 2014 and June 2019. The 33 reports considered a total of 123 pharmaceutical interventions and comparators for 29 diseases. All reports included a CEA, but two reports did not include a BIA. In the 33 ICER reports, 407 RWE uses were identified in total. The description and categorization of model inputs informed by RWE are summarized in Table 2. The results show that RWE was most commonly used to inform non-drug-specific clinical inputs, such as disease progression/mortality rates (28.7%), prevalence (10.8%), and patient characteristics (5.7%). RWE was also widely used for economic inputs, including healthcare costs (21.1%), non-healthcare costs (6.6%), and treatment pattern/market share (3.2%). In addition, RWE was leveraged for utility inputs, mainly non-drug specific utility (9.3%). However, it was rarely used for drug-specific clinical inputs such as effectiveness (1.5%), discontinuation rates (1.2%), and adverse drug event rates (0.5%). The most frequently used study design was a retrospective cohort (207/407, 50.9%), and the most frequently used data source was registry data (163/407, 40.0%) (Tables 3 and 4). About a third (30.2%) of RWE was industry-sponsored.

Table 3 Frequency and distribution of study design of RWE

Study design	Percentage (%)
Retrospective cohort	50.9
Prospective cohort	17.0
Cross-sectional for surveys	12.8
Utility study	9.1
Cross-sectional	3.4
Meta-analysis/Network meta-analysis	2.2
Systematic review	1.7
Unidentifiable	2.9

Source : Lee et al. [13]

Table 4 Frequency and distribution of data source of RWE

Data source	Percentage (%)
Registry data	40.0
Administrative claims data	18.2
Patient survey data/patient diary	18.2
EHRs	9.6
Other observational data	9.3
Synthesis of multiple previous RWE studies	3.7
Healthcare provider survey data	1.0

Source : Lee et al. [13]

Table 5 Use of RWE in HTA Submissions to NICE during 2018–2021

	2018	2019	2020	2021	Total
RWE not included	36	8	6	11	61
Negative recommendation	6	1	1		8
Positive or positive with restrictions	30	7	4	10	51
No recommendation			1	1	2
RWE included	20	43	37	62	162
Negative recommendation	2	3	2	6	13
Positive or positive with restrictions	18	40	35	56	149
Total	56	51	43	73	223
Proportion of total supported by RWE	35.7%	84.3%	86.0%	84.9%	72.6%

Source : Segwagwe et al. [14]

More recent research by Segwagwe et al. [14] investigated the HTA submissions to NICE between 2018 and 2021 utilizing IQVIA's proprietary HTA-Accelerator tool to quantify the use of RWE in HTA submissions. Among the 223 submissions analyzed, 36% in 2018 included RWE, while from 2019 to 2021 the proportion were 84%, 86% and 85% respectively. During these years, 92% of the submissions including RWE made positive recommendations, while 84% of those excluding RWE made positive recommendations. More details can be found in Table 5. These findings suggest that there was an increasing trend of leveraging RWE for HTA submissions.

To sum up, HTA agencies are increasingly accepting RWE to inform the market access and reimbursement decisions. RWE can provide valuable information for various uses in CEAs and economic modeling. The principal uses of RWE for CEA mainly include:

- Estimating the burden (e.g., prevalence and incidence) of disease
- Understanding local treatment pathways
- Determining comparative effectiveness and safety in the real world vs. relevant comparators in the absence of RCT data
- Determining comparative effectiveness in relevant patient populations
- Demonstrating long-term effectiveness and safety

The principal use of RWE in economic modeling are as follows:

- Providing inputs for costs and healthcare resource utilization (HRU)
- Collecting data on quality of life and utility
- Providing model inputs such as transition probability

Overall, RWE has experienced an explosion of interest within the last decade. HTA agencies worldwide are currently exploring the possibilities of using RWE to supplement and enrich evidence. However, the use of RWE to inform drug-specific effectiveness and safety in many countries is still limited. One of the barriers to using RWE is the absence of drug-specific RWD available at the time of assessment, especially during the initial technology assessment. Since the assessment usually takes place soon after regulatory approval, there might be insufficient time to collect drug-specific RWD from registries or observational studies. At this stage, RWD are more likely to be utilized to provide insights on the natural history of disease, disease burden, and unmet medical needs. Another barrier could be the lack of guidance of the use and analysis of RWD for HTA purposes. We are expecting fast progress in guidance development in the future.

3 Value and Strength of RWE for HTA Purposes

3.1 Efficacy–Effectiveness Gap and Strength of RWE

Evidence on drug effectiveness informing HTA submissions conventionally relies on RCTs. RCTs remain the preferred source of evidence among HTA agencies. However, the potential shortcomings of RCTs include the following:

- They are conducted under tightly controlled conditions that often do not reflect the realities of treating patients in routine practice.
- The patient populations included in RCTs often do not reflect the general population in real-world clinical practice. They may exclude or under-present certain types of patient populations, such as pregnant women or women who are breastfeeding, children, and elderly patients [15–17].

– The treatment and follow-up periods are often limited, potentially preventing measurement of long-term health outcomes [18].
– HTA generally prefers hard clinical outcomes which may not be well captured in RCTs due to their rarity.
– HTA preferred active control may not be practical in RCTs.

Consequently, extrapolation of drug efficacy from RCTs to drug effectiveness in real-world clinical practice is difficult. Many RCTs have limited generalizability, and the conclusions drawn from RCTs may only be applicable to selected patients. This discrepancy is frequently referred to as the efficacy–effectiveness gap [2, 19, 20]. In the context of HTA decision-making, which focuses on the value of a drug, initial drug funding decisions are based on the clinical benefits observed from RCTs and the value of a drug is estimated using economic models such as willingness to pay for quality-adjusted life-year gain (QALY). The reliance on RCTs may leave a large efficacy–effectiveness gap, increasing the risk and uncertainty of HTA decisions [21].

In contrast, an inherent strength of RWE is its consideration of the unselected patient population, which may be more relevant to routine practice [20, 22]. The large diversity in inclusion and exclusion criteria provides information on treatments in patient groups that are usually excluded from RCTs [23]. Another advantage is that RWE is more feasible and far less expensive to gather than information from RCTs and can provide long-term longitudinal data on safety and effectiveness. Therefore, RWE constitutes a bridge between the evidence generated in RCTs and routine clinical practice to fill the evidence gap. RWE could be a potentially valuable source of evidence to provide complementary data that are important to decision makers.

Given the strength of RWE, RWE can add value throughout the HTA process. Before assessment, RWE can be used early in the scoping stage to identify disease burden on both patients and the healthcare system, explore treatment patterns, assess disease epidemiology, understand standard of care as a comparator for future analysis. During assessment, RWE can be used to facilitate the identification of relevant subpopulations, provide additional evidence concerning long-term treatment effectiveness and safety, permit the inclusion and analysis of clinical endpoints not included in RCTs but observed in real life, gather information on the safety and effectiveness when drugs are used more broadly in the real-world clinical setting, model cost-effectiveness, illustrate how the product will fit in the current clinical practice, and much more. After assessment, RWE can be used to understand real-world treatment patterns and risks of outcomes, demonstrate real-world benefits, and gain leverage in pricing negotiations. RWE studies that demonstrate effectiveness and acceptability in new patient groups can help inform coverage expansion decisions for broader access.

3.2 Case Studies

To understand the value RWE can bring to market access and reimbursement decision, we examine several case studies. One example is the use of RWE to support the HTA re-submission of ipilimumab for malignant melanoma in Australia [24] (the information of this case study can be found in the public reports provided in PBAC website [25–27]). In July 2011, the initial submission presented a single phase III, randomized double-blind trial. The median survival for the ipilimumab monotherapy arm was 10.12 months compared to 6.44 months in the control group [25]. The submission described ipilimumab as "superior in terms of comparative effectiveness" but the PBAC considered this claim may not be reasonable as ipilimumab may be considered "inferior in terms of immune related adverse events" [25]. Consequently, the PBAC did not recommend the drug because of an "uncertain extent of clinical benefit, uncertain clinical place of therapy, high and uncertain cost-effectiveness ratio and uncertain financial costs" [25]. In the first resubmission, no additional evidence was added but additional exploratory analyses relating to safety were presented. The company also proposed a reduced price but PBAC did not change its decision. In the second resubmission in November 2012, Bristol–Myers Squibb presented new evidence related to the durability of ipilimumab's effect, including three recent real-world post-registration data: the most recent Periodic Safety Update Report (PSUR), Italian "real-world" data relating to efficacy, safety, and rates of re-induction from the European Expanded Access Programme (EAP) and submitted to the European Society for Medical Oncology (ESMO) conference in October 2012, and Australian "real-world" data resulting from the Patient Access Program (which existed between August 1, 2011, and April 15, 2012) [27]. The PBAC agreed that the results from the EAP in Italy supported the results seen in the other trials. This new data demonstrated that the drug increased survival and confirmed the durability of the clinical effect. Finally, PBAC recommended ipilimumab with this new evidence and a decrease in price.

Another example is the HTA submission of aflibercept for treatment of adults with colorectal cancer to SMC in Scotland [24]. The initial submission contained one randomized, placebo-controlled phase III study. Results of the study demonstrated significantly longer overall survival. However, in June 2013, SMC did not recommend aflibercept because it lacked a sufficiently robust economic analysis. In the February 2014 resubmission, Sanofi included two on-going, open-label, single-arm studies to assess safety and quality of life. This RWE was used to revise the utility score within the economic model. In the resubmission, the QALY gain slightly increased, especially due to the two open-label studies providing a stable estimate of quality of life under the treatment [28]. The agency recommended aflibercept because this new data demonstrated a substantial improvement in quality of life in the patient population.

RWE can also be used to support reimbursement in a patient subgroup. One example is the submission of Brentuximab vedotin for treating CD30-positive Hodgkin lymphoma to NICE. To identify the high-risk group reflecting clinical

practice in England, the company provided additional evidence citing a retrospective study including multiple centers across England. The company redefined high-risk patient criteria and identified a subgroup of patients which were accepted by committee [29].

However, there are some uncertainties in the impact of RWE in reimbursement decision-making. Jao et al. [30] conducted a wider review to look beyond melanoma drugs across HTA agencies in seven markets (Germany, France, England, Scotland, Canada, Australia, and South Korea). They found that the use of RWE varied from none to 9% of HTAs between 2012 and 2017. They concluded that not only is RWE infrequently used in HTA, but that it has rarely been influential in decision-making. Their analysis of HTA recommendations with and without RWE in Canada, Germany, France, England, and Scotland did not find a direct correlation between RWE and the recommendations.

Briefly, RWE has been utilized with various purposes in HTA submissions/resubmissions. There are variations of acceptance and influence of RWE across agencies. For resubmission, additional evidence from RWD may play an important role in decision-making. For initial submission, the influence on decision-making may be uncertain.

4 Guidelines for Use of RWE in HTA and Collaborative RWE Standard Development

There is growing interest in leveraging RWE to complement evidence from RCTs for the purpose of decision-making for regulatory approval, reimbursement, and pricing. Regulatory agencies such as the FDA [31, 32] and European Medicines Agency (EMA) [33] are leading the way in developing regulatory frameworks and guidance documents for RWE use. Some HTA agencies are also taking initiatives to advance the use of RWE in HTA, including developing good practice guidance. In addition, some regulatory agencies and HTA agencies have already recognized the need for collaboration on RWE standard development and evidentiary alignment. In this section, we provide an overview of the most recent guidance on the use of RWE for HTA, focusing on several influential HTA agencies. Furthermore, we discuss this regulatory and HTA synergy trend and how it is affecting pharmaceutical companies.

4.1 NICE's New RWE Framework

The latest guidance on the use of RWE in HTA is the RWE framework developed by NICE in June 2022 [4]. The RWE framework is a part of NICE Strategy 2021–2026

[34], a 5-year strategic plan focusing on the use of RWE to fill evidence gaps and facilitate patient access to innovative healthcare interventions [4].

The aims of the RWE framework are to identify when and how RWD can be used to improve recommendations and to describe "best practices for planning, conducting, and reporting RWE studies to improve the quality and transparency of the evidence" [4]. The core principles in the guidance for generating high-quality and trusted RWE include [4] the following:

- Data suitability: Ensure data is of good provenance, sufficient quality and relevance to answer the research question.
- Transparency: Generate evidence in a transparent way and with integrity from study planning through to study conduct and reporting.
- Methods: Use analytical methods that minimize the risk of bias and characterize uncertainty.

The NICE RWE framework formalizes the acceptability of RWE as a source of evidence and outlines the role of RWE in HTA submissions. It acknowledges that RWD and RWE are already widely used for various purposes in NICE decision-making and discusses their potential use within NICE guidance.

The framework provides guidance and a tool for assessing data suitability. It emphasizes that the data used to inform NICE decision-making should be reported transparently and be of good provenance and fit-for-purpose to address the research question. NICE has created the Data Suitability Assessment Tool (DataSAT) for researchers to justify their data source selection.

The framework also has a dedicated section on methods for real-world studies of comparative effects. It suggests that non-randomized studies can be used to provide evidence on comparative effects in the absence of RCTs or to complement trial evidence to answer a broader range of questions about the effects of intervention [4]. It provides specific recommendations for conducting non-randomized studies, including traditional observational studies, as well as clinical trials that use RWD to form external control arms. The framework recommends that RWE study developers follow the "target trial" approach when designing an RWE study. The framework specifies the analysis to be conducted for RWE studies to limit bias, control the confounders, and assess the robustness of the findings. To address the risk of confounding bias, potential confounders should be identified based on a transparent, systematic approach, and causal assumptions should be clearly articulated. Confounding bias should be controlled by using a variety of statistical methods considering both observed and unobserved confounders. In addition, information bias from informative censoring, missing data, and measurement error should be addressed appropriately if needed. Sensitivity analysis should be used to assess the robustness of the studies. Key sensitivity analyses should vary across multiple dimensions, such as follow-up time, model specifications, use of covariates, and algorithms for defining outcomes and covariates [4].

NICE has specified that the framework will be a "living framework," so it will be constantly updated to reflect evolving processes and methodologies in the future. NICE's RWE framework is among the most thorough RWE guidance to date

and attempts to fill multiple gaps in prior recommendations for RWE generation. It provides a great resource for researchers to develop RWE studies, especially comparative effectiveness studies.

4.2 ICER's 2020–2023 Value Assessment Framework

In February 2020, ICER in the US announced that it will generate new RWE for its 2020–2023 assessments [35]. ICER has adopted a new platform to generate RWE and is making an effort in advancing standards for the use of transparent, replicable RWE in technology assessment. In a pilot program, ICER will revisit assessment 24 months after initial publication to analyze RWE for drugs that received accelerated approval. ICER's goal is to determine how the drug is performing in the real-world and if any changes should be made to the cost-effectiveness model.

At the time of product launch, especially for drugs that received accelerated approval, the evidence for comparative effectiveness is often limited. ICER is looking for additional evidence post-launch to address uncertainties and provide a more comprehensive view of the drug's comparative effectiveness and cost-effectiveness at multiple time points. This creates great opportunities for RWE to play a more influential role in the decision-making. This expanding potential of RWE could also drive a new drug-development paradigm. Pharmaceutical companies can actively plan RWE studies, shape the role of RWE in drug development, and influence the way RWE is incorporated into decision-making.

4.3 REALISE Guidance

The REAL World Data In ASia for HEalth Technology Assessment in Reimbursement (REALISE) working group is a collaboration between global experts and 11 Asian health systems. It published REALISE Guidance in 2021 [36]. The detailed framework provides guidance on the use of RWD/RWE in a consistent and efficient way for drug reimbursement decision-making in Asia. The guidance provided directions on the following: (1) When is it appropriate to consider RWD/RWE for reimbursement decisions? (2) What types of RWD should we collect? (3) What are the data sources for RWD? (4) How should we collect RWD? (5) Who should collect RWD? (6) How will RWD be analyzed or processed to generate RWE? (7) How should we use RWE in decision-making? (8) What are the potential biases and how to deal with these biases? and (9) What are the ethical considerations in collecting RWD and generating RWE? [36]. The guidance can increase the quality of RWD/RWE collected and its usage in HTA in Asian countries.

4.4 HAS's Methodology Guidance

HAS, the national HTA agency in France, published a methodology guidance regarding use of real-world studies in June 2021 [37]. The guidance document consists of three chapters. The first chapter discusses the research questions that may be raised during the clinical development of a medicinal product or a medical device and which may justify the implementation of a real-world study. The second chapter summarizes the main HAS recommendations for how to conduct a real-world study for the (re)evaluation by HAS of a medicinal product or a medical device. It covers six aspects: (1) draft a protocol, with the support of a scientific committee; (2) propose a study design consistent with the research questions identified; (3) use pre-existing data; (4) collect good-quality data; (5) integrate patient-reported outcome measures; and (6) guarantee data transparency [37]. The third chapter provides the international methodological references to be taken into account when conducting a real-world study.

This document does not provide a "ready-made formula" that can be applied under all scenarios; rather, it is considered as a high-level methodology guide.

4.5 Collaboration Between CADTH and Health Canada

CADTH, the HTA body in Canada, has also been actively engaged in supporting and providing guidance on the optimal use of RWE. In addition, CADTH and Health Canada, the federal regulatory authority in Canada, have already recognized the need for collaboration and RWE standard development and have partnered on incorporating RWE into both regulatory and reimbursement decision-making.

In 2018, Health Canada and CADTH held a joint workshop launching an initiative to integrate RWE throughout the life cycle of drugs. At this workshop, they announced their intention to codevelop an action plan to optimize and formalize the process for the systematic use and integration of RWE into both regulatory and reimbursement decision-making in Canada [38]. This resulted in joint work on the use of RWE across the product lifecycle, which was published in 2019. In April 2019, Health Canada published the document "Optimizing the Use of Real World Evidence to Inform Regulatory Decision-Making" [39], acknowledging that the use of RWE in regulatory decision-making is increasing globally in the assessment of drug safety, efficacy, and effectiveness. An accompanying document, "Elements of Real World Data/Evidence Quality throughout the Prescription Drug Product Life Cycle" [40], was also published. This document provided key principles to guide the generation of RWE with respect to protocol development, data quality, and prospective and retrospective data collection [40]. The document also identified that certain diseases (such as rare diseases) placed constraints on the conduct of RCT and that studies based on RWE could offer appropriate supporting evidence [39]. In March 2020, a strategy document was published announcing how Health Canada, in

collaboration with CADTH, will be formalizing the integration of RWD/RWE into decision-making [41].

To sum up, RWE is being leveraged throughout the drug lifecycle from early discovery and clinical development, to supporting regulatory approval, all the way through to market access and reimbursement. The landscape of recommendations and guidance on use of RWE evolves from fragmented position pieces to comprehensive guidance, although develops at different stages across regulatory and HTA agencies. The trend toward greater use of RWE by both regulatory and HTA agencies has great implications for pharmaceutical companies, offering them new opportunities to complement RCTs and inform regulatory and HTA decisions. It is essential for the industry to keep abreast of the evolving environment and optimize their RWE strategies. Pharmaceutical companies need to identify, collect, and analyze RWE across a drug's lifecycle in order to support submissions. Careful attention should be paid to ensure that data are effectively leveraged to drive product approval and facilitate patient access to innovative medicines that demonstrate value in the real world [42].

5 Discussion

Innovation and ever-growing research capabilities have been driving the introduction of new heath technologies and interventions in the past several decades. As countries around the world seek to deliver universal health coverage subject to budget constraints, HTA bodies are under increasing pressure to restrict patient access to health technologies. Meanwhile, patients' expectations of effective health care are growing. Consequently, the process of deciding which health technologies and medical interventions to invest in has become not only increasingly imperative but getting more and more rigorous. Correspondingly, fit-in-purpose evidence is much in demand from HTA bodies. The use and potential benefit or risk of health technology in the real world is a major piece of information based on which the HTA bodies make decisions.

As discussed in Sects. 2 and 3, RWE researchers investigated the use of RWE in HTA submissions and decision-making from the period around 2011 to 2018 among selected agencies. The use of RWE varies significantly among agencies, is generally limited, and remains supplementary to RCTs but not influential in decision-making.

HTA bodies are facing several challenges when using RWE in their decision-making. The quality and credibility of the effectiveness estimates reported in real-world studies are among the hurdles for HTA bodies to use RWE in decision-making. More importantly, HTA bodies prefer evidence from RCTs over RWE when appraising health technologies. This is known as the evidence hierarchy [43]. RCTs remain the gold standard to demonstrate efficacy not only in regulatory review but also in technology assessment. However, the difference between the outcomes received from RCTs and those observed in real-world clinical practices has been recognized. The value of RWE lies in its foundation in routine clinical practice,

thus conquers the limitation of RCTs in terms of representativeness of real-world patient population and clinically relevant situations in real-world setting. Evidence demonstrating a health technology's real-world performance should be the key pillar of assessment. What matters in an optimal decision-making framework is the relevance and fit-for-purpose of evidence, not the source of data.

Optimizing the use of RWE in HTA decision-making might still take some time. However, we are witnessing the significant progress that have been made for all healthcare stakeholders including HTA bodies to recognize the value RWE provides in technology assessment. RWE research never stops the pace of progress. Following regulatory agencies' steps in developing guidelines for engaging RWE to support regulatory decision-making, HTA agencies are making efforts in advancing standards for the use of transparent and replicable RWE in technology assessment. The impact of well-developed and high-quality RWE in HTA decision-making will be strengthened.

It is critical for the stakeholders including pharmaceutical companies to collaborate to develop a framework for evidence planning, gathering, and value demonstration to meet the needs from the HTA bodies around the world.

References

1. Hamouzadeh P, Sadrollahi A, Yousefvand M. The basis for calculating costs in health technologies assessment studies 2021.
2. Nordon C, Karcher H, Groenwold RH, Ankarfeldt MZ, Pichler F, Chevrou-Severac H, et al. The "efficacy-effectiveness gap": historical background and current conceptualization. Value in Health 2016;19:75–81.
3. FDA. Framework for FDA's real-world evidence program. Silver Spring, MD: US Department of Health and Human Services Food and Drug Administration 2018.
4. NICE. Real world evidence framework. London: National Institute for Health and Care Excellence (NICE), 2022. Available from: https://www.nice.org.uk/corporate/ecd9/resources/nice-realworld-evidence-framework-pdf-1124020816837.
5. Makady A, Ten Ham R, de Boer A, Hillege H, Klungel O, Goettsch W. Policies for use of real-world data in health technology assessment (HTA): a comparative study of six HTA agencies. Value in Health 2017;20:520–32.
6. CADTH. Use of real-world evidence in single-drug assessments. Ottawa: Canadian Agency for Drugs and Technologies in Health, 2018. Available from: https://www.cadth.ca/sites/default/files/pdf/es0323-rwe-in-single-drug-appraisal.pdf.
7. Simpson A, Ramagopalan SV. R WE ready for reimbursement? A round up of developments in real-world evidence relating to health technology assessment: part 6. Journal of Comparative Effectiveness Research 2022;11:473–5.
8. The G-BA implements post launch real world evidence generation for Zolgensma®. 2021. Available from: https://www.lightning.health/news/the-g-ba-implements-post-launch-real-world-evidence-generation-for-zolgensma/.
9. Makady A, van Veelen A, Jonsson P, Moseley O, D'Andon A, de Boer A, et al. Using real-world data in health technology assessment (HTA) practice: a comparative study of five HTA agencies. Pharmacoeconomics 2018;36:359–68.
10. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford University Press; 2015.

11. Sullivan SD, Mauskopf JA, Augustovski F, Caro JJ, Lee KM, Minchin M, et al. Budget impact analysis—principles of good practice: report of the ISPOR 2012 Budget Impact Analysis Good Practice II Task Force. Value in Health 2014;17:5–14.
12. Lee, W., Dayer, V., Jiao, B., Carlson, J.J., Devine, B. and Veenstra, D.L., 2021. Use of real-world evidence in economic assessments of pharmaceuticals in the United States. Journal of Managed Care & Specialty Pharmacy, 27(1), pp. 5-14.
13. Lee W, Dayer V, Jiao B, Carlson JJ, Devine B, Veenstra DL. Use of real-world evidence in economic assessments of pharmaceuticals in the United States. Journal of Managed Care & Specialty Pharmacy 2021;27:5–14.
14. Segwagwe, Curtin and Berg 2022 The impact of the use of Real-World Evidence (RWE) for NICE submissions. Available at https://www.iqvia.com/locations/united-kingdom/blogs/2022/08/the-impact-of-the-use-of-real-world-evidence-rwe-for-nice-submissions.
15. Kim H-S, Lee S, Kim JH. Real-world evidence versus randomized controlled trial: clinical research based on electronic medical records. Journal of Korean Medical Science 2018;33.
16. Nazha B, Mishra M, Pentz R, Owonikoko TK. Enrollment of racial minorities in clinical trials: old problem assumes new urgency in the age of immunotherapy. American Society of Clinical Oncology Educational Book 2019;39:3–10.
17. Stang A. Randomized controlled trials—an indispensible part of clinical research. Deutsches Ärzteblatt International 2011;108:661.
18. Nazha B, Yang JC, Owonikoko TK. Benefits and limitations of real-world evidence: lessons from EGFR mutation-positive non-small-cell lung cancer. Future Oncology 2021;17:965–77.
19. Eichler H-G, Abadie E, Breckenridge A, Flamion B, Gustafsson LL, Leufkens H, et al. Bridging the efficacy–effectiveness gap: a regulator's perspective on addressing variability of drug response. Nature Reviews Drug Discovery 2011;10:495–506.
20. Bartlett VL, Dhruva SS, Shah ND, Ryan P, Ross JS. Feasibility of using real-world data to replicate clinical trial evidence. JAMA Network Open 2019;2:e1912869–e1912869.
21. CADTH. Procedures for CADTH drug reimbursement reviews. Ottawa: Canadian Agency for Drugs and Technologies in Health, October 2020. Available from: https://cadth.ca/sites/default/files/Drug_Review_Process/CADTH_Drug_Reimbursement_Review_Procedures.pdf.
22. Dai WF, Arciero V, Craig E, Fraser B, Arias J, Boehm D, et al. Considerations for developing a reassessment process: report from the Canadian real-world evidence for value of cancer drugs (CanREValue) Collaboration's Reassessment and Uptake Working Group. Current Oncology 2021;28:4174–83.
23. McNair D, Lumpkin M, Kern S, Hartman D. Use of RWE to inform regulatory, public health policy, and intervention priorities for the developing world. Clinical Pharmacology & Therapeutics 2022;111:44–51.
24. Jaksa A. Does real world evidence matter in Health Technology Assessments? 2015. Available from: https://pharmaphorum.com/articles/does-real-world-evidence-matter-in-health-technology-assessments/.
25. Submission. Ipilimumab, concentrate solution for I.V. infusion, 50 mg in 10 mL, 200 mg in 40 mL, Yervoy® - July 2011. Available from: https://www.pbs.gov.au/info/industry/listing/elements/pbac-meetings/psd/2011-07/pbac-psd-ipilimumab-july11.
26. Resubmission 1. Ipilimumab, concentrate solution for I.V. infusion, 50 mg in 10 mL, 200 mg in 40 mL, Yervoy® - March 2012. Available from: https://www.pbs.gov.au/info/industry/listing/elements/pbac-meetings/psd/2012-03/ipilimumab.
27. Resubmission 2. Ipilimumab, concentrate solution for I.V. infusion, 50 mg in 10mL, 200 mg in 40 mL, Yervoy® - November 2012. Available from: https://www.pbs.gov.au/info/industry/listing/elements/pbac-meetings/psd/2012-11/ipilimumab.
28. SMC. Aflibercept (Zaltrap®) is accepted for use within NHS Scotland. 2014. Available from: https://www.scottishmedicines.org.uk/medicines-advice/aflibercept-zaltrap-resubmission-87813/.
29. NICE. Technology appraisal guidance [TA524] Brentuximab vedotin for treating CD30-positive Hodgkin lymphoma. June 2018. Available from: https://www.nice.org.uk/guidance/ta524/resources/brentuximab-vedotin-for-treating-cd30positive-hodgkin-lymphoma-pdf-82606840474309.

30. Jao R, Jaksa A, Pontynen A, Wang X. Health technology assessment (HTA) agencies consideration of real world evidence (RWE). Value in Health 2018;21:S7.
31. Food and Drug Administration. Framework for FDA's real-world evidence program. December 2018. Available at https://www.fda.gov/media/120060/download.
32. FDA. Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drug and Biological Products. Silver Spring, MD: Food and Drug Administration. September 2022. Available from: https://www.fda.gov/media/124795/download.
33. EMA. Guideline on registry-based studies. Amsterdam: European Medicines Agency. 2021. Available from: https://www.ema.europa.eu/en/guideline-registry-based-studies-0.
34. NICE. The NICE Strategy 2021 to 2026. London: National Institute for Health and Care Excellence (NICE), 2021. Available from: https://www.nice.org.uk/about/who-we-are/corporate-publications/the-nice-strategy-2021-to-2026.
35. ICER. 2020–2023 Value Assessment Framework. Boston: Institute for Clinical and Economic Review, 2020. Available from: https://icer.org/wp-content/uploads/2020/10/ICER_2020_2023_VAF_102220.pdf.
36. Lin LW, Ahn J, Bayani DBS, Chan K, Choiphel D, Isaranuwatchai W, et al. Use of real-world data and real-world evidence to support drug reimbursement decision-making in Asia.
37. Real-world studies for the assessment of medicinal products and medical devices, 2021, available at https://www.has-sante.fr/upload/docs/application/pdf/2021-06/real-world_studies_for_the_assessment_of_medicinal_products_and_medical_devices.pdf.
38. Tadrous M, Ahuja T, Ghosh B, Kropp R. Developing a Canadian real-world evidence action plan across the drug life cycle. Healthcare Policy 2020;15:42.
39. Health Canada. Optimizing the use of real world evidence to inform regulatory decision-making. Ottawa: Government of Canada. April 2019. Available from: https://www.canada.ca/en/health-canada/services/drugs-health-products/drug-products/announcements/optimizing-real-world-evidence-regulatory-decisions.html#_blank.
40. Health Canada. Elements of Real World Data/Evidence Quality throughout the Prescription Drug Product Life Cycle. March 2019. Available at https://www.canada.ca/en/services/health/publications/drugs-health-products/real-world-data-evidence-drug-lifecycle-report.html.
41. Health Canada and the Pan-Canadian health technology assessment collaborative. A strategy to optimize the use of real-world evidence across the medical device life cycle in Canada. Available from: https://www.canada.ca/en/health-canada/corporate/transparency/regulatory-transparency-and-openness/improving-review-drugs-devices/real-world-evidence-medical-device-strategy.html.
42. The case for real-world evidence in health technology assessment. Available at https://aetion.com/evidence-hub/the-case-for-real-world-evidence-in-health-technology-assessment/.
43. Griffiths EA, Macaulay R, Vadlamudi NK, Uddin J, Samuels ER. The role of noncomparative evidence in health technology assessment decisions. Value in Health 2017;20:1245–51.

Part IV
Application and Case Studies

Examples of Applying Causal-Inference Roadmap to Real-World Studies

Yixin Fang

1 Introduction

ICH E9(R1) [1] states that "a central question for drug development and licensing is to establish the existence and estimate the magnitude of treatment effects: how the outcome of treatment compares to what would have happened to the same subjects under alternative treatment." The question is asked in terms of potential or counterfactual outcomes, which are a concept used in causal inference [2]. To derive real-world evidence (RWE) from the analysis of real-world data (RWD) generated from real-world studies, we follow the causal-inference roadmap described in chapter "Causal Inference with Targeted Learning for Producing and Evaluating Real-World Evidence" of this book and other papers such as [3–5].

How to form a sound research question in real-world setting is discussed in chapter "Key Considerations in Forming Research Questions and Conducting Research in Real-World Setting" of this book, and the research question will be driving the choice of data, design, and analytic methods. The causal-inference roadmap consists of six key steps:

(i) Describe the observed data and the data generating experiment
(ii) Specify a realistic model for the distribution of the observed data
(iii) Define the target estimand of the observed data distribution
(iv) Propose an estimator of the target estimand
(v) Obtain estimate, uncertainty measurement, and statistical inference
(vi) Conduct sensitivity analysis and interpret the statistical results

Y. Fang (✉)
Data and Statistical Sciences, AbbVie, North Chicago, IL, USA
e-mail: yixin.fang@abbvie.com

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_18

In chapters "Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods" and "Sensitivity Analysis in the Analysis of Real-World Data" of this book, we review methods for confounding adjustment and for conducting sensitivity analysis, respectively. In this chapter, we demonstrate the application of the roadmap to the following scenarios:

1. Cohort studies with continuous or binary outcomes
2. Single-arm studies with external controls
3. Cohort studies with intercurrent events (ICEs)

We utilize a subset of the NHEFS study (National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study) to generate real-data examples, combined with simulated outcomes, missing data, and intercurrent events. A detailed description of the NHEFS study can be found at www.cdc.gov/nchs/nhanes/nhefs/. The subset of the NHEFS data (including the subject ID, outcome variable, and 9 baseline variables) was prepared by Hernan and Robins for their book [6]. The dataset can be downloaded from their book website, https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/. The authors of book [6] used the dataset illustrating the analyses discussed in the book and stated the following statement: "encourage readers to improve upon and refine our analyses."

The remaining of this chapter is organized as follows. In Sects. 2–4, we provide examples of applying the causal-inference roadmap to the above three scenarios, respectively. We conclude with a summary in Sect. 5.

2 Cohort Studies with Continuous or Binary Outcomes

2.1 Describe the Observed Data and the Data Generating Experiment

The NNEFS dataset described in the introduction section can be considered as a dataset from a cohort study, where A indicates whether or not the subject quit smoking between 1971 and 1982, which is named $qsmk$ in the dataset. There is one limitation in considering $qsmk$ as the point-exposure variable because the exact time of quitting smoking is unknown. For the purpose of this chapter, like in [6], we consider it as a binary point-exposure variable measured in 1971. To fit into the pharmaceutical setting, we may consider quitting smoking as one kind of "behavioral treatment," and non-quitting as the comparator. The outcome variable Y_{con} is the weight change measured in kilograms (kg) defined as the body weight in 1982 minus that in 1971, which is named $wt82\_71$ in the dataset. In addition, we dichotomize this outcome variable into a binary outcome, setting $Y_{bin} = 1$ if $Y_{con} > 5$ kg and $Y_{bin} = 0$ if $Y_{con} \leq 5$ kg.

The vector W includes 9 baseline characteristics measured in 1971: (1) sex (0: male, 1: female), (2) race (0: white, 1: other), (3) age, (4) education (1: 8th grade or less, 2: high school dropout, 3: high school, 4: college dropout, 5: college or

Table 1 Means and SMDs of baseline characteristics W between two cohorts, along with treatment variable A and outcome variable Y

Variable notation	Variable name	Treated $n_1 = 403$	Control $n_0 = 1163$	SMD
W	age	46.17	42.79	0.28
	sex	0.45	0.53	−0.16
	$race$	0.09	0.15	−0.20
	$exercise$	1.25	1.18	0.11
	$active$	0.69	0.63	0.09
	$education$	2.79	2.68	0.08
	$smokeintensity$	18.60	21.19	−0.21
	$smokeyrs$	26.03	24.09	0.15
	$wt71$	72.35	70.30	0.13
A	$qsmk$			
Y	$Y_{con} = wt82\_71$; $Y_{bin} = wt82\_71\_bin$			

more), (5) cigarettes per day, (6) years of smoking, (7) exercise (0: much exercise, 1: moderate exercise, 2: little or no exercise), (8) active (0: very active, 1: moderately active, 2: inactive), and (9) weight in 1971 in kg. There are $n_1 = 403$ subjects in the quitter cohort of $A = 1$ (also referred to as the treated cohort) and $n_0 = 1163$ subjects in the non-quitter cohort of $A = 0$ (also referred to as the control cohort).

Table 1 summarizes the baseline characteristics comparison between the treated cohort ($n_1 = 403$) and the control cohort ($n_0 = 1163$), where sex and race are binary variables; education, exercise, and active are ordinal variables; and the others are quantitative variables. Standardized mean difference (SMD) is used for balance checking, defined as the mean difference between two cohorts divided by the standard deviation among the subjects in two cohorts combined. Seven variables have absolute SMD bigger than 0.1, indicating the need to adjust for potential confounding bias in the estimation of treatment effect [10].

We also summarize the unadjusted comparisons of Y_{con} and Y_{bin} between two cohorts. The averages of Y_{con} in the treated and control cohorts are 4.53 kg and 1.98 kg, respectively, and the unadjusted 95% confidence interval (CI) estimate of the treatment effect is (1.58, 3.50) in kg. The proportions of Y_{bin} in the treated and control cohorts are 43.92% and 30.44%, respectively, and the unadjusted 95% CI estimate of the treatment effect is (7.96%, 19.00%).

When conducting causal inference, we should describe not only the observed data but also the data generating process. The same Fig. 1 of chapter "Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods" describes the data generating process of baseline vector W, treatment variable A, and outcome variable Y, which could be either Y_{con} or Y_{bin}. Baseline vector was measured in 1971, treatment variable was measured between 1971 up to 1982 and was dependent on W, and outcome variable was measured in 1982 and was dependent on both W and A. Missing data are omitted in the preparation of the original dataset. For the purpose of demonstration, we will consider the original dataset in some examples and create missing data in some examples.

2.2 Specify a Realistic Model for the Observed Data

For each subject, the observed data are $O = (W, A, Y)$, where $Y = Y_{con}$ or Y_{bin} depending on which outcome variable is considered as the primary outcome variable. The analysis set consists of $n = n_1 + n_0$ independent and identically distributed copies, $O_i, i = 1, \cdots, n$, following true distribution P_0. We specify our structural causal model (SCM) according to Fig. 1 of chapter "Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods",

$$\begin{aligned} W &= f_W(U_W), \\ A &= f_A(W, U_A), \\ Y &= f_Y(W, A, U_Y), \end{aligned} \tag{1}$$

where $U = (U_W, U_A, U_Y)$ are the exogenous variables following joint distribution P_U, and no assumptions are made on f_W, f_A, f_Y, and P_U.

2.3 Define the Target Estimand

With this SCM, for each subject, the potential outcomes are $Y^1 = f_Y(W, 1, U_Y)$ and $Y^0 = f_Y(W, 0, U_Y)$ under two treatment conditions. This implies the consistency assumption, $Y = Y^1 A + Y^0(1 - A)$. We also make the no unmeasured confounder (NUC) assumption, $U_A \perp\!\!\!\perp U_Y$, and the positivity assumption, $P_0(A = 1, W = w) > 0$ and $P_0(A = 0, W = w) > 0$, for each possible realization w of W.

Under the consistency, NUC, and positivity assumptions, the average treatment effect (ATE) can be expressed in terms of the distribution P_0 of the observed data O,

$$\begin{aligned} & E(Y^1) - E(Y^0) \\ &= E_0\{E_0(Y|A = 1, W) - E_0(Y|A = 0, W)\} \triangleq \theta_{ATE}(P_0). \end{aligned} \tag{2}$$

2.4 Propose an Estimator of the Target Estimand

To estimate the estimand of interest, $\theta_{ATE}(P_0)$, we propose to consider the targeted maximum likelihood estimator (TMLE; [7]). First, TMLE is a plugin estimator; that is, it is of form $\theta_{ATE}(\widehat{P})$ with P_0 replaced by its estimator $\widehat{P}$. Second, TMLE is a doubly robust estimator; that is, it is consistent if either the estimator of the propensity score function $g(a|w) = P_0(A = a|W = w)$ or the estimator of the regression function $Q(a, w) = E_0(Y|A = a, W = w)$ is consistent. Third, the estimator is asymptotically efficient; roughly speaking, the variance of the

estimator achieves the Cramer-Rao bound asymptotically. Refer to book [7] for more discussion on the properties of TMLE.

2.5 Obtain Estimate, Uncertainty Measurement, and Inference

We use R function "tmle" of R package "tmle" to implement the TMLE estimation, with the following excerpt of R codes:

```
# wt82_71 is continuous outcome so family=gaussian is set
# Q.SL.library is set for fitting regression Q(a,w)
# g.SL.library is set for fitting propensity g(a|w)
nhefs_tmle_con <- tmle(Y=data0$wt82_71, A=data0$qsmk, W=data0[,W.
     vec],
   Q.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
        randomForest"),
   g.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
        randomForest"),
   family = "gaussian")

# point estimate and 95% confidence interval
nhefs_tmle_con$estimates$ATE$psi
nhefs_tmle_con$estimates$ATE$CI

# wt82_71_bin is binary outcome so family=binomial is set
nhefs_tmle_bin <- tmle(Y=data0$wt82_71_bin, A=data0$qsmk, W=data0
     [,W.vec],
   Q.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
        randomForest"),
   g.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
        randomForest"),
   family = "binomial")

nhefs_tmle_bin$estimates$ATE$psi
nhefs_tmle_con$estimates$ATE$CI
```

For continuous outcome Y_{con}, the point estimate of $\theta_{ATE}(P_0)$ is 3.42, with SE $=$ 0.43 and 95% CI $=$ $(2.58, 4.25)$. For binary outcome Y_{bin}, the point estimate of $\theta_{ATE}(P_0)$ is 16.70%, with SE $=$ 2.63% and 95% CI $=$ (11.54%, 21.86%).

2.6 Conduct Sensitivity Analysis and Interpret the Results

For continuous outcome Y_{con}, the unadjusted estimate of $\theta_{ATE}(P_0)$ is 2.55 with 95% CI $=$ $(1.58, 3.50)$. Due to non-randomization and presence of confounding bias, we adjust for 9 confounders and assume there is no unmeasured confounder, providing the TMLE estimate of $\theta_{ATE}(P_0)$, 3.42 with 95% CI $=$ $(2.58, 4.25)$.

We apply E-value [8] to conduct sensitivity analysis for exploring the robustness of the result to unmeasured confounders. E-Value was originally proposed for binary outcome. For continuous outcome, we first calculate the point estimate of the standardized mean difference (i.e., Cohen's d), using the standard deviation of the outcome variable, which is equal to SD $=$ 7.88. Then the point estimate of Cohen's d is $3.42/7.88 = 0.4340$ with SE $= 0.43/7.88 = 0.0546$. We use R function "evalues.MD" of R package "EValue" to calculate the E-value, with the following R codes:

```
> evalues.MD(est=0.4340, se=0.0546)
```

From the R output, the E-value corresponding to the point estimate is 2.33 and the E-value corresponding to the lower end of 95% CI is 2.03. This means that, in order to explain away the estimated treatment effect, unmeasured confounder(s) would need to more than double the probability of a subject's being exposed versus not being exposed and would also need to more than double the probability of being high versus low on the outcome. This sensitivity analysis shows that the estimated treatment effect is robust.

For binary outcome Y_{bin}, the unadjusted estimate of $\theta_{ATE}(P_0)$ is 13.48%, with 95% CI $= (7.96\%, 19.00\%)$. Due to non-randomization and presence of confounding bias, we adjust for 9 confounders and assume there is no unmeasured confounder, providing the TMLE estimate of $\theta_{ATE}(P_0)$, 16.70% with 95% CI $=$ $(11.54\%, 21.86\%)$.

Similarly, we apply E-value to conduct sensitivity analysis. For binary outcome, E-Value was originally proposed in terms of relative risk. Therefore, we obtain TMLE of the treatment effect in terms of relative risk, which is 1.56 with 95% CI $= (1.37, 1.78)$. Then, we use R function "evalues.RR" of R package "EValue" to calculate the E-value, with the following R codes:

```
> evalues.RR(est=1.56, lo=1.37)
```

From the R output, we see that the E-value corresponding to the point estimate is 2.49 and the E-value corresponding to the lower end of 95% CI is 2.08. This means that, in order to explain away the estimated treatment effect, unmeasured confounder(s) would need to more than double the probability of a subject's being exposed versus not being exposed and would also need to more than double the probability of being high versus low on the outcome. This sensitivity analysis shows that the estimated treatment effect is robust.

3 Single-arm Studies with External Controls

3.1 Describe the Observed Data and the Data Generating Experiment

This subsection is the same as Sect. 2.1, except that now we consider the treated cohort of $A = 1$ as the data from a single-arm study and the control cohort of $A = 0$ as the external-control group. Moreover, the data generating process is the same as that in Sect. 2.1.

One popular method for analyzing data from a single-arm study with external controls is to construct a subset of the external-control group that is matched with the single-arm dataset using the propensity score matching [9]. However, as discussed in chapter "Recent Statistical Development for Comparative Effectiveness Research Beyond Propensity-Score Methods", the propensity score matching method is not efficient. In the following steps, we apply the targeted learning method.

3.2 Specify a Realistic Model for the Observed Data

This subsection is the same as Sect. 2.2, except that now $A = 1$ indicates the single-arm study and $A = 0$ indicates the external-control group.

3.3 Define the Target Estimand

This subsection is the same as Sect. 2.3, except that now we are interested in the average treatment effect among the treated (ATT). Under the consistency, NUC, and positivity assumptions, ATT can be expressed in terms of the distribution P_0 of the observed data O,

$$\begin{aligned} & E(Y^1|A=1) - E(Y^0|A=1) \\ &= E_{0,W|A=1}\{E_0(Y|A=1,W) - E_0(Y|A=0,W)\} \triangleq \theta_{ATT}(P_0). \end{aligned} \tag{3}$$

3.4 Propose an Estimator of the Target Estimand

To estimate the estimand of interest, $\theta_{ATT}(P_0)$, we propose to consider the TMLE estimator, which has good properties as discussed in Sect. 2.4.

3.5 Obtain Estimate, Uncertainty Measurement, and Inference

The R function "tmle" of R package "tmle" can implement both the ATE and ATT estimation, with the following R codes that output the ATT estimates:

```
nhefs_tmle_con$estimates$ATT$psi
nhefs_tmle_con$estimates$ATT$CI

nhefs_tmle_bin$estimates$ATT$psi
nhefs_tmle_bin$estimates$ATT$CI
```

For continuous outcome Y_{con}, the point estimate of $\theta_{ATT}(P_0)$ is 3.46, with $chap11SE = 0.46$ and 95% CI $= (2.55, 4.37)$. For binary outcome Y_{bin}, the point estimate of $\theta_{ATT}(P_0)$ is 16.46%, with SE $= 2.71\%$ and 95% CI $= (11.15\%, 21.76\%)$.

3.6 Conduct Sensitivity Analysis and Interpret the Results

This subsection is similar to Sect. 2.6, so here we only show the sensitivity analysis for continuous outcome Y_{con}. The point estimate of Cohen's d is 3.46/7.88=0.4391 with SE $= 0.46/7.88 = 0.0584$. Using the following R codes:

```
> evalues.MD(est=0.4391, se=0.0584)
```

we obtain that the E-value corresponding to the point estimate is 2.34 and the E-value corresponding to the lower end of 95% CI is 2.02. This means that, in order to explain away the estimated treatment effect, unmeasured confounder(s) would need to more than double the probability of a subject's being exposed versus not being exposed and would also need to more than double the probability of being high versus low on the outcome. This sensitivity analysis shows that the estimated treatment effect is robust.

4 Cohort Studies with Intercurrent Events

In the above two examples, we consider scenarios where there are no ICEs. In this section, we consider examples where there are ICEs. ICH E9(R1) proposes five strategies for dealing with ICEs: (1) hypothetical strategy, (2) treatment policy strategy, (3) composite variable strategy, (4) while on treatment strategy, and (5) principal stratum strategy. Therefore, we consider five examples in this section, respectively, for applying these five strategies.

4.1 Example of Using Hypothetical Strategy

We consider the same dataset described in Table 1, along with a newly simulated variable E (1 for ICE occurrence; 0 for no ICE), with the following log-odds:

$$\text{logit}(P(E=1)) = 0.5 \times A - age/50 - 1.5 \times (wt71/100)^2, \tag{4}$$

using the following R codes:

```
set.seed(1)
# ice=1 means ICE and ice=0 means no ICE
ice <- rbinom(nrow(data0), size=1, prob=plogis(0.5*data0$qsmk
                - 2*(data0$age/100)-1.5*(data0$wt71/100)^2))

# data after ice=1 are considered missing
wt82_71.m <- ifelse(ice, NA, data0$wt82_71)
```

4.1.1 Describe the Observed Data and the Data Generating Experiment

In the treated cohort, 90 out of 403 subjects have ICE occurrence, while in the control cohort, 204 out of 1163 subjects have ICE occurrence.

According to ICH E9(R1), if the hypothetical strategy is applied to handle the ICE, "a scenario is envisaged in which the intercurrent event would not occur." We envisage a scenario that $E = 1$ would not occur and the subject would be treated by the initial treatment A throughout. Figure 1 displays the data generating experiment, where $Y^{E=0}$ is the potential outcome variable under the envisage scenario where the subject would be treated by the initial treatment A throughout. Let $Y = Y^{E=0}$ be the outcome variable of interest using the hypothetical strategy to handle ICE. We see that if $E = 0$, then $Y = Y^{E=0}$ is observed in the real-world, while if $E = 1$, then counterfactual outcome $Y^{E=0}$ is considered as "missing data." Figure 1 shows that A depends on W, $Y^{E=0}$ depends on W and A, E depends on W and A, and there is no direct path between E and $Y^{E=0}$.

4.1.2 Specify a Realistic Model for the Observed Data

For each subject, the observed data are $O = (W, A, E, (1-E)Y)$. Note that $Y = Y^{E=0}$ is only observed if $E = 0$, while $(1-E)Y$ is equal to 0 if $E = 1$, indicating "missing data." The analysis set consists of $n = n_1 + n_0$ independent and identically distributed copies, $O_i, i = 1, \cdots, n$, following true distribution P_0. We specify our SCM according to Fig. 1,

$$W = f_W(U_W),$$
$$A = f_A(W, U_A),$$

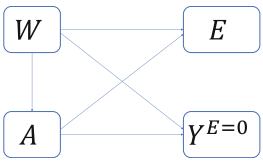

Fig. 1 The data generating process using the hypothesis strategy

$$E = f_E(W, A, U_E),$$
$$Y = f_Y(W, A, U_Y), \tag{5}$$

where $U = (U_W, U_A, U_E, U_Y)$ are the exogenous variables following joint distribution P_U, and no assumptions are made on f_W, f_A, f_E, f_Y, and P_U.

4.1.3 Define the Target Estimand

Besides the consistency, NUC, and positivity assumptions that are made in Sect. 2.3, we assume that U_E and U_Y are independent, that is, the missing at random (MAR) assumption. For each subject the potential outcomes are $Y^{A=1,E=0} = f_Y(W, 1, U_Y)$ and $Y^{A=0,E=0} = f_Y(W, 0, U_Y)$ under two treatment conditions. Under these assumptions, we have

$$\begin{aligned} & E(Y^{A=1,E=0}) - E(Y^{A=0,E=0}) \\ = & E_0\{E_0(Y|W, A = 1, E = 0) - E_0(Y|W, A = 0, E = 0)\} \triangleq \theta_h(P_0). \end{aligned} \tag{6}$$

4.1.4 Propose an Estimator of the Target Estimand

To estimate the estimand of interest, $\theta_h(P_0)$, we propose to consider the TMLE estimator, which has good properties as discussed in Sect. 2.4.

4.1.5 Obtain Estimate, Uncertainty Measurement, and Inference

The R function "tmle" of R package "tmle" can handle missing data, with "delta" argument (1 - observed, 0 - missing), using the following R codes:

```
# delta=0 means missing outcome and delta=1 means no missing
delta <- 1-ice
# g.Delta.SL.library is set for propensity of not-missing
nhefs_tmle_fit.h <- tmle(Y=data0$wt82_71.m, A=data0$qsmk, W=data0
     [,W.vec],
   Delta=data0$delta,
   Q.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
        randomForest"),
```

```
    g.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
         randomForest"),
    g.Delta.SL.library = c("SL.rpart", "SL.glmnet", "SL.gam"),
    family = "gaussian")

nhefs_tmle_fit.h$estimates$ATE$psi
nhefs_tmle_fit.h$estimates$ATE$CI
```

From the output, the point estimate of $\theta_h(P_0)$ is 3.28, with SE $= 0.50$ and 95% CI $= (2.31, 4.28)$.

4.1.6 Conduct Sensitivity Analysis and Interpret the Results

We can use E-value to conduct sensitivity analysis for the NUC assumption, similar to Sect. 2.6. Here we only show sensitivity analysis for the MAR assumption, using two reference-based imputation methods [11], copy-reference (CR), and jump-to-reference (J2R) imputation methods. In this example, there is only one follow-up time point in 1982, so CR and J2R are equivalent. We use R package "mice" to generate 100 imputations, with an excerpt of R codes,

```
data.ref0<-data0[(data0$qsmk==0)|(data0$delta==0), ]
data.ref <-data.ref0
data.ref[, c("wt82_71", "wt82", "qsmk", "delta")] <- NULL
data.refim <- mice(data.ref, m=100, seed=500)
```

For each completed dataset, we obtain an TMLE estimate. Then we use Rubin's rule [12] to combined 100 TMLE estimates, providing pooled estimate 3.35 with SE $= 0.49$. Hence the estimate under CR or J2R is similar to the estimate under the MAR assumption and the result is robust.

4.2 Example of Using Treatment Policy Strategy

We consider the same dataset described in Table 1, along with a newly simulated variable E, using the following R codes:

```
set.seed(1)
ice <- (data0$wt82_71 < -5)
data0$ice <- ice
data0$wt82_71.tp <-data0$wt82_71
data0$wt82_71.tp[ice] <- data0$wt82_71.tp[ice]
                          + rnorm(sum(ice), 5, 1)
```

To understand the above R codes, assume that the ICE is taking rescue medication if the intermediate outcome shows non-response. If the original outcome variable $wt82\_71$ is less than -5, then let $ice = 1$ and add a random number generated from normal distribution $N(5, 1)$ to the outcome variable, generating the final outcome variable $wt82\_71.tp$, denoted as Y.

4.2.1 Describe the Observed Data and Data Generating Experiment

In the treated cohort, 42 out of 403 subjects have $E = 1$, while in the control cohort, 152 out of 1163 subjects have $E = 1$. According to ICH E9(R1), if the treatment policy strategy is applied, "the intercurrent event is considered to be part of the treatments being compared." Therefore, instead of considering treatment variable A, now we consider a dynamic treatment regime $A^*(A, E)$, where $A^*(A, E) = 1$ means "starting with $A = 1$ and taking rescue medication if $wt82\_71 < -5$ kg" and $A^*(A, E) = 0$ means "starting with $A = 0$ and taking rescue medication if $wt82\_71 < -5$ kg."

Figure 2 displays the data generating experiment, where Y is the outcome variable using the treatment policy strategy, E is incorporated into the dynamic treatment regime $A^*(A, E)$. Figure 2 also shows that A^* depends on W, and Y depends on W and A^*. In addition, note that $A^*(A, E) = A$ in this example.

4.2.2 Specify a Realistic Model for the Observed Data

This subsection is the same as Sect. 2.2, except that the dynamic treatment regime $A^*(A, E)$ is of interest and Y is the observed outcome variable regardless of whether or not the ICE occurs. For each subject, the observed data are $O = (W, A, E, Y)$. The analysis set consists of $n = n_1 + n_0$ independent and identically distributed copies, $O_i, i = 1, \cdots, n$, following true distribution P_0. We specify our SCM according to Fig. 2 in the following:

$$\begin{aligned} W &= f_W(U_W), \\ A^* &= f_{A^*}(W, U_{A^*}), \\ Y &= f_Y(W, A^*, U_Y), \end{aligned} \tag{7}$$

where $U = (U_W, U_{A^*}, U_Y)$ are the exogenous variables following joint distribution P_U, and no assumptions are made on f_W, f_{A^*}, f_Y, and P_U.

4.2.3 Define the Target Estimand

For each subject, the potential outcomes are $Y^1 = f_Y(W, 1, U_Y)$ and $Y^0 = f_Y(W, 0, U_Y)$ under two treatment conditions. Under the same consistency, NUC (i.e., U_Y and U_{A^*} are independent), and positivity assumptions that are described

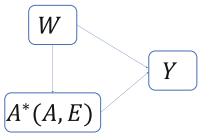

Fig. 2 The data generating process using the treatment policy strategy

Sect. 2.3, we have

$$\begin{aligned} & E(Y^1) - E(Y^0) \\ &= E_0\{E_0(Y|W, A^* = 1) - E_0(Y|W, A^* = 0)\} \triangleq \theta_{tp}(P_0). \end{aligned} \tag{8}$$

4.2.4 Propose an Estimator of the Target Estimand

To estimate the estimand of interest, $\theta_{tp}(P_0)$, we propose to consider the TMLE estimator, which has good properties as discussed in Sect. 2.4.

4.2.5 Obtain Estimate, Uncertainty Measurement, and Inference

Similar to Sect. 2.5, the R function "tmle" of R package "tmle" can implement the proposed method, using the following R codes:

```
nhefs_tmle_fit.tp <- tmle(Y=data0$wt82_71.tp, A=data0$qsmk, W=
     data0[,W.vec],
   Q.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
        randomForest"),
   g.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
        randomForest"),
   family = "gaussian")

nhefs_tmle_fit.tp$estimates$ATE$psi
nhefs_tmle_fit.tp$estimates$ATE$CI
```

From the output, the point estimate of $\theta_{tp}(P_0)$ is 3.19, with $SE = 0.41$ and 95% CI = (2.393.99).

4.2.6 Conduct Sensitivity Analysis and Interpret the Results

Similar to Sect. 2.6, we apply E-value to conduct sensitivity analysis for exploring the robustness of the result to unmeasured confounders. We first calculate the point estimate of Cohen's d, using the standard deviation of the outcome variable, which is equal to SD = 6.98. Then the point estimate of Cohen's d is $3.19/6.98 = 0.4570$ with SE $= 0.41/6.98 = 0.0587$. Using the following R codes:

```
> evalues.MD(est=0.4570, se=0.0587)
```

the E-value corresponding to the point estimate is 2.39 and the E-value corresponding to the lower end of 95% CI is 2.07. This means that, in order to explain away the estimated treatment effect, unmeasured confounder(s) would need to more than double the probability of a subject's being exposed versus not being exposed and would also need to more than double the probability of being high versus low on

the outcome. This sensitivity analysis shows that the estimated treatment effect is robust.

4.3 Example of Using Composite Variable Strategy

We consider the same dataset described in Table 1, along with the same variable E simulated in Sect. 4.1 and the binary outcome variable $Y = Y_{bin}$.

4.3.1 Describe the Observed Data and the Data Generating Experiment

In the treated cohort, 90 out of 403 subjects have ICE occurrence, while in the control cohort, 204 out of 1163 subjects have ICE occurrence. According to ICH E9(R1), if the composite variable strategy is applied, "an intercurrent event is considered in itself to be informative about the patient's outcome and is therefore incorporated into the definition of the variable." Therefore, we incorporate the ICE occurrence as another mode of failure; that is, we define $Y^* = Y^*(Y, E)$, letting $Y^* = 0$ if $Y_{bin} = 0$ or $E = 1$ and letting $Y^* = 1$ if $Y_{bin} = 1$ and $E = 0$. The proportions of $Y^* = 1$ in the treated cohort and control cohort are 34.24% and 25.11%, respectively.

Figure 3 displays the data generating experiment, where Y^* is the outcome variable, which is composite variable combining the binary outcome variable Y_{bin} and ICE indicator E. Figure 3 also shows that A depends on W and Y^* depends on W and A.

4.3.2 Specify a Realistic Model for the Observed Data

For each subject, the observed data are $O = (W, A, Y^*)$. The analysis set consists of $n = n_1 + n_0$ independent and identically distributed copies, $O_i, i = 1, \cdots, n$, following true distribution P_0. We specify our SCM according to Fig. 3,

$$\begin{aligned} W &= f_W(U_W), \\ A &= f_A(W, U_A), \\ Y^* &= f_{Y^*}(W, A, U_{Y^*}), \end{aligned} \tag{9}$$

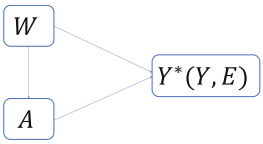

Fig. 3 The data generating process using the composite variable strategy

where $U = (U_W, U_A, U_{Y^*})$ are the exogenous variables following joint distribution P_U, and no assumptions are made on f_W, f_A, f_{Y^*}, and P_U.

4.3.3 Define the Target Estimand

For each subject the potential outcomes are $Y^{*1} = f_{Y^*}(W, 1, U_{Y^*})$ and $Y^{*0} = f_{Y^*}(W, 0, U_{Y^*})$ under two treatment conditions. Under the consistency, NUC (i.e., U_{Y^*} and U_A are independent), and positivity assumptions, we have

$$\begin{aligned} & E(Y^{*1}) - E(Y^{*0}) \\ &= E_0\{E_0(Y^*|W, A = 1) - E_0(Y^*|W, A = 0)\} \triangleq \theta_{cv}(P_0). \end{aligned} \tag{10}$$

4.3.4 Propose an Estimator of the Target Estimand

To estimate the estimand of interest, $\theta_{cv}(P_0)$, we propose to consider the TMLE estimator, which has good properties as discussed in Sect. 2.4.

4.3.5 Obtain Estimate, Uncertainty Measurement, and Inference

The R function "tmle" of R package "tmle" can implement the proposed method, using the following R codes:

```
# Define Y.star as composite variable
data0$Y.star <- ifelse(ice, 0, data0$Y.bin)

nhefs_tmle_fit.cv <-tmle(Y=data0$Y.star, A=data0$qsmk, W=data0[,W
     .vec],
 Q.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
      randomForest"),
 g.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
      randomForest"),
 family = "binomial")

nhefs_tmle_fit.cv$estimates$ATE$psi
nhefs_tmle_fit.cv$estimates$ATE$CI
```

From the output, the point estimate of $\theta_{cv}(P_0)$ is 11.20%, with SE = 2.72% and 95% CI = (5.86%, 16.53%).

4.3.6 Conduct Sensitivity Analysis and Interpret the Results

We apply E-value to conduct sensitivity analysis. We first obtain the TMLE of the treatment effect in terms of relative risk, which is 1.46 with 95% CI=(1.23, 1.72). Using the following R codes:

```
> evalues.RR(est=1.46, lo=1.23)
```

we see that the E-value corresponding to the point estimate is 2.27 and the E-value corresponding to the lower end of 95% CI is 1.76. This means that, in order to explain way the estimated treatment effect, unmeasured confounder(s) would need to increase by about 1.76 times the probability of a subject's being exposed versus not being exposed and would also need to increase by about 1.76 times the probability of being high versus low on the outcome. This sensitivity analysis shows that the estimated treatment effect is only slightly robust.

4.4 Example of Using While on Treatment Strategy

We consider the same dataset described in Table 1, along with variable E simulated similarly as in Sect. 4.1 but with a bigger event rate, a newly simulated time of ICE occurrence, denoted as $t(E)$, and a newly simulated outcome variable measured at the time, denoted as $Y_{t(E)}$. Specially, we simulate $t(E)$ using a uniform distribution between 72 and 81 and simulate $Y_{t(E)}$ using the projection of Y_{con} from year 82 to year $t(E)$, using the following model:

$$\begin{aligned} t(E) &\sim \text{Unif}(72, 81), \text{ if } E = 1; t(E) = 82, \text{ if } E = 0, \\ Y_{t(E)} &= wt82\_71 \times (t(E) - 71)/(82 - 71). \end{aligned} \tag{11}$$

We use R codes to simulate the above three variables, ice, $t.ice$, and $Y.t$,

```
set.seed(1)
ice<- rbinom(nrow(data0), size=1, prob=plogis(1 + 0.5*data0$qsmk
                - 2*(data0$age/100)-1.5*(data0$wt71/100)^2))
t.ice <- rep(82, nrow(data0))
t.ice[ice==1] <- sample(72:81, sum(ice), replace=TRUE)
data0$Y.t <- data0$wt82_71*(t.ice-71)/(82-71)
```

4.4.1 Describe the Observed Data and Data Generating Experiment

In the treated cohort, 177 out of 403 subjects have ICE occurrence, while in the control cohort, 425 out of 1163 subjects have ICE occurrence. The ICE event rate is 38.44%, which is bigger than the previous three examples.

According to ICH E9(R1), if the while on treatment strategy is applied, "response to treatment prior to the occurrence of the intercurrent event is of interest".

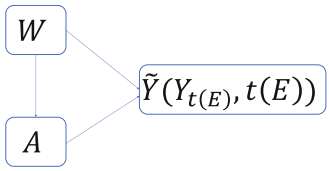

Fig. 4 The data generating process using the while on treatment strategy

Therefore, we incorporate the time of the ICE occurrence into the definition of the primary outcome variable,

$$\widetilde{Y} = \widetilde{Y}(Y_{t(E)}, t(E)) = Y_{t(E)}/(t(E) - 71). \tag{12}$$

The means of $Y_{t(E)}$ are 3.60 kg and 1.53 kg in the treated and control cohorts, respectively. The means of $\widetilde{Y}$ are 0.41 kg/year and 0.18 kg/year the treated and control cohorts, respectively.

Figure 4 displays the data generating experiment, where $\widetilde{Y}$ is the outcome variable, which is defined as the rate of change between $t(E)$ and 1971. Figure 4 also shows that A depends on W and $\widetilde{Y}$ depends on W and A.

4.4.2 Specify a Realistic Model for the Observed Data

For each subject, the observed data are $O = (W, A, \widetilde{Y})$. The analysis set consists of $n = n_1 + n_0$ independent and identically distributed copies, $O_i, i = 1, \cdots, n$, following true distribution P_0. We specify our SCM according to Fig. 4,

$$\begin{aligned} W &= f_W(U_W), \\ A &= f_A(W, U_A), \\ \widetilde{Y} &= f_{\widetilde{Y}}(W, A, U_{\widetilde{Y}}), \end{aligned} \tag{13}$$

where $U = (U_W, U_A, U_{\widetilde{Y}})$ are the exogenous variables following joint distribution P_U, and no assumptions are made on f_W, f_A, $f_{\widetilde{Y}}$, and P_U.

4.4.3 Define the Target Estimand

For each subject the potential outcomes are $\widetilde{Y}^1 = f_{\widetilde{Y}}(W, 1, U_{\widetilde{Y}})$ and $\widetilde{Y}^0 = f_{\widetilde{Y}}(W, 0, U_{\widetilde{Y}})$ under two treatment conditions. Under the same consistent, NUC ($U_{\widetilde{Y}}$ and U_A are independent), and positivity assumptions, we have

$$\begin{aligned} &E(\widetilde{Y}^1) - E(\widetilde{Y}^0) \\ &= E_0\{E_0(\widetilde{Y}|W, A = 1) - E_0(\widetilde{Y}|W, A = 0)\} \triangleq \theta_{wot}(P_0). \end{aligned} \tag{14}$$

4.4.4 Propose an Estimator of the Target Estimand

To estimate the estimand of interest, $\theta_{wot}(P_0)$, we propose to consider the TMLE estimator, which has good properties as discussed in Sect. 2.4.

4.4.5 Obtain Estimate, Uncertainty Measurement, and Inference

The R function "tmle" of R package "tmle" can implement the proposed method, using the following R codes:

```
data0$Y.tilde <- data0$Y.t/(t.ice-71)

nhefs_tmle_fit.wot <- tmle(Y=data0$Y.tilde, A=data0$qsmk, W=data0
     [,W.vec],
 Q.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
      randomForest"),
 g.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
      randomForest"),
 family = "gaussian")

nhefs_tmle_fit.wot$estimates$ATE$psi
nhefs_tmle_fit.wot$estimates$ATE$CI
```

From the output, the point estimate of $\theta_{wot}(P_0)$ is 0.3071 kg/year, with $SE = 0.0425$ and 95% $CI = (0.2238, 0.3905)$.

4.4.6 Conduct Sensitivity Analysis and Interpret the Results

Two main assumptions are made in the primary analysis: (1) the NUC assumption and (2) the linearity assumption made implicitly in the definition of $\widetilde{Y}$ according to the while on treatment strategy.

Similar to Sect. 2.6, we apply E-value to conduct sensitivity analysis for exploring the robustness of the result to the deviation of the NUC assumption. We first calculate the standard deviation of the outcome variable $\widetilde{Y}$, which equals SD $=$ 0.7164. Then the point estimate of Cohen's d is $0.3071/0.7164 = 0.4287$ with SE $= 0.0425/0.7164 = 0.0593$. We use R function "evalues.MD" of R package "EValue" to calculate the E-value, with the following R codes:

```
> evalues.MD(est=0.4287, se=0.0593)
```

From the R output, the E-value corresponding to the point estimate is 2.32 and the E-value corresponding to the lower end of 95% CI is 1.99. This means that, in order to explain away the estimated treatment effect, unmeasured confounder(s) would need to double the probability of a subject's being exposed versus not being exposed and would also need to double the probability of being high versus low on the outcome. This sensitivity analysis shows that the estimated treatment effect is robust to the NUC assumption.

Now we conduct sensitivity analysis for the linearity assumption. We introduce the following parameter to tune the deviation from the linearity assumption:

$$\eta = \frac{Y_{82} - Y_{t(E)}}{82 - t(E)} - \widetilde{Y}, \tag{15}$$

where Y_{82} is the potential value of $wt82\_71$ that would be measured if there were not ICE. Hence the perturbed rate of change between 1971 and 1982 becomes

$$\widetilde{Y}' = [(\widetilde{Y} + \eta)(82 - t(E)) + \widetilde{Y} \times (t(E) - 71)]/(82 - 71). \tag{16}$$

Similar to reference-based imputation, we only add perturbations in terms of η to the subjects in the treated cohort. Because the point estimate of $\theta_{wot}(P_0)$ under the linearity assumption is about 0.3, we select 13 values for η, from -0.3 to 0.3 by 0.05. For each value, we compute perturbed outcome variable $\widetilde{Y}'$, implement same R codes in Sect. 4.4.5, and obtain a version of point estimate and 95% CI of $\theta_{wot}(P_0)$. Figure 5 displays 95% CIs of $\theta_{wot}(P_0)$ corresponding to 13 values of η,

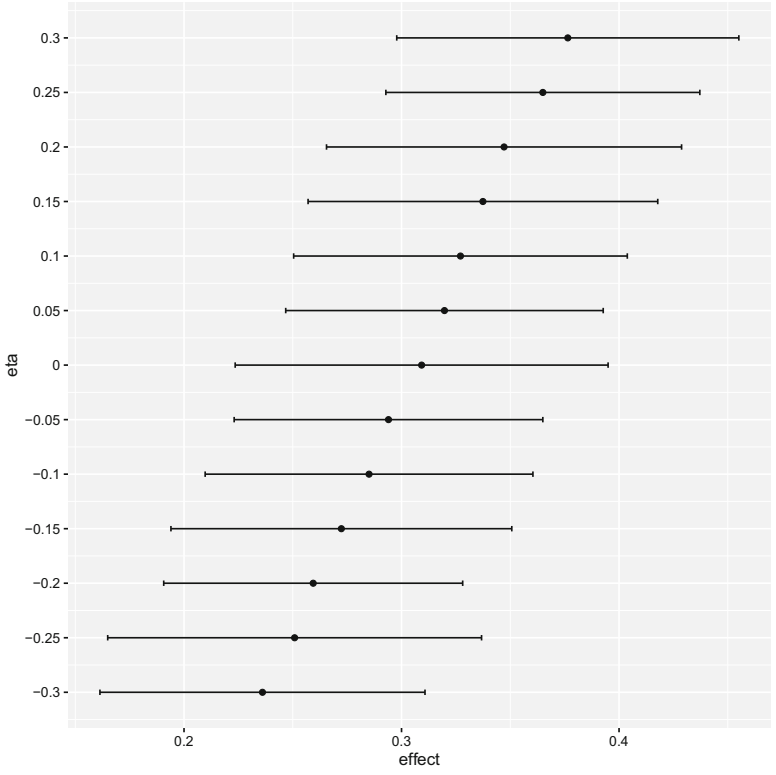

Fig. 5 Sensitivity analysis for the linearity assumption

where $\eta = 0$ is corresponding to the original result, and the others are corresponding to perturbed data. Even for $\eta = -0.3$, the point estimate and 95% are 0.2361 and (0.1613, 0.3108). Hence we can conclude that the result is robust to the deviations of the linearity assumption.

4.5 Example of Using Principal Stratum Strategy

We consider the same dataset described in Table 1, except that we simulate variable E using the following R codes:

```
set.seed(1)
ice <- rbinom(nrow(data0), size=1, prob=plogis(-1.5+2*data0$qsmk
                    + 2*(data0$age/10)-1.5*data0$wt71/10))$
```

4.5.1 Describe the Observed Data and the Data Generating Experiment

According to ICH E9(R1), if the principal stratum strategy is applied, "(t)he target population might be taken to be the principal stratum in which an intercurrent event would occur. Alternatively, the target population might be taken to be the principal stratum in which an intercurrent event would not occur." In this example, we are interested in the principal stratum in which the ICE would not occur. We refer to this principal stratum as PS_0. Within PS_0, A depends on W and Y depends on W and A.

The proportions of ICEs in the treated cohort and the control cohort are $167/403 = 41.44\%$ and $190/1163 = 16.34\%$, respectively. The proportion of ICEs among all the subjects is $(167 + 190)/(403 + 1163) = 22.80\%$. Consider a hypothetical world in which the treated subjects had been untreated and the untreated subjects had been treated. Therefore, PS_0 is the new target population of subjects who have no ICE in the real-world and would have no ICE in the hypothetical world either. A marginal estimate of the proportion of ICEs among all the subjects in the hypothetical world is $41.44\%(1163) + 16.34\%(403) = 34.98\%$.

4.5.2 Specify a Realistic Model for the Observed Data

Define two potential outcomes for the ICE occurrence, $E^{a=1}$ and $E^{a=0}$, where $E^{a=1} = 1$ is the indicator that an ICE would occur if the subject was treated by $a = 1$ and $E^{a=0} = 1$ is the indicator that an ICE would occur if the subject was treated by $a = 0$. Therefore,

$$PS_0 = \{E^{a=1} = 1, E^{a=0} = 1\}. \tag{17}$$

Within PS_0, we specify the same SCM as in Sect. 2.2.

4.5.3 Define the Target Estimand

Within PS_0, for each subject, the potential outcomes are $Y^1 = f_Y(W, 1, U_Y)$ and $Y^0 = f_Y(W, 0, U_Y)$ under two treatment conditions. Under the same consistent, NUC, and positivity assumptions that are made in Sect. 2.3, we have

$$\begin{aligned} & E_{PS_0}(Y^1) - E_{PS_0}(Y^0) \\ = & E_{PS_0}\{E_{PS_0}(Y|W, A = 1) - E_{PS_0}(Y|W, A = 0)\} \triangleq \theta_{ps}(P_0), \end{aligned} \tag{18}$$

where the expectation is over principal stratum PS_0.

4.5.4 Propose an Estimator of the Target Estimand

To estimate the estimand of interest, we propose to apply logistics regression model to predict the membership of PS_0 and the TMLE estimator to estimate $\theta_{ps}(P_0)$.

4.5.5 Obtain Estimate, Uncertainty Measurement, and Inference

The R function "glm" can be used to estimate the PS_0 membership of each subject, using the covariates that are believed to be predictive for the membership and the following R codes:

```
# Fit a logistic regression model with ICE as outcome
fit.ice = glm(formula = ice ~ qsmk + age + wt71,
  data = data0, family = binomial)

# Predictive probability of ICE in the hypothetical world
data0.new <- data0
data0.new$qsmk <- ifelse(data0$qsmk, 0, 1)
prob.new <- predict(fit.delta, data0.new, type="response")

# Use marginal ICE rate 0.3498 as threshold for prediction
threshold <- quantile(prob.new, 1 - 0.3498)
pred.new <- (prob.new>threshold)

# Obtain estimated PS0, which has 920 subjects
data0.ps <- data0[(1-delta)&(1-pred.new),]
nrow(data0.ps) # 920
```

The R function "tmle" of R package "tmle" can implement the proposed method with the estimated PS_0, using the following R codes:

```
nhefs_tmle_fit.ps <- tmle(Y=data0.ps$wt82_71, A=data0.ps$qsmk,
  W=data0.ps[,W.vec],
  Q.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
       randomForest"),
  g.SL.library = c("SL.glm", "SL.rpart", "SL.glmnet", "SL.
       randomForest"),
  family = "gaussian")

nhefs_tmle_fit.ps$estimates$ATE$psi
nhefs_tmle_fit.ps$estimates$ATE$CI
```

From the output, the point estimate of $\theta_{ps}(P_0)$ is 3.22, with SE $= 0.67$ and 95% CI $= (1.90, 4.55)$.

4.5.6 Conduct Sensitivity Analysis and Interpret the Results

Two main assumptions are made in the primary analysis: (1) the NUC assumption and (2) the underlying assumption made implicitly in the prediction of the membership of PS_0 according to the marginal estimate of the proportion of subjects with ICEs in the hypothetical world in which subjects had taken the alternative treatment, denoted as p_e, which is 34.98%. We omit the application of E-value for the NUC assumption. Here we demonstrate sensitivity analysis assuming different proportions than $p_e = 34.98\% \doteq 0.35$.

For this aim, we select 7 values for p_e around 0.35, say, from 0.2 to 0.5 by 0.05. For each value of p_e, we have an estimated PS_0 and the corresponding estimate of $\theta_{ps}(P_0)$. Figure 6 displays 95% CIs of $\theta_{ps}(P_0)$ corresponding to 7 values of p_e,

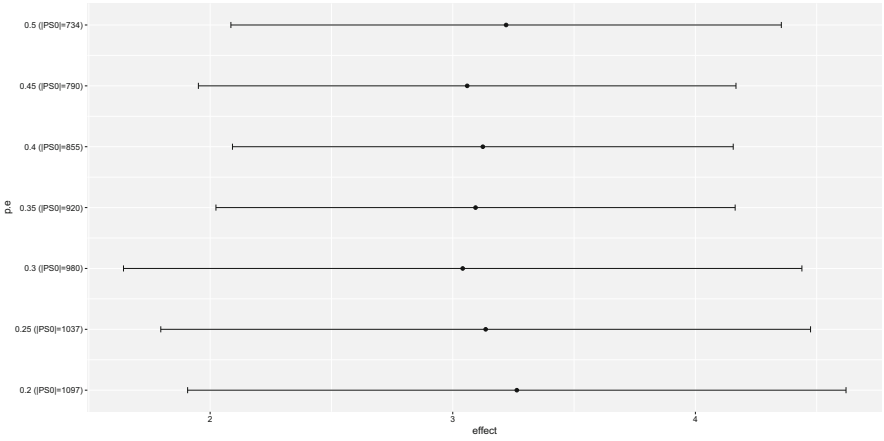

Fig. 6 Sensitivity analysis to a range of values of p_e

Table 2 Summary of features of five strategies

Strategy	Variables	Key feature	PROTECT
Principal stratum	(W, A, Y)	Principal stratum PS_0	"P"
Composite variable	$(W, A, Y^*(Y, E))$	Composite outcome Y^*	"R/O"
Treatment policy	$(W, A^*(A, E), Y)$	Treatment regime A^*	"T/E"
While on treatment	$(W, A, Y_{t(E)})$	Outcome measured at $t(E)$	"T"
Hypothetical	$(W, A, Y^{E=0})$	Counterfactual outcome $Y^{E=0}$	"C"

along with the size of each estimated PS_0 showed on the y-axis, where $p_e = 0.35$ is corresponding to the main result. We see all the intervals are similar to each other. Hence we can conclude the result is robust to a wide range of different versions of the estimation of PS_0.

5 Summary

We demonstrate the application of the causal-inference roadmap, which consists of six key steps, using seven real-data examples. The first two examples are for the ATE estimand and the ATT estimand, respectively. In these two examples, we demonstrate the application of the roadmap to deal with the confounding bias in non-randomized real-world studies.

The remaining five examples are corresponding to five strategies of ICE handling. In these examples, we demonstrate the application of the roadmap to deal with both the confounding bias due to non-randomization and the challenge due to the existence of ICEs. Table 2 provides a summary of the features of these five examples. The first column is the strategy in each example. The second column is the vector of confounders, treatment variable, and outcome variable in each example. The third and fourth columns are the key feature and the corresponding PROTECT element as discussed in chapter "Key Considerations in Forming Research Questions and Conducting Research in Real-World Setting" of this book.

References

1. FDA Guidance Document (2021): E9(R1) Statistical Principles for Clinical Trials; Addendum: Estimand and Sensitivity Analysis in Clinical Trials, https://www.fda.gov/media/148473/download
2. Imbens, G., Rubin, D.: Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, (2015)
3. Ho, M., Laan, M., Lee, H., Chen, J., Lee, K., Fang, Y., He, W., Irony, T., Jiang, Q., Lin, X., Meng, Z., Mishra-Kalyani, P., Rockhold, F., Song, Y., Wang, H., White R.: (2021) The current landscape in biostatistics of real-world data and evidence: Causal inference frameworks for study design and analysis. *Statistics In Biopharmaceutical Research*. pp. 1–14 (2021)

4. Fang, Y., Wang, H., He, W.: A statistical roadmap for journey from real-world data to real-world evidence. *Therapeutic Innovation & Regulatory Science*. **54**, 749–757 (2020)
5. Petersen, M., Laan, M.: Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology*. **25**, 418 (2014)
6. Hernan, M., Robins, J: Causal Inference: What If. Boca Raton: Chapman & Hall/CRC (2020)
7. van der Laan, M., Rose, S.: Targeted Learning: Causal Inference for Observational and Experimental Data. Springer (2011)
8. VanderWeele, T., Ding, P.: Sensitivity analysis in observational research: introducing the E-value. *Annals Of Internal Medicine*. **167**, 268–274 (2017)
9. Rosenbaum, P., Rubin, D.: The central role of the propensity score in observational studies for causal effects. *Biometrika*. **70**, 41–55 (1983)
10. Austin, P.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*. **46**, 399–424 (2011)
11. O'Kelly, M., Ratitch, B.: Clinical Trials with Missing Data: A Guide for Practitioners. John Wiley & Sons (2014)
12. Rubin, D.: Multiple imputation after 18+ years. *Journal Of The American Statistical Association*. **91**, 473–489 (1996)

Applications Using Real-World Evidence to Accelerate Medical Product Development

Weili He, Tae Hyun Jung, Hongwei Wang, and Sai Dharmarajan

1 Introduction

In recent years, we have seen an increasing usage of RWE/RWD in clinical development and life-cycle management. Especially encouraged by legislations and guidance released by regulators and special interest groups, sponsors have been actively seeking guidance and application use cases. Sponsors are now faced with the tasks of determining regulatory contexts where RWD can be used for regulatory decisions, proposing appropriate RW study designs to address specific research questions, including determining and assessing fit-for-purpose data sources, be prospectively collected or using existing RWD sources, using appropriate statistical methods to determine causal inference from the RWD, and seeking regulatory guidance and decisions on the application.

Although in the last few years there have been a few RWE/RWD use cases, there is still a paucity of real examples in literature, especially lacking in-depth analyses of such cases in a systemic manner for practitioners to draw best practices and lessons learned. In Sect. 2, we describe six real examples with information available in the public domain, where the background of diseases and case studies along with the study findings are provided. Where information in the public domain is available, our analysis in Sect. 3 is devoted to delineating the regulatory contexts, key regulatory review issues, whether the use of RWE/RWD is pivotal or supplemental for the regulatory decisions, assessment of fit-for-use data sources, statistical methods employed, and whether substantial evidence of effectiveness as

W. He (✉) · H. Wang
Medical Affairs and Health Technology Assessment Statistics, Data and Statistical Sciences, AbbVie, North Chicago, IL, USA
e-mail: weili.he@abbvie.com

T. H. Jung · S. Dharmarajan
Office of Biostatistics, CDER, Food and Drug Administration, White Oak, MD, USA

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_19

stated in Regulations 21 CFR 314.126 is met for the specific case study. Section 4 is devoted to discussions summarizing the lessons learned and best practices from these case studies. The final section provides concluding remarks.

2 RWE/RWD Case Studies by Regulatory Purposes

In this section, we provide background of six RWE/RWD use cases for which we will provide further analysis on the rationale of regulatory decisions in Sect. 3. We categorize these use cases by regulatory purposes.

2.1 RWE/RWD as Part of the Original Marketing Application

2.1.1 Avelumab

Merkel cell carcinoma (MCC) is a rare, aggressive skin cancer with 5-year survival rate of 78% for localized disease, 52% for regional disease, and 19% for advanced or metastatic disease [1]. MCC is chemo-sensitive, and patients with newly diagnosed metastatic MCC (mMCC) can achieve objective response rates (ORRs) between 50% and 60%. However, early development of resistance to chemotherapy leads to shorter duration of response (DOR) with median around 85 days [2].

Avelumab is an anti-PD-L1 monoclonal antibody developed for the treatment of mMCC. It received accelerated approval from the FDA in 2017 as the first therapy in this patient population. The approval is based on JAVELIN Merkel 200 trial which was an open label, single arm, and multicenter Phase 2 study. It enrolled mMCC patients with disease progression after prior chemotherapy. The primary endpoint was confirmed ORR per Response Evaluation Criteria in Solid Tumors V.1.1 and secondary endpoints included DOR, progression-free survival, overall survival, and safety. The efficacy analyses were conducted based on Part A when all patients had a minimal 12-month follow-up. Out of the 88 patients enrolled, the centrally adjudicated ORR was 33% (95% Confidence Interval: 23%, 44%). Among the 29 patients that responded, the median DOR was not reached, meaning that $=>50\%$ of study patients (72%) continued to have response. This treatment effect was consistent across clinically important subgroups [3].

To quantify the natural history of disease mainly ORR and DOR under standard of care, a retrospective chart review study in the US Oncology Network (USON) was included in the submission package. USON includes about 1 million cancer patients from over 470 sites across 25 states. A total of 39 patients with potential mMCC who received second-line chemotherapy were identified using the oncology-specific iKnowMed electronic health record (EHR). Of those, 14 were determined to meet similar inclusion/exclusion criteria of JAVELIN. These 14 patients had an ORR of 28.6% (95% CI: 8.4%, 58.1%) and median DOR of 1.7 months (95%: 0.5,

3.0) [3]. Although the sample size is small, the findings are consistent with other literatures showing transient response.

The accelerated approval of avelumab covers both adults and pediatric patients 12 years and older. It does not require patients to receive prior systemic therapies either. This is based on the biology of mMCC, pharmacokinetics, and pharmacodynamics (PK/PD) modeling and lack of therapy in this life-threatening disease. Confirmatory trial was required to verify the avelumab's clinical benefit in the extrapolated patient population, that is, pediatric patients 12–18 years old and patients who have not received systemic therapies, as part of the approval.

2.1.2 Tafasitamab

Diffuse large B-cell lymphoma (DLBCL) is the most common high-grade non-Hodgkin lymphomas (NHLs). The first-line standard of care regardless of stage is the combination therapy R-CHOP (rituximab [Rituxan], cyclophosphamide [Cytoxan], doxorubicin [Adriamycin], vincristine [Oncovin], and prednisone). Majority of relapsed or refractory (R/R) DLBCL patients are not eligible to receive intensive immunochemotherapy or autologous stem cell transplantation (ASCT) and have survival times ranging from 6 to 12 months, representing a high unmet medical need.

Tafasitamab is an Fc-modified antibody that binds to CD19 antigen. It received the designations of fast-track review, orphan drug, and breakthrough therapy. Its combination with lenalidomide received accelerated approval from FDA for the treatment of adult R/R DLBCL ineligible for ASCT in 2019. The approval is based on favorable benefit-risk profile observed from the open-label, single-arm, Phase 2 L-MIND study. A total of 71 patients out of 81 enrolled had confirmed DLBCL, received at least one dose of both drugs and formed the primary efficacy analysis population. The primary objective of best overall response rate (BORR) was 55% (95% CI: 43–67%) where 37% of patients had a complete response (CR). The key secondary objective of median DOR was 21.7 months [4]. This efficacy profile was confirmed in the long-term L-MIND study with $\geq$35 months follow-up where ORR was 57.5% including a CR of 40.0%, median DOR was 43.9 months and median overall survival was 33.5 months [5]. The contribution in efficacy effect from tafasitamab in this population was also confirmed by MOR208C201. In this Phase 2a study, ORR was 26% (95% CI: 13–43%) among 35 patients received the single agent of tafasitamab [6].

Results from L-MIND of the combo therapy were also considered in the context of a retrospective observational cohort study MOR208C206 (RE-MIND) where all patients received a single agent lenalidomide. Key eligibility criteria of RE-MIND were aligned with L-MIND and the observed ORR was 32% (95% CI: 21–43%). However, FDA deemed "formal statistical comparisons unfeasible" because of several limitations [4]. A recommendation of post market requirement (PMR) to confirm the benefit of tafasitamab in a RCT was issued.

2.2 RWE/RWD as Primary Data Source for Label Expansion

2.2.1 Prograf

Prograf (tacrolimus) is an immunosuppressant indicated for the prevention of organ rejection in adult and pediatric patients receiving transplantation. The FDA first approved Prograf in 1994 for the prevention of organ rejection in liver transplantation and then approved for kidney (1997) and heart (2006) transplantations based on scientific evidence from randomized clinical trials. Prograf has been used in combination with other immunosuppressants such as mycophenolate mofetil (MMF) or azathioprine (AZA). Until the FDA approved Prograf for the prevention of rejection of lung transplantation on July 16, 2021, no immunosuppressant had been approved for this indication. However, many patients had been treated with Prograf off-label of which approximately 86% of lung transplant recipients are treated with Prograf, mostly in combination with MMF [7].

The primary objective of the study was to evaluate transplant-related outcomes and use of tacrolimus immediate release (TAC IR) and other immunosuppressive agents over time in United States [8]. The applicant conducted a non-interventional study using RWD from the US Scientific Registry of Transplant Recipients (SRTR) database to evaluate transplant-related outcomes and the use of Prograf including other immunosuppressants in lung transplant recipients [8]. From the SRTR database, the study identified patients who received a primary lung transplantation (not re-transplantation) between January 1, 1999, and December 31, 2017. The primary efficacy endpoint was the composite endpoint of time to graft failure (GF) or death (due to any cause) within 1 year after transplantation. Patients were followed for up to 3 years post transplant. The study was descriptive and used the Kaplan–Meier estimates of the cumulative incidence. The primary efficacy endpoint of 1-year graft failure or death was used for primary analysis.

The study identified 15,478 adults and 450 pediatric patients receiving Prograf in combination with MMF [8]. The 1-year graft survival estimates from time of discharge were 90.9% (adult) and 91.7% (pediatric), respectively. The study also identified 4263 adults and 72 pediatric patients receiving Prograf in combination with AZA. The 1-year graft survival estimates from time of discharge were 90.8% (adult) and 84.7% (pediatric). Although this study did not include a comparator, these outcomes were very unlikely to have occurred by chance alone compared to the adequately documented natural history of a transplanted lung with no or minimal immunosuppressant, which yielded no graft or overall survival at 1 year and survival was largely limited to several weeks [9, 10].

In this study, the applicant generated RWE by the analyses of RWD extracted from the SRTR database [8]. SRTR is a national transplant registry operated under a federal contract by Hennepin Healthcare Research Institute [11]. This registry has a well-established and robust operational structure for collecting rigorous data on all US solid-organ transplant recipients, as required by the National Organ Transplantation Act of 1984 [12]. SRTR directly collects the primary source of data

from the Organ Procurement and Transplantation Network (OPTN), and they are supplemented by data from the Centers for Medicare & Medicaid Services (CMS) and the National Technical Information Service's (NTIS) Death Master File [11]. SRTR releases Standard Analytic files (SAF) for bona fide research or analysis purpose under a Data Use Agreement, and it contains data element on all solid organ transplant candidates, recipients, and donors in the United States from Oct 1, 1987, to the present. For this application, the applicant submitted SAF including patients receiving primary lung transplantation from Jan 1, 1999, to Dec 31, 2017 [8].

2.2.2 SurgiMend

Acellular dermal matrices (ADMs) have become a vital part of breast reconstruction procedures because they address the tissue deficiencies resulting from mastectomy: ADM provides reinforcement for weakened tissue, supplements thin and overly dissected tissue, and repairs the breast boundaries that were eliminated during the procedure. Although previous investigators have evaluated its risks, few studies have assessed the impact of ADM on other outcomes, including patient-reported measures.

The MROC Study (Mastectomy Reconstruction. Outcomes Consortium) was a prospective, observational cohort study and included women undergoing first-time reconstruction following mastectomy, for breast cancer treatment or prophylaxis, from 11 centers and 58 participating surgeons in the United States and Canada [13]. Eligible patients were enrolled between 2012 and 2015 and included those undergoing tissue expander placement for immediate unilateral or bilateral reconstruction following mastectomy for breast cancer treatment or prophylaxis. All patients subsequently underwent expander exchange for saline-or silicone-filled reconstructive implants. Study patients were divided into two cohorts: (1) those undergoing expander reconstruction with ADM and (2) those receiving expander reconstruction without ADM. After obtaining informed consent, patient demographic and clinical information was gathered from electronic medical records (EMRs) by the site coordinators and included age, body mass index (BMI), laterality (unilateral vs. bilateral), indication for mastectomy (treatment vs. prophylactic), mastectomy type (nipple sparing, simple or modified radical), smoking status, diabetes, lymph-node management, adjuvant chemotherapy, and radiation. The MROC study was funded and run by the U.S. National Cancer Institute (NCI). Data collection relied on protocol-specified Patient Reported Outcome (PRO) measures, including the BREAST-Q—a validated PRO instrument designed specifically for patients who undergo breast reconstruction surgery [14, 15]—and other PRO questionnaires, as well as data in the EMRs and billing records. Patients completed questionnaires, including PRO instruments, before surgery and at 1 week, 3 months, 1 year, and 2 years after initial surgery. Postoperative complications, which were prespecified in the protocol, were retrospectively identified in the EMR at 1 year and 2 years after

the subject's breast reconstruction. As of September 1, 2021, multiple peer-reviewed publications have been published based on the MROC Study [13].

SurgiMend Acellular Bovine Dermal Matrix for Soft Tissue Reconstruction is intended for implantation to reinforce soft tissue where weakness exists and for the surgical repair of damaged or ruptured soft tissue membranes. Due to enrollment challenges, the sponsor (Integra LifeSciences Corporation) and FDA agreed that RWE could provide the best path forward to demonstrate the safety and effectiveness of SurgiMend PRS ABDM in breast reconstruction surgeries [16]. Integra entered discussions with FDA to consider the use of RWE generated from the RWD in the MROC Study sponsored by the University of Michigan. The SurgiMend study was an analysis of a subset of MROC study data using a prospectively developed analysis plan to compare SurgiMend vs No-Acellular Dermal Matrix (No-ADM) in immediate, two-stage, submuscular implant-based breast reconstruction (IBBR). The inclusion/exclusion criteria were incorporated for subject selection from the raw MROC datasets, so that the treatment group would include only subjects who underwent immediate, two-stage submuscular IBBR with the use of SurgiMend, and the control group would include subjects who underwent immediate, two-stage submuscular IBBR with total submuscular coverage, i.e., no ADM. Among 4306 MROC study subjects enrolled from January 2012 to February 2016, per the pre-specified inclusion/exclusion criteria, 1792 subjects were identified to have undergone immediate, two-stage submuscular IBBR, among which 987 subjects were treated with either SurgiMend or control No-ADM and were selected into the SurgiMend study. There were 119 subjects from two investigational sites in the treatment group (SurgiMend) and 868 subjects from nine investigational sites in the control group (noADM).

The pre-specified primary endpoint in this study was the composite clinical success (CCS). A subject achieves the composite clinical success if both two criteria are satisfied: (1) an assessment of BREAST-Q Physical Well-Being, Chest score $\geq (-4)$ point change from baseline at 1-year post implant and (2) absence of major complications through year 2 or through year 1 (if year 2 data are not available). Propensity score-based stratification approach was used to reduce potential confounding to estimate the average treatment effect on the treated (ATT) using weights based on the number of SurgiMend treated subjects within each propensity score stratum [17]. Multiple imputations were used to handle missing data issues in the analyses of the primary and secondary endpoints with the propensity score stratification method.

Using the pre-specified primary ATT approach, the primary estimated CCS rate was 32.4% for the SurgiMend group and 21.1% for the control group. The estimated difference for CCS rate between the SurgiMend and control groups was 11.2% with a 95% confidence interval of (1.7%, 20.8%), excluding 0. The CCS rate for SurgiMend group was statistically significantly higher than that for the control group with a two-sided p-value of 0.02. For the primary endpoint CCS, approximately 25% of data are missing. The Breast Q—Physical well-being, chest at year 1 had 44.1%

missing data for No ADM control and 34.5% for SurgiMend group. The Breast Q—Physical well-being, chest at year 2 had 62.9% missing data for No ADM control and 58% for SurgiMend group. Additional details can be found in SurgiMend FDA Briefing document 2021 [16].

2.3 RWE/RWD as One of the Data Sources for Label Expansion

2.3.1 Orencia

Acute graft-versus-host disease (aGVHD) is a potentially fatal complication that can occur after stem cell transplantation when the donor's immune cells (the graft) view the recipient's body (the host) as foreign and the donated cells attack the body. The chances of developing aGVHD increase when the donor and recipient are not related or are not a perfect match. Severe (grade 3–4) aGVHD is a major cause of death after unrelated-donor (URD) hematopoietic cell transplant (HCT), resulting in particularly high mortality after human leukocyte antigen (HLA)-mismatched transplantation. There are no approved agents for aGVHD prevention, underscoring the critical unmet need for novel therapeutics.

Abatacept, sold under the brand name Orencia, is a medication used to treat autoimmune diseases like rheumatoid arthritis, by interfering with the immune activity of T cells.

GVHD-1 trial, also known as ABA2, was a phase II trial to rigorously assess safety, efficacy, and immunologic effects of adding T-cell costimulation blockade with abatacept to calcineurin inhibitor (CNI)/methotrexate (MTX)-based GVHD prophylaxis, to test whether abatacept could decrease aGVHD. Watkins et al. (2021) reported details in the study design, methods, and results [18]. To summarize the study outcome, the study measured severe (grade III–IV) aGVHD-free survival, overall survival and moderate-severe (grade II–IV) aGVHD-free survival 6 months after transplantation. The study reported 73.6% (80% confidence interval [CI]: 62.0–82.2) overall survival (OS) at 2 years in recipients of a 7/8 HLA-single mismatch unrelated donor (7/8 MMUD) HSCT following treatment with abatacept + a standard aGVHD prophylaxis regimen (calcineurin inhibitor [CNI] + methotrexate [MTX] without [−] antithymocyte globulin [ATG]), compared with 45.3% (80% CI: 39.3%–51.1%; $P = 0.002$) in a standard treatment cohort (CNI + MTX − ATG) of matched controls from the Center for International Blood and Marrow Transplant Research (CIBMTR) [19].

GVHD-2 was a real-world study to further evaluate the impact of abatacept on survival of 7/8 MMUD HSCT recipients, treated with CNI + MTX − ATG with or without abatacept, from CIBMTR database of all allogeneic HSCTs performed in the United States in recent years [20]. In this observational study, patients (≥ 6 years of age with leukemia, lymphoma, or myelodysplastic syndrome, whose first allogenic HSCT was with a 7/8 MMUD between 2011 and 2018) were treated with

CNI + MTX − ATG with or without abatacept. OS (defined as time between date of transplant and documented date of death) was evaluated at 181 days post-transplant by weighted log-rank test with inverse propensity scores, obtained using logistic regression models as weights (inverse probability of treatment weighting [IPTW]) to reduce bias due to confounding. The marginal hazard ratio (HR) and 95% CI were estimated by a weighted Cox proportional hazards model. Exploratory analyses of OS were evaluated in 7/8 MMUD HSCT recipients treated with abatacept + CNI + MTX (without ATG) versus CNI + MTX + ATG, and versus those treated with post-transplant cyclophosphamide-based (PT-Cy) GVHD prophylaxis.

For the primary analysis, 216 patients (54 [25%] abatacept + CNI + MTX − ATG and 162 [75%] CNI + MTX − ATG) were included. Key patient demographics and characteristics were generally similar across treatment. Most patients were male and had performance scores of 90–100; had acute myeloid leukemia; and had received myeloablative conditioning, peripheral blood stem cell grafts, and tacrolimus. Kaplan–Meier OS rates at day 181 post-transplant by weighted log-rank test with inverse propensity scores (95% CI) were 98% (78–100%) for patients treated with abatacept + CNI + MTX − ATG and 75% (67–82) for those treated with CNI + MTX − ATG ($P = 0.0028$). The marginal HRs (95% CI) were 0.06 (0.01–0.27) and 0.07 (0.01–0.30) using treatment only and treatment plus disease status as covariates, respectively.

2.4 RWE/RWD as Supplemental Information for the Regulatory Decision

2.4.1 Ibrance

In terms of cancer death, breast cancer is the second leading cause among women and the fourth leading cause overall. Metastatic breast cancer (MBC) is incurable and the treatment of patients with MBC is palliative in nature, aiming to prolong survival, and/ or improve disease-related symptoms. While male breast cancer is rare, it is a serious and life-threatening condition that was estimated to affect 2620 men and kill nearly 520 in 2020 [21]. As the prognosis for men is similar to that for women with a comparable stage of disease [22], current clinical practice standards for the treatment of male patients with breast cancer mirror those for women with breast cancer [23]. However, as of 2019, there were no therapies approved specifically for the treatment of male patients with MBC.

For patients with hormone receptorpositive, human epidermal growth factor receptor 2 (HER2)-negative metastatic breast cancer, endocrine therapy represents the main initial therapeutic strategy [23] . Other treatment options for these patients include endocrine therapy in combination with mTOR inhibitors (everolimus) or an inhibitor of cyclin-dependent kinases (CDK) 4 and 6 inhibitors such as palbociclib, ribociclib, and abemaciclib [23].

Palbociclib (IBRANCE®), originally granted accelerated approval on February 3, 2015, for use in combination with letrozole in postmenopausal women was granted regular approval on March 31, 2017 as initial endocrine-based therapy [24]. The drug's indication was expanded from allowing palbociclib in combination with only letrozole to allowing it in combination with any aromatase inhibitor [24].

Updated results from 666 patients in the randomized, double-blind, placebo-controlled phase 3 trial (Study PALOMA-2) in women with HR-positive, HER2-negative advanced or metastatic breast cancer whose disease was not previously treated continued to demonstrate a clinically meaningful benefit for Palbociclib [25]. The estimated median PFS in the palbociclib plus letrozole arm was 27.6 months (95% Cl = 22.4, 30.3) compared to 14.5 months (95% Cl: 12.3, 17.1) in the placebo plus letrozole arm (HR = 0.563 95% Cl: 0.461, 0.687; $p < 0.001$) [25].

However, male patients with breast cancer were ineligible to participate in PALOMA-2 and in studies that provided the data to support prior approvals of Palbociclib. Therefore, the applicant provided the results of an analysis of RWD from EHRs from the Flatiron Health Analytic Database (FHAD), to support the request for broadening the palbociclib indication to include male patients [23]. The Flatiron Database is generated from the EHR data that is collected within the Flatiron Provider Network of cancer care providers in the United States and includes cancer patients who are actively receiving treatment [23]. Data were de-identified and provisions put in place to prevent re-identification before a retrospective cohort study was initiated. The study included males aged 18 years or older with the following inclusion criteria met in the study period from January 1, 2011 to July 31, 2017: an ICD diagnosis of breast cancer; two or more clinical visits; an MBC diagnosis and HR-positive/HER2-negative disease (in a 60-day window before or after MBC diagnosis date), both confirmed through unstructured data. Two cohorts of patients were defined to compare with respect to clinical characteristics and outcomes. Patients in Cohort A (palbociclib-treated cohort) received a palbociclib-based regimen in any line of therapy (LOT), whereas patients in Cohort B (non-palbociclib-treated cohort) received an endocrine therapy-based regimen in any LOT and were never treated with a palbociclib-containing regimen [23].

The primary outcome of interest for this study was real-world response, defined as the treating clinician's assessment of radiological evidence for change in burden of disease over the course of treatment with a given LOT [23]. The assessment was mapped to one of the following categories: CR, partial response, stable disease, or progressive disease. Only 12 patients in cohort A and 29 patients in cohort B had radiological follow-up visits necessary to make on-study tumor assessments. Furthermore, in Cohort B, 13 patients whose endocrine therapy only included a tamoxifen agent were excluded from the analysis, as tamoxifen is not approved for use in combination with palbociclib, decreasing the size of Cohort B to 16 patients. The observed response rates were 25% (3/12) in Cohort A and 12.5% (2/16) in Cohort B. No formal statistical comparisons were made [23].

3 Analysis of Key Considerations in the Regulatory Decisions

In this section, to understand the contexts for the decision, the past regulatory interactions and approval status for each use case will be described. Furthermore, we will provide analysis of key considerations on regulatory decisions. These may include but not limited to the regulatory contexts, regulatory quality data source, statistical methods employed to minimize potential biases and confounding, and any regulatory opinions for the submission and regulatory decision.

3.1 RWE/RWD Supporting the Original Marketing Application

3.1.1 Avelumab

MCC is a rare skin cancer and each year about 2000 patients were diagnosed with it in the United States [1]. Metastatic MCC is incurable, and there was no FDA-approved therapy when avelumab was under review. Coupled with poor 5-year survival rate, there is a high unmet medical need. These factors contributed to the avelumab's orphan drug designation.

The real-world study that was included in the avelumab FDA submission came from USON which includes about one million cancer patients annually. The iKnowMed EHR system was specifically designed for oncologists and hematologists. It tracks outpatient encounter across the practice of 1200 affiliated physicians. The relevant patients' characteristics and clinical outcomes can be abstracted from the structured data fields as well by chart review from unstructured data. It can also be linked with external data source such as the Social Security Administration's Limited Death Master File (SSALDMF) to retrieve patients' vital status.

The real-world study identified 39 potential eligible mMCC patients where 14 were included in the final analysis to mimic comparable JAVELINE Merke 200 trial enrollees. The FDA review recognized the small sample size and inherent selection bias in using historical control data. It concluded that the data were exploratory in nature and was considered only to further characterize the natural history of disease in target population for benefit-risk assessment. This is a challenge in using RW historic control data especially for rare disease. However, the findings are consistent with the limited literatures [2] which reported a median DOR of 101 days based on 30 distant metastatic MCC patients who received second-line chemotherapy. FDA review did indicate that the single-arm study "demonstrated a clinically meaningful ORR that was significantly more durable than response rates observed for salvage chemotherapy, which is the current treatment standard." And for the management of mMCC, durable ORR is considered to be a valid surrogate endpoint for clinical benefit such as survival, how patients function or feel. There was no advisory committee meeting before FDA granted avelumab accelerated approval.

Following approval of avelumab for mMCC, a more recent publication [26] reported clinical outcomes in patients with locally advanced (la) or mMCC initiating first-line avelumab from the same RW data source USON. All patients were required to have a minimal of 6-month follow-up or evidence of death at the time of analysis. A total of 9 laMCC and 19 mMCC patients were identified with response rates of 66.7% and 63.2%, respectively. This patient population is similar to those enrolled in JAVELIN Merkel 200 Part B, which is stage IV mMCC patients who had not received prior systemic therapy. A total of 39 patients were enrolled in Part B and a pre-planned interim analysis based on 29 patients with at least 3-month follow-up reported similar ORR of 62.1% [27]. After linking with SSALDMF, this newer RW study reported that median progression-free survival and overall survival were not reached in laMCC and were 10.0, 20.2 months in mMCC, respectively. These results showed clinical benefit of avelumab in US RW clinical setting with the limitation of small sample size.

3.1.2 Tafasitamab

The unmet medical need for R/R DLBCL is high where limited treatment options exist. As the accelerated approval is based on a single-arm Phase 2 study where all patients received the combination therapy of tafasitamab and lenalidomide, a real-world study RE-MIND was included in the FDA submission package for comparison with monotherapy of lenalidomide. The real-word data was retrospectively collected from health records in academic hospitals, public hospitals, and private practice in North America, Europe, and Asia Pacific region using electronic data capture. The eligibility criteria mimic that of the L-MIND study "patients aged $\geq$18 years with histologically confirmed DLBCL and who had received $\geq$2 prior systemic therapies for R/R DLBCL (including $\geq$1 anti-CD20 therapy)." The primary endpoint follows similar definition which was best ORR as assessed by the investigator.

During a dedicated FDA Type C meeting for the use of RWD in this submission, the following limitations were reported [4].

1. Data quality: Specifically, there may be systematic differences in type of data being collected, covariates (measured or unmeasured), validity of outcome assessment, amount of missing, and duration of follow-up. Evidence to support RWD collection being adequate, accurate, and non-differential need to be demonstrated.
2. Data completeness: As much as possible, clinically important covariates should be included comprehensively. Further imputation of missing data could not be accepted for the purpose of estimating propensity score. The agency recognized that this would result in sample size attrition but felt that it is necessary to enable assessment of comparability.
3. Population comparability: Only patients from comparable geographic regions and relevant initial dose of lenalidomide at index date should be included

in the monotherapy cohort. And the overlap weight-based approach was not recommended.

To address these concerns, the following revisions were adopted and subsequently reported in Zinzani et al. [28]:

1. Only study sites from the EU and the United States were selected to be consistent with L-MIND. Similarly, only patients initiating a lenalidomide dose of 25 mg/day were included.
2. Only patients with complete data on nine prespecified baseline covariates of clinical importance were included. A tenth covariate of Easter Cooperative Oncology Group performance status (ECOG PS) was included as a prespecified sensitivity analysis.
3. Outcome of ORR was validated for a subset of patients by an independent committee following both radiological and clinical review. To address the intercurrent event of treatment change, the ORR status has to be assessed between lenalidomide initiation and starting a new anti-DLBCL medication or death.
4. To enable accurate assessment of response rate, all patients were required to have a minimal 6-month follow-up period. The 6-month requirement was met if a patient responded to treatment or progressed or died at any time within 6 months of treatment initiation without a documented response. Patients with unknown response status or nonresponding with first tumor assessment being after 6 months were excluded from lenalidomide cohort.
5. Propensity score matching with 1:1 ratio and nearest neighbor algorithm was employed to balance the two comparison groups. Standardized mean difference (SMD) with threshold of 0.2 was prespecified to assess balance after matching.
6. Standard statistical inference such as Fisher's exact test, logistic regression, Kaplan–Meier curve, and log-rank test was conducted to compare the two matched cohorts for categorical and time-to-event outcomes.
7. Multiple sensitivity analyses were included to assess the robustness of results, including adoption of doubly robust method to relax underlying assumptions and address residual imbalance.

Among 524 potential patient charts collected from RWD source, a total of 140 fulfilled all inclusion criteria. Out of the 81 patients enrolled in L-MIND, 76 patients met the minimal 6-month follow-up requirement and received the combination therapy. All 76 patients were successfully matched with a control. The two comparison cohorts were largely comparable in terms of baseline characteristics with seven out of nine prespecified covariates having SMD < 0.2. Best ORR was 67.1% (95% CI: 55.4–77.5%) for the combination therapy versus 34.2% (95% CI: 23.7–46.0%) in the lenalidomide monotherapy. Fisher's exact test indicates high statistical significance ($P < 0.0001$). Among patients who responded, DOR was 20.5 versus 6.6 months in the combination and monotherapy cohorts, respectively. Sensitivity analyses confirmed the findings from primary analyses.

Overall, the real-world study RE-MIND was judged to have many limitations and only provided contextual evidence in the original FDA approval. The subsequent publication incorporating the agency comments appears to support the incremental benefits of the combination therapy versus lenalidomide monotherapy.

3.2 RWE/RWD as the Primary Data Source for Label Expansion

3.2.1 Prograf

Since the initial approval in 1994, Prograf has been used as the mainstay of the immunosuppressive regimens in most transplant recipients for the approved indications of liver, kidney, and heart transplantations [29]. Off-label use of Prograf in lung transplantation has increased since 1994 and gradually replaced cyclosporine as the calcineurin inhibitor of choice. According to the 2018 US Annual Data Report published by the SRTR, approximately 85% of the lung transplant patients were treated off-label, with tacrolimus, MMF, and corticosteroids [7]. No immunosuppressant was approved for lung transplant recipients before the FDA approval of Prograf in prevention of rejection of lung transplantation in July 2021.

The SRTR was the primary source of RWD used in the non-interventional study to demonstrate the effectiveness and safety of Prograf use for the new indication. The SRTR were considered fit-for-use because the data was relevant and reliable to the regulatory research questions [29]. The SRTR has a well-established and robust operational structure that captured all solid-organ transplant recipients in United States. As a compulsory registry, the SRTR collects all U.S. population of lung transplant recipients [30]. Thus, limitations related to a non-representative study population did not apply in this submission. Also, the granular capture of relevant clinical variables on each patient regarding transplantation and graft survival outcomes enhanced the relevance of the data. The accuracy of the SRTR could be improved through a linkage with external sources such as CMS and NTIS Death Master File. In this submission, NTIS Death Master File supplemented the SRTR death outcomes as a trusted repository of mortality data thus the primary outcome of interest, death, was considered reliable. The comprehensive capture of clinical information resulted in low percentage of missing data, which enhanced the reliability of the data.

The FDA stated that this regulatory approval "reflects how a well-designed, non-interventional study relying on fit-for-purpose RWD, when compared with a suitable control, can be considered adequate and well-controlled under FDA regulations [31]." This implies that even a non-interventional study that might have challenges from confounding and other biases can provide robust scientific findings comparable to a randomized clinical trial (RCT) and meet the regulatory requirements depending on the fit-for-use data source (including robust data collection method), study design, statistical analysis approach in minimizing potential confounding

biases. An evident improvement in outcomes was observed among lung transplant patients receiving Prograf in combination with other immunosuppressants and these outcomes were very unlikely to have occurred by chance alone compared to the previous natural history of a transplanted lung with no or minimal immunosuppressants, where no patients survived to 1 year and the median survival was limited to several weeks [9, 10].

In addition to the RWE generated from the non-interventional study, confirmatory evidence of efficacy comes from randomized controlled trials in other solid organ transplants. Additional support for the application came from existing journal publications relevant to lung transplantation [29].

3.2.2 SurgiMend

In the past decade prior to 2017, surgeons have begun utilizing surgical mesh products to assist with breast reconstructive procedures, and ADM mesh products are now used in most implant-based breast reconstruction procedures in the United States [13]. However, the FDA has not cleared or approved any surgical mesh device—whether synthetic, animal collagen derived, or human collagen derived—specifically indicated for breast reconstruction. In March 2019, the FDA's General and Plastic Surgery Advisory Committee discussed the evidentiary requirements needed to assess surgical mesh benefit versus risk in breast reconstruction. Trial design considerations identified by FDA at the March 2019 Advisory Committee meeting as critical for assessing surgical mesh of device safety and effectiveness for breast reconstruction [32].

SurgiMend PRS Acellular Bovine Dermal Matrix, indicated for use in post-mastectomy breast reconstruction, has not been marketed in the United States or any foreign country. The sponsor states that the device is the same as the SurgiMend device that was cleared under K071807 for plastic and reconstructive surgery on August 6, 2007 with different device configurations, sizes, and thickness and is legally marketed in the United States as well as the EU, Canada, Colombia, Israel, Korea, Mexico, New Zealand, Panama, Peru, and Thailand. To date, the sponsor states this SurgiMend device has not been withdrawn from marketing in any country for any reason related to the safety or effectiveness of the device. While the sponsor believes that only devices cleared under K071807 were used in the MROC study based on marketing, there is no way to confirm this as fact based on the MROC dataset. Thus, FDA included a description of all SurgiMend devices that were available during the MROC study. There were two 510 k cleared devices that were available between 2012 and 2015, which would align with the enrollment of patients into the MROC Study.

The FDA supports the use of relevant and reliable RWD for regulatory decisions [16]. FDA conducted a preliminary assessment of the MROC study data regarding its potential relevance and reliability. This analysis concluded that the dataset was of sufficient quality to proceed with analyses of the prespecified outcomes and specific manufacturers' device performance. Because FDA has access to de-

identified patient-level MROC study data, but the sponsor does not, FDA conducted the analysis using the prospectively defined statistical analysis plan. Integra's PMA relies on this prospective analysis of existing observational study data to support a reasonable assurance of safety and effectiveness of the subject device (SurgiMend PRS ABDM).

In the FDA's Executive Summary [16] at the October 2021 FDA Advisory meeting, FDA discussed limitations of using MROC datasets for the SurgiMend study. They included lack of clinical studies to support a reasonable assurance of safety and effectiveness of their subject device. Furthermore, MROC Study was not designed to evaluate the safety and effectiveness of the SurgiMend PRS ABDM device. Some of the additional key drawbacks included:

1. The MROC study was an observational, non-randomized study. The study results are prone to confounding bias. The SurgiMend study data were a subset of the MROC study data.
2. The propensity score study design is applied in the SurgiMend study to mitigate the biases caused by observed confounders; however, potential biases may remain due to unmeasured confounders. For example, clinical site information is deidentified and surgeon-level data were not provided. Therefore, the SurgiMend study could not consider differences in region (i.e., United States and Canada sites), site-to-site variability, and surgeon performance.
3. There is large amount of missing data for the primary and secondary endpoints.
4. MROC followed patients for 2 years after tissue expander and SurgiMend placement. Thus, there is a lack of information on long-term adverse events (AEs) including cancer recurrence.
5. Limited information on AEs, serious AEs, and other AEs (for example causes of death).
6. Limited information on patient accounting/disposition.

FDA posed three questions for the advisory committee panel to vote on:

1. There is reasonable assurance that the SurgiMend PRS ABDM is safe for the proposed indications for use.
2. There is reasonable assurance that the SurgiMend PRS ABDM is effective for the proposed indication for use.
3. The benefits of SurgiMend PRS ABDM outweigh the risks for the proposed indications for us.

Of the 12 panel members, 7 voted yes and 5 no for the first question. For the second question, 5 voted yes, 6 no, and 1 abstention, and for the third question, 5 voted yes, 7 no. The main concerns from the panel on voting no included the following reasons:

- The dataset is small with 119 patients on SurgiMend ADM. In some categories of outcome measures, a few patients may even flip the results.
- Some of the potential confounding variables, such as institutions and physicians were not captured and adjusted for in the analysis.

- Even though the BREAST-Q is a validated endpoint, as well as the complications, it's not validated putting the two together, so the composite endpoint was not a validated endpoint.
- Propensity score can only adjust for measured variables. The sensitivity analyses using average treatment effect on the treated patients showed that the endpoint was no longer significant.
- There was differential missing data between the two groups on the BREAST-Q, which again brings in biases; therefore, the results were hard to interpret.

3.3 RWE/RWD as One of the Data Sources for Label Expansion

3.3.1 Orencia

Orencia was originally approved by the FDA in 2005 for the treatment of adult rheumatoid arthritis (RA). Orencia is also approved for the treatment of polyarticular juvenile idiopathic arthritis and adult psoriatic arthritis [33]. In December 2021, FDA approved Orencia for the prophylaxis (prevention) of aGVHD. Orencia may be used in adults and pediatric patients 2 years of age or older undergoing hematopoietic stem cell transplantation (commonly known as bone marrow transplantation or stem cell transplantation) from an unrelated donor. The approval was based on results from two key studies (GVHD-1 and GVHD-2) in patients undergoing stem cell transplant as described in Sect. 2.1 for this case study.

This is the first FDA drug approval for aGVHD prevention and incorporates RWE as supported by GVHD-2 as one component of the determination of clinical effectiveness. GVHD-2 study is based on data from CIBMTR, which maintains one of the world's largest observational databases of clinical information on hematopoietic cell transplantation (HCT). It has been collecting HCT outcomes data for 50 years, resulting in a Research Database with information from more than 585,000 patients. These data are freely available to investigators with an interest in HCT and treatments for cancer and other life-threatening diseases. As mentioned in chapter "Assessment of Fit-for-Use Real-World Data Sources and Applications" of this book [34], fit-for-purpose RWD sources can be assessed from data relevancy and reliability perspectives. CIBMTR Research Database contains relevant baseline recipient and donor information [19] along with follow-up recipient and donor information including outcomes, such as survival and aGVHD for recipients and adverse events for donors. The quality of RWD source in GVHD-2 study from a well-known source CIBMTR was recognized as fit-for-use. To reduce bias due to confounding, the analysis of OS employed weighted log-rank test with inverse propensity scores, obtained using logistic regression models as weights (inverse probability of treatment weighting [IPTW]). The marginal HR and 95% CI were estimated by a weighted Cox proportional hazards model [20].

The approval of Orencia showed that RWE/RWD was used as pivotal for a regulatory approval and addressed a critical unmet medical need. With the initial approval of Orencia in 2005 in patients with RA and other conditions, the previous approvals afforded appropriate regulatory contexts for this approval. Dr. Pazdur, director of the FDA's Oncology Center of Excellence, said at the press conference for the approval, "Acute graft versus host disease can affect different parts of the body and become a serious post-transplant complication", and "by potentially preventing the disease, more patients may successfully undergo bone marrow or stem cell transplantation with fewer complications."

3.4 RWE/RWD as Supplemental Information for the Regulatory Decision

3.4.1 Ibrance

The applicant submitted a supplemental new drug application (sNDA) to expand the proposed indication of palbociclib to include male patients with HR-positive, HER2-negative advanced, or metastatic breast cancer. Since male patients with breast cancer were ineligible in randomized controlled studies that provided the data to demonstrate the clinical benefit to support prior approvals, in their submission, the applicant provided the results of an analysis of RWD from EHRs as additional supportive data. The objective of this real-world analysis was to characterize the use of palbociclib in combination with endocrine therapy (aromatase inhibitor or fulvestrant) in male patients based on observed tumor responses in this rare subset of patients with breast cancer.

The FDA reviewers noted several limitations in the real-world analyses submitted to support the approval of palbociclib for male MBC patients. Mainly, it was pointed out that the interpretation of the findings of the retrospective cohort study in FHAD were limited since only 1% of the initial sample of 2500 patients were ultimately available for analysis [35]). Furthermore, owing to the small sample size, baseline prognostic factors could not be balanced in the palbociclib cohort and the comparator cohort with the help of matching, weighting, or other propensity score methods [35]. For example, the reviewers noted that the patients on the palbociclib cohort were younger and substantially more refractory and, thus, no direct comparison between cohorts could be made leading the palbociclib cohort being considered as a single-arm analysis. With these considerations, the conclusion from the reviewers seems to have been that while RWD provided some evidence that palbociclib in combination with endocrine therapy has antitumor activity (as measured by the real-world response rate) in men with MBC, the data did not isolate the effect of palbociclib. The FDA relied on the large, randomized studies in women for the isolation of the effect of palbociclib [35].

Ultimately, with the favorable benefit-risk profile of palbociclib in women based on several large, randomized studies combined with the supportive RWD along with

safety information from review of two phase 1 studies, the Pfizer global database and postmarketing reports, the FDA found a favorable benefit-risk profile for palbociclib in men [34]. To this end, the indication of palbociclib was expanded to include use in addition to aromatase inhibitors or fulvestrant in male patients with HR-positive, HER2-negative advanced, or metastatic breast cancer.

4 Lessons Learned and Best Practices

The six case studies demonstrated complex and multifaceted aspects with the use of RWE in regulatory sciences. Among them, a couple of these case studies were used in original submissions and subsequent approvals, further demonstrating the critical and contemporary use of RWE for regulatory decisions. As shown, these case studies were developed in rare disease area, where there are substantial unmet medical needs. We believe that the six case studies made clear of and showcased key considerations when utilizing RWE. We observed that most of the feedback from regulators and/or FDA Advisory Committee panel on the six case studies centered around fit-for-purpose RWD sources as shown in the analysis narratives in Sect. 3. For details on assessing data elements related to relevancy and reliability of a RWD source, please refer to chapter "Assessment of Fit-for-Use Real-World Data Sources and Applications" of this book [34] and Levenson et al. (2022) [36].

With the Avelumab case study, the FDA review identified limitations of the small sample size and inherent selection bias in the historical control data. It concluded the data were exploratory in nature and considered only to further characterize the natural history of disease in the target population for benefit-risk assessment. As a result, the RWD was not pivotal for the market application and approval.

With the Tafasitamab case study, the key comments were focused on data quality and data completeness as assessed in the data reliability dimension and population comparability as in the data relevancy dimension. The applicant made several modifications to their plan subsequently, such as restricting the study sites to align with L-MIND study, modifying the study SAP on only including patients with complete baseline covariates in the analysis, validating the outcome variable via an independent committee, and including a minimum follow-up period. Overall, though, the real-world study RE-MIND was judged to have many limitations and only provided contextual evidence in the original FDA approval.

For each case of Prograf and Orencia, a well-known and well-established registry data source was utilized—SRTR and CIBMTR, respectively. SRTR was considered a fit-for-purpose data source, and the success of Prograf's approval paves the way that a non-interventional study with potential challenges from confounding and other biases could still provide robust scientific findings comparable to a RCT and meet the regulatory requirements for approval. CIBMTR is another well-known and long-term running registry data source, and the FDA also recognized it as fit-for-purpose. The approval of Orencia showed that RWE/RWD was used as pivotal for a regulatory approval and addressed a critical unmet medical need.

In the SurgiMend case, however, although the FDA supported the use of MROC registry study data for the SurgiMend regulatory submission, both the FDA and the Advisory Committee panel identified several data issues as not fit-for-purpose. The majority of the Advisory Committee members voted against the approval of SurgiMend for the intended indication. Some of the issues include absence of key potential confounding variables in the dataset, large amount of missing data on key components of the primary endpoint and differential missingness among the two treatment groups, the composite endpoint with two components not being validated, sensitivity analysis leading to a different conclusion, and limited information on adverse events of interest, such as death.

With the Ibrance case, the small sample size of the retrospective cohort study in FHAD precluded any statistical comparisons to make any meaningful impact in the consideration for regulatory approval. Ultimately, with the favorable benefit-risk profile of palbociclib in women based on several large, randomized studies combined with the supportive RWD along with safety information from review of two phase 1 studies, the postmarketing reports by the sponsor, the FDA found a favorable benefit-risk profile for palbociclib in men.

5 Conclusions

The key considerations with the use of RWE require an inter-disciplinary approach and collaboration. This book provides rich discussions on key considerations from a range of topics. We argue that the regulatory guidance, key considerations by stakeholders, and current gaps in applications and best practices identified through these case studies in this chapter will provide additional insight and bring to bear the importance of the following key considerations, although may not be comprehensive, such as designing studies to minimize potential biases and confounding with the use of RWE, especially in observational research, assessing fit-for-purpose RWD sources and providing justification of the same, employing appropriate statistical methods to adjust for potential confounding factors, selecting appropriate outcome measures with relevant clinical impacts, quantifying uncertainty around the study findings, and understanding the limitations of the research using RWE/RWD. We especially recommend applicants proactively engage and communicate with regulatory agencies for discussions either at study design or pre-submission stages, since there are still great uncertainties in the use of RWE/RWD for specific applications.

Disclaimer This chapter reflects the views of the authors and should not be construed to represent FDA's views or policies.

References

1. About Merkel Cell Skin Cancer. https://www.cancer.org/cancer/merkel-cell-skin-cancer/about.html (2022). Accessed.
2. Iyer JG, Blom A, Doumani R, Lewis C, Tarabadkar ES, Anderson A, et al. Response rates and durability of chemotherapy among 62 patients with metastatic Merkel cell carcinoma. Cancer Med. 2016;5(9):2294–301. https://doi.org/10.1002/cam4.815.
3. FDA. Multi-Discipline review of avelumab (BLA 761049). https://www.accessdata.fda.gov/drugsatfda_docs/nda/2017/761049Orig1s000MultidisciplineR.pdf 2017.
4. FDA. Multi-Discipline review of tafasitamab-cxix (BLA 761163). https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/761163Orig1s000MultidisciplineR.pdf 2020.
5. Duell J, Maddocks KJ, Gonzalez-Barca E, Jurczak W, Liberati AM, De Vos S, et al. Long-term outcomes from the phase II L-MIND study of tafasitamab (MOR208) plus lenalidomide in patients with relapsed or refractory diffuse large B-cell lymphoma. Haematologica. 2021. https://doi.org/10.3324/haematol.2021.279802.
6. Jurczak W, Zinzani PL, Gaidano G, Goy A, Provencio M, Nagy Z, et al. Phase IIa study of the CD19 antibody MOR208 in patients with relapsed or refractory B-cell non-Hodgkin's lymphoma. Ann Oncol. 2018;29(5):1266–72. https://doi.org/10.1093/annonc/mdy056.
7. Valapour M, Lehr CJ, Skeans MA, Smith JM, Uccellini K, Goff R, et al. OPTN/SRTR 2018 Annual Data Report: Lung. Am J Transplant. 2020;20 Suppl s1:427–508. https://doi.org/10.1111/ajt.15677.
8. Erdman J, Wolfram J, Nimke D, Croy R, Wang X, Weaver T, et al. Lung Transplant Outcomes in Adults in the United States: Retrospective Cohort Study Using Real-world Evidence from the SRTR. Transplantation. 2022;106(6):1233–42. https://doi.org/10.1097/TP.0000000000004011.
9. Hardy JD, Webb WR, Dalton ML, Jr., Walker GR, Jr. Lung Homotransplantation in Man. JAMA. 1963;186:1065–74. https://doi.org/10.1001/jama.1963.63710120001010.
10. Veith FJ, Koerner SK. The present status of lung transplantation. Arch Surg. 1974;109(6):734–40. https://doi.org/10.1001/archsurg.1974.01360060004002.
11. Scientific Registry of Transplant Recipients. In: Administration HRaS, editor.
12. Leppke S, Leighton T, Zaun D, Chen SC, Skeans M, Israni AK, et al. Scientific Registry of Transplant Recipients: collecting, analyzing, and reporting data on transplantation in the United States. Transplant Rev (Orlando). 2013;27(2):50–6. https://doi.org/10.1016/j.trre.2013.01.002.
13. Sorkin M, Qi J, Kim HM, Hamill JB, Kozlow JH, Pusic AL, et al. Acellular Dermal Matrix in Immediate Expander/Implant Breast Reconstruction: A Multicenter Assessment of Risks and Benefits. Plast Reconstr Surg. 2017;140(6):1091-100. https://doi.org/10.1097/PRS.0000000000003842.
14. Pusic AL, Klassen AF, Scott AM, Klok JA, Cordeiro PG, Cano SJ. Development of a new patient-reported outcome measure for breast surgery: the BREAST-Q. Plast Reconstr Surg. 2009;124(2):345–53. https://doi.org/10.1097/PRS.0b013e3181aee807.
15. Pusic AL, Matros E, Fine N, Buchel E, Gordillo GM, Hamill JB, et al. Patient-Reported Outcomes 1 Year After Immediate Breast Reconstruction: Results of the Mastectomy Reconstruction Outcomes Consortium Study. J Clin Oncol. 2017;35(22):2499-506. https://doi.org/10.1200/JCO.2016.69.9561.
16. FDA. General and Plastic Surgery Devices Panel of the Medical Devices Advisory Committee Meeting: Surgimend Briefing Document. https://www.fda.gov/advisory-committees/advisory-committee-calendar/october-20-2021-general-and-plastic-surgery-devices-panel-medical-devices-advisory-committee-meeting 2021.
17. Yue LQ, Campbell G, Lu N, Xu Y, Zuckerman B. Utilizing national and international registries to enhance pre-market medical device regulatory evaluation. J Biopharm Stat. 2016;26(6):1136–45. https://doi.org/10.1080/10543406.2016.1226336.

18. Watkins B, Qayed M, McCracken C, Bratrude B, Betz K, Suessmuth Y, et al. Phase II Trial of Costimulation Blockade With Abatacept for Prevention of Acute GVHD. J Clin Oncol. 2021;39(17):1865–77. https://doi.org/10.1200/JCO.20.01086.
19. Center for International Blood and Marrow Transplant Research (CIBMTR) Database. https://www.cibmtr.org/Data/Available/Pages/index.aspx 2004.
20. Kean LS, Burns LJ, Kou TD, Kapikian R, Tang X-Y, Zhang M-J, et al. Improved Overall Survival of Patients Treated with Abatacept in Combination with a Calcineurin Inhibitor and Methotrexate Following 7/8 HLA-Matched Unrelated Allogeneic Hematopoietic Stem Cell Transplantation: Analysis of the Center for International Blood and Marrow Transplant Research Database. Blood. 2021;138:3912.
21. Society AC: Key Statistics for Breast Cancer in Men https://www.cancer.org/cancer/breast-cancer-in-men/about/key-statistics.html (2020). Accessed December 1, 2020.
22. Fentiman IS, Fourquet A, Hortobagyi GN. Male breast cancer. Lancet. 2006;367(9510):595–604. https://doi.org/10.1016/S0140-6736(06)68226-3.
23. Kraus AL, Yu-Kite M, Mardekian J, Cotter MJ, Kim S, Decembrino J, et al. Real-World Data of Palbociclib in Combination With Endocrine Therapy for the Treatment of Metastatic Breast Cancer in Men. Clin Pharmacol Ther. 2022;111(1):302–9. https://doi.org/10.1002/cpt.2454.
24. Pfizer Inc NY, NY: IBRANCE® capsules (palbociclib). Full Prescribing Information. https://labeling.pfizer.com/showlabeling.aspx?id=2191 (2019). Accessed.
25. Rugo HS, Finn RS, Dieras V, Ettl J, Lipatov O, Joy AA, et al. Palbociclib plus letrozole as first-line therapy in estrogen receptor-positive/human epidermal growth factor receptor 2-negative advanced breast cancer with extended follow-up. Breast Cancer Res Treat. 2019;174(3):719–29. https://doi.org/10.1007/s10549-018-05125-4.
26. Cowey CL, Liu FX, Kim R, Boyd M, Fulcher N, Krulewicz S, et al. Real-world clinical outcomes with first-line avelumab in locally advanced/metastatic Merkel cell carcinoma in the USA: SPEAR-Merkel. Future Oncol. 2021;17(18):2339–50. https://doi.org/10.2217/fon-2020-1250.
27. D'Angelo SP, Russell J, Lebbe C, Chmielowski B, Gambichler T, Grob JJ, et al. Efficacy and Safety of First-line Avelumab Treatment in Patients With Stage IV Metastatic Merkel Cell Carcinoma: A Preplanned Interim Analysis of a Clinical Trial. JAMA Oncol. 2018;4(9):e180077. https://doi.org/10.1001/jamaoncol.2018.0077.
28. Zinzani PL, Rodgers T, Marino D, Frezzato M, Barbui AM, Castellino C, et al. RE-MIND: Comparing Tafasitamab + Lenalidomide (L-MIND) with a Real-world Lenalidomide Monotherapy Cohort in Relapsed or Refractory Diffuse Large B-cell Lymphoma. Clin Cancer Res. 2021;27(22):6124–34. https://doi.org/10.1158/1078-0432.CCR-21-1471.
29. FDA: Approval Date(s) and History, Letters, Labels, Reviews for NDA 050708. https://www.accessdata.fda.gov/scripts/cder/daf/index.Cfm?event=overview.process&ApplNo=050708 Accessed Aug 20 2022.
30. Hanto DW. Reliability of voluntary and compulsory databases and registries in the United States. Transplantation. 2003;75(12):2162–4. https://doi.org/10.1097/01.TP.0000080273.83998.C4.
31. FDA. FDA approves new use of transplant drug based on real-world evidence. https://www.fda.gov/drugs/news-events-human-drugs/fda-approves-new-use-transplant-drug-based-real-world-evidence 2021.
32. FDA. General and Plastic Surgery Devices Advisory Committee meeting. https://www.fda.gov/advisory-committees/advisory-committee-calendar/march-25-26-2019-general-and-plastic-surgery-devices-panel-medical-devices-advisory-committee 2019.
33. FDA. Orencia FDA drug approval package https://www.accessdata.fda.gov/drugsatfda_docs/nda/2005/125118_s0000_OrenciaTOC.cfm#:~:text=Approval%20Date:%2012/23/2005 2005.
34. He W, Zhang Z, Dharmarajan S. Assessment of fit-for-use real-world data sources and applications. In: He W, Fang Y, Wang H, editors. Practical Approaches and Considerations for Translating Real-World Data into Robust Real-World Evidence for Regulatory Decisions. Springer Nature; 2023.

35. Wedam S, Fashoyin-Aje L, Bloomquist E, Tang S, Sridhara R, Goldberg KB, et al. FDA Approval Summary: Palbociclib for Male Patients with Metastatic Breast Cancer. Clin Cancer Res. 2020;26(6):1208–12. https://doi.org/10.1158/1078-0432.CCR-19-2580.
36. Levenson M, He W, Dharmarajan S, Izem R, Meng Z, Pang H, et al. Statistical consideration for fit-for-use real-world data to support regulatory decision making in drug development. Statistics in Biopharmaceutical Research 2022. https://doi.org/10.1080/19466315.2022.2120533.

The Use of Real-World Data to Support the Assessment of the Benefit and Risk of a Medicine to Treat Spinal Muscular Atrophy

Tammy McIver, Muna El-Khairi, Wai Yin Yeung, and Herbert Pang

1 Introduction

1.1 Spinal Muscular Atrophy

Spinal muscular atrophy (SMA) is a rare, debilitating genetic neuromuscular disease. It affects approximately 1 in 10,000 individuals and when untreated is the leading genetic cause of infant mortality [1]. SMA is characterized by progressive loss of motor neurons (nerve cells that control muscle movement). The disease is caused by mutations or deletions in the survival of motor neuron 1 (*SMN1*) gene, which leads to a deficiency in SMN protein [2]. SMN protein is found throughout the body and is essential for the function of nerves that control muscles and movement. Without SMN protein, motor neurons cannot function properly, which in turn leads to muscle wasting over time [3]. Depending on the type of SMA, an individual's physical strength and their ability to walk, eat, or breathe can be significantly diminished or lost [4].

Patients with SMA are typically classified into types 1–4 based on the age of symptom onset and highest motor milestone achieved [5–7], with types 1, 2, and 3 SMA representing approximately 99% of the SMA population [8]. Table 1 shows a summary of the primary SMA types.

Since 2016, the US Food and Drug Administration (FDA) has approved three medications to treat SMA: nusinersen (SPINRAZA®), onasemnogene abeparvovec-xioi (ZOLGENSMA®), and risdiplam (EVRYSDI®). The goal of these disease-

T. McIver · M. El-Khairi · W. Y. Yeung
PD Data Sciences – Biostatistics, Roche Products Limited, Welwyn Garden City, UK

H. Pang (✉)
PD Data Sciences – Biostatistics, Genentech, South San Francisco, CA, USA
e-mail: pathwayrf@gmail.com

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6_20

Table 1 Summary of primary SMA types

Type	Age at onset	Impact
1	Before 6 months	Never sit independently Life expectancy less than 2 years
2	6–18 months	Able to sit and may stand with assistance Never walk
3	18 months onward	Able to stand and walk Often lose the ability to walk in early life

SMA spinal muscular atrophy

modifying treatments is to increase the availability of SMN protein, leading to clinically meaningful improvements in muscle function.

1.2 Risdiplam

At the time of initiating the risdiplam clinical development program in 2016, there was no approved treatment for SMA. Risdiplam was developed by Roche/Genentech (the sponsor) in partnership with PTC Therapeutics and the SMA Foundation to help address the unmet needs for children and adults with SMA. Risdiplam is a small molecule administered daily at home in liquid form by mouth or feeding tube. It is a selective *SMN2* gene splicing modifier that increases the production of full-length SMN protein in the central nervous system and peripheral tissues [9]. It was hypothesized that increasing the amount of SMN protein would reduce motor neuron degeneration thereby limiting muscle atrophy.

A series of clinical studies on risdiplam were designed to represent a broad spectrum of people with SMA, from birth to 60 years of age.

- FIREFISH (NCT02913482): an open-label, single-arm, two-part study in infants aged 1–7 months with type 1 SMA ($N = 62$).
- SUNFISH (NCT02908685): a randomized, placebo-controlled, two-part study in children and young adults aged 2–25 years with type 2 or 3 SMA ($N = 231$).
- JEWELFISH (NCT03032172): an open-label, single-arm study in children and adults aged 6 months to 60 years who have taken part in clinical trials for SMA or received other investigational or approved SMA therapies ($N = 174$).
- RAINBOWFISH (NCT03779334): an open-label, single-arm study in infants genetically diagnosed with SMA and not yet presenting symptoms ($N = 26$).

Table 2 shows a summary of key milestones in the development of risdiplam.

As of September 2022, risdiplam has been approved in more than 90 countries and the dossier is under review in 18 countries. More than 7000 people have been treated with risdiplam across clinical trials, through the Compassionate Use Program/Pre-Approval Access and in the commercial setting.

Table 2 Summary of key milestones in the development of risdiplam

Year	Milestone
2016	First patient dosed with risdiplam in a clinical study (SUNFISH)
2017	Risdiplam was granted Orphan Drug designation by the FDA
2018	Risdiplam was granted PRIME designation by the EMA
2020	The FDA approved risdiplam for the treatment of SMA in adults and children aged 2 months and older
2021	The EMA approved risdiplam for the treatment of SMA in patients from 2 months old with type 1, 2, or 3 SMA, or those who have up to four copies of a gene known as *SMN2*
2022	The FDA approved a label extension for the use of risdiplam in infants with SMA under 2 months of age

EMA European Medicines Agency, *FDA* US Food and Drug Administration, *SMA* spinal muscular atrophy, *SMN2*, survival of motor neuron 2, *PRIME* Priority Medicines

The focus of this case study will be on the two pivotal studies in the risdiplam clinical development program, FIREFISH and SUNFISH. Both these studies had an operationally seamless design, with an exploratory dose-finding part (Part 1) and a confirmatory part (Part 2). Real-world data (RWD) were critical to support the clinical development planning, data interpretation, and registration of risdiplam. Here, we describe how RWD from publications were used to define performance criteria for key clinical endpoints in FIREFISH and benchmark the results for success in patients with type 1 SMA. In addition, RWD from individual patients were used to perform a robust statistical comparison and contextualize the results from SUNFISH in patients with type 2 and 3 SMA. In this chapter, we will discuss why we used RWD, how we used RWD and the impact of using RWD, including a summary of challenges and lessons learned.

2 FIREFISH Study: External Control Data from Publications

2.1 Design and Methods

2.1.1 Study Design

FIREFISH was an operationally seamless, two-part, open-label, multicenter Phase 2/3 study to investigate the safety, tolerability, pharmacokinetics, pharmacodynamics, and efficacy of risdiplam in infants with type 1 SMA aged 1–7 months at enrolment. Figure 1 shows a summary of the FIREFISH study design.

FIREFISH Part 1 was an exploratory, dose-finding study conducted in 21 infants with type 1 SMA, which determined the dose for use in Part 2 [10]. FIREFISH Part 2 was a confirmatory study conducted in 41 infants with type 1 SMA [11]. Part

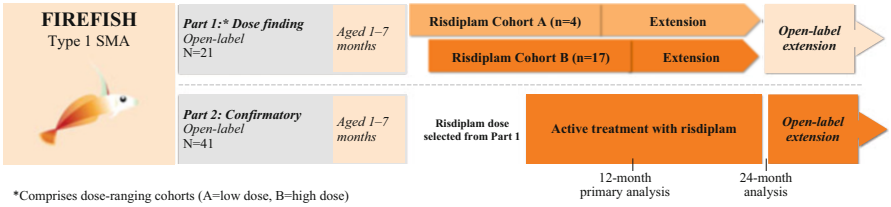

Fig. 1 Study Design of Part 1 and Part 2 of FIREFISH. *SMA* spinal muscular atrophy

1 and Part 2 enrolled different infants. The primary endpoint for the confirmatory Part 2 of the study was the proportion of infants sitting without support for at least 5 seconds (s) after 12 months of treatment, as assessed by Item 22 of the Bayley Scales of Infant and Toddler Development, third edition (BSID-III) Gross Motor Subscale [12]. Sitting without support was selected as the primary endpoint because achieving this milestone illustrates a divergence from the natural course of type 1 SMA as these infants would never achieve this motor milestone without treatment. Besides, sitting independently is clinically meaningful for infants, as it allows them to use their upper limbs to reach for objects, grasp objects, and feed themselves. In addition, event-free survival, defined as being alive without permanent ventilation, was a key secondary endpoint. The overall study design and choice of endpoints incorporated health authority advice from both the European Medicines Agency (EMA) Committee for Medicinal Products for Human Use (CHMP) and the FDA. The efficacy outcome measures analyzed in FIREFISH Part 1 were consistent with those in Part 2, but no formal hypothesis testing was planned for Part 1.

RWD as a Component of Study Design in FIREFISH

The natural history of type 1 SMA is well defined and has been described in numerous studies [13, 14]. Untreated infants with type 1 SMA are never able to sit independently [15]. In addition, natural history shows that 50% of infants will have died or required permanent non-invasive ventilation support by 10.5 months of age and 92% of infants by 20 months of age [13].

Given the severity, rapid decline, and high mortality and morbidity in infants with type 1 SMA, a placebo comparator group was not included in the FIREFISH study. In the absence of a control arm, RWD played an important role in the design of the study and the analysis and interpretation of the study results. Natural history data were used to define thresholds of achievement, or "performance criteria," against which to assess the efficacy of risdiplam treatment in Part 2 of the study. This approach can be acceptable in the development of new treatments for serious and rare diseases such as type 1 SMA if the natural course of the disease is well understood, the external comparator group is similar to the treatment group (e.g., with regard to patient characteristics and endpoints), and a large treatment effect is expected with the study drug. A treatment is considered to be effective if the

threshold for success is crossed, and when the observed treatment effect is large, it is reasonable to exclude chance or bias as a possible explanation [16].

The FIREFISH study included objective endpoints that facilitated the comparison with RWD. The primary endpoint (sitting without support for at least 5 s) was assessed using strict criteria to minimize bias following the BSID-III manual, objectively measured by trained clinical evaluators, video recorded and scored by two independent reviewers. As sitting without support is never achieved in untreated infants with type 1 SMA [13], a large divergence from natural history would be expected when treated with an efficacious drug. Secondary endpoints included motor function, achievement of other developmental motor milestones, survival, and event-free survival. Performance criteria were defined for the primary and key secondary endpoints in Part 2 of the FIREFISH study.

Identification and Selection of Historical Data Sources for Defining Performance Criteria

First, a literature search was performed to identify publications in infants with type 1 SMA that included an endpoint reported in FIREFISH. A PubMed search was undertaken using the following search terms: (spinal muscular atrophy [Title/Abstract]) AND (observational OR cohort OR natural history OR registr* OR association OR describe OR description OR match* or control*) AND ("2000/01/01"[Date – Publication]: "3000/01/01"[Date – Publication]), which gave 938 hits (February 1, 2018). Titles and abstracts were reviewed to identify articles that appeared to report on the outcome measures included in the FIREFISH study in an observational setting, from both retrospective and prospective data collection, in infants with type 1 SMA. This left a total of 35 articles that could be used for comparison with the FIREFISH study. From these 35 articles, publications were excluded for the following reasons related to usability of the data, patient characteristics, and standard of care:

- No data relating to any of the endpoints in the FIREFISH study
- No extractable data to derive a performance criterion for any of the endpoints in the FIREFISH study, e.g., no longitudinal data were provided to calculate change from baseline values
- No infants with a genetic confirmation of SMA
- Standard of care not reflective of guidelines described in the consensus statement for standards of care in SMA [17], i.e., no consistent use of non-invasive ventilation or gastrostomy tube

Of the 35 observational studies identified, 18 were excluded for meeting one or more of these criteria. In addition to the 17 remaining observational studies, the untreated, matched cohort selected retrospectively as a comparator group in a Phase 1/2 study of valproic acid and carnitine in infants with type 1 SMA [18] was also included as a potential data source. Endpoints defined in the FIREFISH study were sometimes described in more than one of these studies. The identified studies were

ranked based on the level of similarity of the patient population to the expected population in the FIREFISH study. When more than one source was available for an endpoint, the published study cohort with baseline characteristics most similar to those targeted by the FIREFISH study inclusion and exclusion criteria (i.e., the published study with the highest ranking) was selected to set the performance criterion. The following characteristics were considered when determining the similarity of the historical cohorts to the FIREFISH study population:

- *SMN2* copy number
- Age at onset of symptoms
- Age at enrolment (start of follow-up in study or presentation at treating center)
- Type 1 SMA classification
- Standard of care
- Time period
- Region
- Type of treating center

Multicenter, prospective studies were given a higher ranking, while studies were given a lower ranking if the data collection occurred prior to the publication of the consensus statement for standards of care in SMA [17]; if the data were collected over a long period of time (e.g., 15 years) over which the standard of care would be expected to change; and if there was no information provided for important infant characteristics such as *SMN2* copy number and age at onset of symptoms. Based on these criteria, the population in the NeuroNEXT SMA infant biomarker study [14, 19] was judged to be most similar to the expected population in the FIREFISH study. Whenever possible, the performance criterion derived from this study was selected as the benchmark to be used for hypothesis testing. When data for an endpoint were not available from the NeuroNEXT SMA infant biomarker study (e.g., for the Hammersmith Infant Neurological Examination, Module 2, which is a secondary endpoint not described in this book chapter), the benchmark was derived from the study conducted by De Sanctis et al. [20]. The NeuroNEXT SMA infant biomarker study included 16 patients with two copies of the *SMN2* gene, and the study conducted by De Sanctis et al. [20] included 24 infants classified as type 1B. The demographic and baseline characteristics of these two cohorts and of the infants enrolled in FIREFISH Part 2 are presented in Table 3.

2.1.2 Statistical Methodology: Performance Criteria Approach

The performance criterion for the primary endpoint in Part 2 of the FIREFISH study (the proportion of infants sitting without support for at least 5 s after 12 months of treatment) was based on the natural history of the disease in which untreated patients with type 1 SMA never achieve sitting without support [13]. A threshold of 5% was chosen to provide sufficient confidence that any effect seen in the FIREFISH study would not otherwise have occurred in the natural history of infants enrolled in the study. An exact binomial test was performed to test the hypothesis that the

Table 3 Demographic and baseline characteristics of patients in the NeuroNEXT SMA infant biomarker study, the study conducted by De Sanctis et al. and FIREFISH Part 2

	NeuroNEXT SMA Infant Biomarker Study ($N = 16$)	De Sanctis et al. ($N = 24$)	FIREFISH Part 2 ($N = 41$)
Age at enrolment/first visit (months)	≤6	Range: 2–7	Median: 5.3 Range: 2.2–6.9
Age at onset of symptoms (months)	<1 month: 6 (38%) 1–2 months: 5 (31%) 2–3 months: 3 (19%) 4–5 months: 1 (6%) Unknown: 1 (6%)	After first week but before 5–6 months	Median: 1.5 Range: 1.0–3.0
SMN2 copy number 2 Unknown	 16 (100%) –	 – 24 (100%)	 41 (100%) –
Country	United States	Italy, United States	Brazil, China, Croatia, France, Italy, Japan, Poland, Russia, Turkey, United States

Source: Kolb et al. [19], De Sanctis et al. [20], Darras et al. [11]

proportion of infants who sit without support on treatment (p_1) was:

$$H_0 : p_1 \leq 5\% \text{ (null) versus } H_a : p_1 > 5\% \text{ (alternative)}$$

If the one-sided *p*-value was ≤5%, then the null hypothesis would be rejected. If the lower limit of the two-sided 90% confidence interval (CI) (Clopper–Pearson) was above the 5% threshold, then the primary objective of the study would be considered achieved. With a sample size of 41 infants, a minimum of 6 infants sitting without support would be needed for a statistically significant result. Infants were classified as non-responders for the primary endpoint if they were not able to sit without support at Month 12, did not maintain sitting achieved at an earlier time-point, were withdrawn or died prior to Month 12, or had a missing assessment at Month 12.

The performance criterion for the key secondary endpoint of event-free survival was based on the NeuroNEXT SMA infant biomarker study [14]. The point estimate and 90% CI were obtained after selecting infants with two *SMN2* gene copies, similar to the population in the FIREFISH study. The performance criterion was based on the upper limit of the 90% CI, derived using the complementary log-log transformation for the proportion of patients who were alive without permanent ventilation at 18 months of age. The benchmark was set at 18 months of age to reflect the expected average age of infants in Part 2 of the FIREFISH study after 12 months of treatment. The estimated proportion (90% CI) of patients alive without permanent ventilation at 18 months of age based on the available data was 20% (5–42), giving a performance criterion of 42%. Potential performance criteria were also calculated

from other available data sources and documented in the appendix of the Statistical Analysis Plan, along with key selection criteria and rankings for transparency.

When a pre-defined benchmark could be determined for a secondary endpoint, hypothesis testing was performed. For the secondary endpoint of event-free survival, a z-test was performed to test the hypothesis that the proportion of infants alive without permanent ventilation at Month 12 on treatment (p_2) was:

$$H_0 : p_2 \leq 42\% \text{ (null) versus } H_a : p_2 > 42\% \text{ (alternative)}$$

To control for multiplicity across the different endpoints, a hierarchical testing approach was implemented.

2.2 Results

The baseline characteristics of the patients enrolled in Part 1 and Part 2 of the FIREFISH study were representative of a population with well-established, symptomatic type 1 SMA. Of the 41 infants enrolled in Part 2, 22 (54%) were female. At enrolment, the median age of patients was 5.3 months (range: 2.2–6.9 months). The median age at onset of symptoms was 1.5 months (range: 1.0–3.0 months). No infants were able to sit without support at baseline. The results for sitting without support and event-free survival in FIREFISH Part 1 and Part 2 at Month 12 compared with the pre-defined performance criteria are shown in Table 4. The results are presented separately for Part 1 and Part 2 as they included different infants. In addition, the performance criteria were pre-defined for the confirmatory Part 2 only.

The primary efficacy endpoint of the study was successfully met. Twelve of 41 infants (29%; 90% CI 18–43) in Part 2 were able to sit without support for at least 5 s after 12 months of treatment. This proportion was significantly higher than the pre-defined performance criterion of 5% based on well-established natural history data ($p < 0.0001$). At 12 months of treatment, seven of 21 infants in Part 1 (33%; 90% CI 17–54) were able to sit without support for at least 5 s. All infants who were ongoing in the study had an assessment at Month 12. These results were clinically meaningful, as untreated patients with type 1 SMA are unable to sit without support at any age.

In Part 2, the proportion of infants alive without permanent ventilation at Month 12 was 85% (90% CI 73–92). Three infants died within the first 3 months following study enrolment, and three infants met the endpoint of permanent ventilation. One infant who attended the Month 12 visit a few days early and therefore had not yet reached 12 months from enrolment as of the data-cutoff date was censored in the analysis. The proportion of infants alive without permanent ventilation (85%) was significantly higher than the pre-defined performance criterion of 42% ($p < 0.0001$). Figure 2 shows a summary of the results for event-free survival.

Table 4 Results from FIREFISH part 1 and part 2 at month 12

	Part 1	Part 2		
Endpoint	Risdiplam ($N = 21$)	Risdiplam ($N = 41$)	Performance criterion	p-value[a]
Sitting without support for ≥5 s – BSID-III (90% CI)	33% (17–54)	29% (18–43)	5%	<0.0001
Alive without permanent ventilation[b] (90% CI)	90% (73–97)	85% (73–92)	42%	<0.0001

Source: Baranello et al. [10], Darras et al. [11]
BSID-III Bayley Scales of Infant and Toddler Development, third edition, *CI* confidence interval
[a]p-value for sitting without support is based on an exact binomial test; p-value for event-free survival is based on a z-test
[b]Proportions are estimated using Kaplan–Meier methodology
Permanent ventilation is defined as tracheostomy, or ≥16 h of non-invasive ventilation per day for >21 consecutive days or intubation for >21 consecutive days in the absence of, or following the resolution of, an acute reversible event

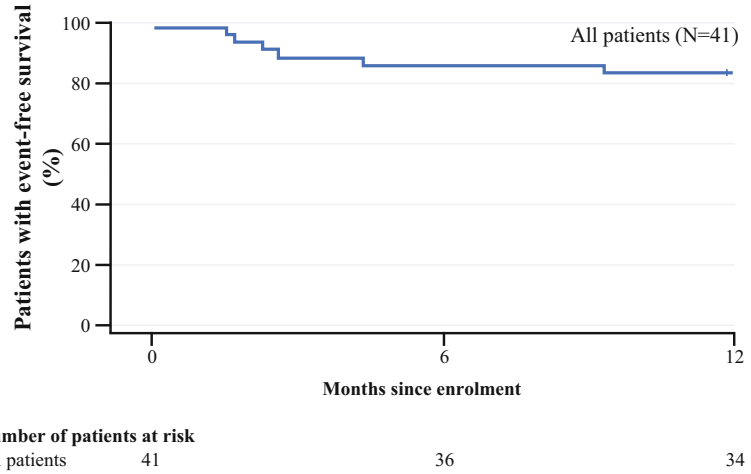

Fig. 2 FIREFISH Part 2: Event-free survival at Month 12 (Intent-to-Treat Population). (Source: Darras et al. [11])

In Part 1, the proportion of infants alive without permanent ventilation at Month 12 was 90% (90% CI 73–97). Two infants died prior to Month 12, and no infants met the definition of permanent ventilation. The median time to death or permanent ventilation was not estimable as few patients had an event. Clinically meaningful and statistically significant improvements were also observed for other key secondary endpoints in FIREFISH Part 2 [11]. The results of these analyses were used to confirm the benefits of risdiplam in type 1 SMA and thus to support the approval and registration of risdiplam in different countries around the world.

The results for Part 1 were used for the initial regulatory filing and are included in the United States prescribing information (EVRYSDI® prescribing information) [21]. This was because we filed early based on the Part 1 results, before the results from Part 2 were available due to the clear divergence shown from natural history and the high unmet medical need.

3 SUNFISH Study: External Control Data from Individual Patient Data

3.1 Design and Methods

3.1.1 Study Design

SUNFISH was an operationally seamless, two-part, multicenter, randomized, double-blind, placebo controlled, Phase 2/3 study, designed to assess the safety, tolerability, pharmacokinetics, pharmacodynamics, and efficacy of risdiplam in a broad patient population including children, teenagers, and adults aged 2–25 years with type 2 and 3 SMA. Figure 3 shows a summary of the SUNFISH study design.

SUNFISH Part 1 was an exploratory, dose-finding study conducted in 51 patients with type 2 and ambulant or non-ambulant type 3 SMA, which determined the dose for use in Part 2. SUNFISH Part 2 was a confirmatory study conducted in 180 patients with type 2 or non-ambulant type 3 SMA [22]. Part 1 and Part 2 had different patients. The primary efficacy endpoint in Part 2 was the change in motor function assessed using the 32-item Motor Function Measure (MFM32) from baseline to Month 12. The MFM32 is a clinician-reported outcome measure that evaluates different levels of motor function in individuals with SMA, from distal fine motor movements of the hands such as using a touch-screen to more complex gross motor function activities such as standing and transfers [23]. The 32 items of this measure were scored using a 4-point Likert scale: 0: cannot initiate the task, 1: can perform the task partially, 2: can perform the task incompletely or completely but imperfectly; 3: can perform the task fully and "normally." The raw score of the

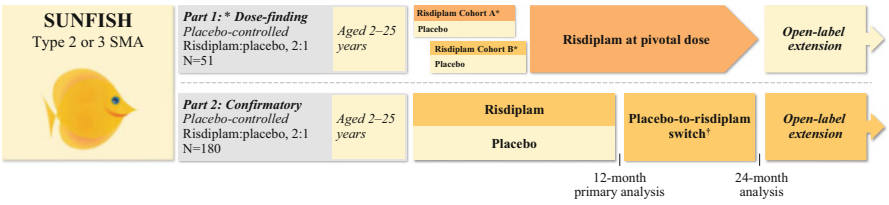

Fig. 3 Study design of Part 1 and Part 2 of SUNFISH [*Comprises two age groups (Cohort A: 2–11 years, two dose levels; Cohort B: 12–25 years, three dose levels). †Placebo-treated patients were switched to risdiplam in a blinded manner]

32 items (range: 0–96) was converted to a 0–100 scale, where lower scores indicate poorer functional ability [24].

The overall study design and choice of endpoints incorporated health authority advice from both the EMA CHMP and the FDA.

RWD as a Component of Study Design in SUNFISH

RWD were important for determining the anticipated treatment effect for the primary endpoint (MFM32) in the SUNFISH study as natural history data demonstrated that patients with type 2 and 3 SMA had a decline in motor function over time. Patients with type 2 SMA are able to sit independently and occasionally stand or take a few steps, but are unable to walk independently [15]. Patients with type 3 SMA are able to sit, stand, and walk independently [15], though nearly a third of patients with type 3 SMA lose their ability to walk between the ages of 3 and 28 years [24]. Natural history studies show that patients with type 2 and 3 SMA have a decline in motor function over time, as reported in a number of publications with different validated motor function measures. For example, natural history data demonstrated that the overall slope of decline over time, using the MFM32 total score, is in the range of −0.9 points/year for patients with type 2 SMA and −0.6 points/year for patients with type 3 SMA [24]. In order to gain additional information on the Motor Function Measure (MFM) endpoint in a broad population, the sponsor co-funded a Natural History Study (NatHis-SMA; NCT02391831) in patients with type 2 and 3 SMA with the Institute of Myology, which was designed in collaboration with Patient Advisory Groups (SMA Europe, Cure SMA, SMA Foundation). This study was also critical to ensure access to good-quality data to perform a robust statistical analysis of the MFM endpoint in SUNFISH compared with RWD. RWD were used to generate an external comparator group of patients with type 2 and 3 SMA to give context to the SUNFISH Part 1 results before the placebo-controlled results from SUNFISH Part 2 were available.

Selection of External Comparator Sources

The external comparator group used to give context to the SUNFISH Part 1 results comprised of untreated patients with SMA from the NatHis-SMA study and the placebo arm of a Phase 2 trial of olesoxime for the treatment of SMA (NCT01302600).

- The *NatHis-SMA Study* was a prospective, multicenter, longitudinal natural history study of patients with type 2 and 3 SMA. The primary objective of this study was to characterize the disease course in patients with type 2 and 3 SMA using standardized evaluations including the MFM. The study included 81 patients aged 2–29 years and was conducted in Europe between 2015 and 2018.

The maximum duration of study participation for each patient was 24 months [25].

- The *Olesoxime Study* was a Phase 2, parallel-group, placebo-controlled, randomized, double-blind, multicenter study, designed to assess the efficacy and safety of olesoxime over a 2-year period in patients with type 2 or non-ambulant type 3 SMA. The study included 165 patients aged 2–25 years, of whom 57 were randomized to placebo and was conducted in Europe between 2010 and 2013 [26]. The development of olesoxime has since been discontinued.

The NatHis-SMA study and olesoxime study were considered as appropriate sources for generating an external comparator group because of the following similarities to SUNFISH:

- Similar patient population with type 2 and 3 SMA
- All studies included the MFM scale as an outcome measure
- Studies were conducted in Europe with an overlap of some study centers
- SUNFISH and the olesoxime study were both conducted in the same controlled clinical setting. The olesoxime study was placebo controlled which provided a robust control arm
- Investigators from SUNFISH and the NatHis-SMA study were trained in the same way with regard to the MFM scale application, hence the assessment was considered similar
- The first year of follow-up in the NatHis-SMA study occurred just before study enrolment started for SUNFISH, hence patients had similar standards of care and calendar time bias (a bias associated with patients treated in the past progressing differently than those treated today due to changes in standard of care over time) was likely small.

Endpoint Used for Comparison with External Control Analysis

Although patient motor function was measured using the MFM in all three studies, the scale was not administered in the same way. In SUNFISH, all patients completed all 32 items (MFM32), whereas in the NatHis-SMA and olesoxime studies, patients aged less than 6 years completed 20 items (MFM20) while patients aged 6 years or older completed all 32 items. MFM total score was chosen to compare motor function between SUNFISH and the external comparator group. MFM total score was derived from the MFM20 total score for all patients aged less than 6 years and from the MFM32 total score for all patients aged 6 years or older. Both scales were transformed to 0–100%. Missing items on the MFM scale were recorded as 0 (i.e., cannot initiate) prior to calculation of total score. Only patients with an MFM assessment at baseline and at least one post-baseline assessment at Month 12 or Month 24 were included in the analysis.

3.1.2 Statistical Methodology

Patients in the external comparator group were weighted using Inverse Probability of Treatment Weighting based upon pre-selected prognostic factors at baseline: age at enrolment; SMA type; *SMN2* copy number; ambulatory status; presence of scoliosis; MFM total score at baseline; and MFM scale used. This allowed a comparison of the treated and untreated groups with similar prognostic factors. A propensity score was estimated for each patient using logistic regression incorporating the pre-selected prognostic factors of treatment assignment (risdiplam vs no risdiplam) as independent variables. Patients with missing prognostic factors were excluded. Trimming, defined as removing extreme values and outliers [27], was applied to include only patients with an overlapping distribution of propensity scores. Inverse Probability of Treatment Weighting (IPTW) was applied to the propensity scores to derive weights only for the external comparator group based on the average effect for treated patients approach. The IPTW approach was chosen because it was considered to be an efficient approach where patients with unknown or missing prognostic factors were not included in the weighting procedure. A weight of 1 was given to each of the patients in the risdiplam-treated group and a weight of $p_j/(1 - p_j)$ was given to the jth patient in the untreated external comparator group, where p_j was the propensity score of the jth patient. In other words, the IPTW was applied to the propensity scores to derive weights only for the external control group based on the average effect for the treated patients (ATT) approach. To control for too much influence of patients with very low propensity scores, weights were truncated at the 99th percentile. The truncation was applied after trimming. The variance balance between the treated and untreated groups was assessed pre- and post-weighting. The standardized mean difference (SMD) was computed for each of the covariates to assess if adequate balance had been achieved between the treated and the untreated groups. Adequate balance was assumed if all SMDs were less than 0.25 [28].

The statistical analysis was performed after weighting was applied. Change from baseline in MFM total score was analyzed using a mixed model for repeated measures (MMRM) with treatment; time; time by treatment; MFM total score at baseline by time; and the prognostic factors (age at enrolment; SMA type; *SMN2* copy number; ambulatory status; presence of scoliosis; MFM total score at baseline; and MFM scale used) as covariates. Estimated treatment differences in least squares mean change from baseline between patients treated with risdiplam in SUNFISH Part 1, and the external comparator were calculated with corresponding 95% CIs and *p*-values. The proportions of patients demonstrating improvement (≥3-points change from baseline in MFM total score) were analyzed via logistic regression. In the responder analyses, only patients with an MFM total score at baseline and the post-baseline time point (Month 12 or Month 24) of interest were included in the analysis. Supplemental analyses were performed on each of the external comparator data sources separately.

3.2 Results

After excluding patients with missing information on selected prognostic factors and trimming, 48 patients from the risdiplam arm of SUNFISH Part 1 and 109 patients from the external comparator group who had a valid MFM total score at baseline and Month 12 or Month 24 were included in this analysis. In particular, with the trimming, two treated patients from SUNFISH Part 1 were excluded due to extreme weights (i.e., the prognostic profile of these two patients was not similar to those in the external control group) and no patients from the external control group were excluded. In addition, no patients from either group were excluded due to truncation. After weighting was applied, weights were summed to generate an external comparator group of 49.3. All patients from SUNFISH Part 1 were given a weight of 1, to give a sum of weights of 48.0. The balance between the treated and untreated groups in terms of their baseline prognostic factors profile (covariate balance) was assessed using the SMDs. The SMDs are presented in Fig. 4 for each of the covariates prior to and after weighting.

Prior to weighting (for "All" and "Region"), except for the MFM total score, all covariates had already achieved balance between the treated patients in SUNFISH Part 1 and the untreated patients in the external comparator group, with their corresponding SMD values lying within the range of −0.25 to +0.25. After weighting (ATT weighted region), variance balance of each covariate was achieved

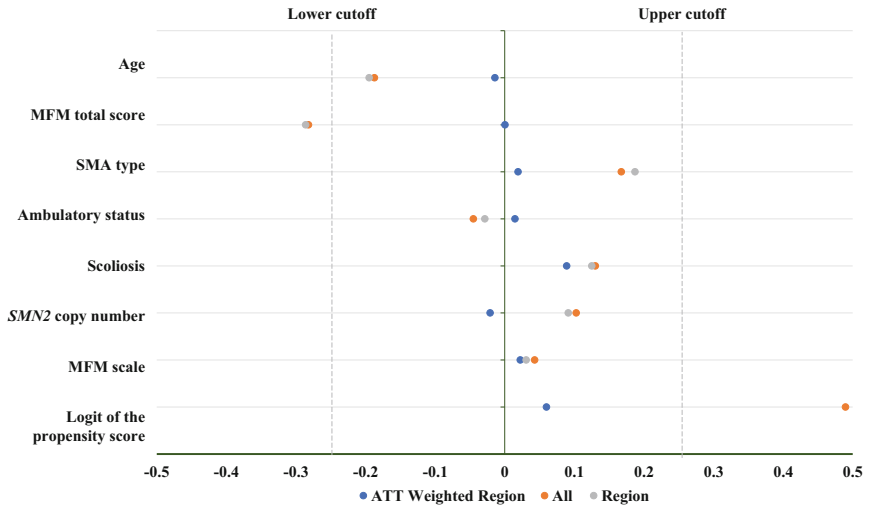

Fig. 4 Results of the covariate balance assessment (standardized mean difference values) of SUNFISH Part 1 compared with an external comparator group. "All" means based on the population from the external comparator group, "Region" means based on the patients included in the analysis and "ATT weighted region" means based on the patients after weighting. ATT average effect for treated patients, *MFM* motor function measure, *SMA* spinal muscular atrophy, *SMN2* survival of motor neuron 2

Table 5 Summary of baseline characteristics in SUNFISH part 1 compared with an external comparator group before and after weighting

	Before weighting		After weighting	
	Risdiplam ($N = 48$)	External comparator ($N = 109$)	Risdiplam (w$N = 48.0$)	External comparator (w$N = 49.3$)
Age at enrolment (years)				
mean (SD)	9.3 (6.1)	10.5 (6.8)	9.3 (6.1)	9.4 (6.3)
median (range)	7 (2–24)	8 (2–28)	7.0 (2–24)	7.0 (2–28)
Age group (years), n (%)				
2–5	17 (35.4)	37 (33.9)	17.0 (35.4)	16.9 (34.3)
6–11	13 (27.1)	27 (24.8)	13.0 (27.1)	16.3 (33.2)
12–18	14 (29.2)	28 (25.7)	14.0 (29.2)	11.5 (23.3)
>18	4 (8.3)	17 (15.6)	4.0 (8.3)	4.5 (9.1)
SMA type, n (%)				
Type 2	35 (72.9)	70 (64.2)	35.0 (72.9)	35.5 (72.0)
Type 3 (ambulant)	7 (14.6)	17 (15.6)	7.0 (14.6)	6.9 (14.1)
Type 3 (non-amb.)	6 (12.5)	22 (20.2)	6.0 (12.5)	6.9 (13.9)
SMN2 copy number, n (%)				
3	44 (91.7)	97 (89.0)	44.0 (91.7)	45.5 (92.3)
4	4 (8.3)	12 (11.0)	4.0 (8.3)	3.8 (7.7)
MFM total score, mean (SD)				
MFM20	($n = 17$) 53.9 (13.6)	($n = 37$) 57.1 (16.0)	($n = 17.0$) 53.9 (13.6)	($n = 16.9$) 55.3 (17.2)
MFM32	($n = 31$) 44.4 (15.4)	($n = 72$) 50.2 (18.0)	($n = 31.0$) 44.4 (15.4)	($n = 32.4$) 43.8 (17.8)
Scoliosis, n (%)	27 (56.3)	68 (62.4)	27.0 (56.3)	29.9 (60.6)

MFM Motor Function Measure, *MFM20* 20-item MFM, *MFM32* 32-item MFM, *SD* standard deviation, *SMA* spinal muscular atrophy, *SMN2* survival of motor neuron 2, *wN* number of patients after weighting, *non-amb* non-ambulant

with all SMDs close to 0 and lying within the −0.25 to +0.25 boundaries. Summary results for the baseline characteristics before and after weighting are shown in Table 5. After weighting, the baseline characteristics became more similar and more comparable between SUNFISH Part 1 and the external comparator group. For example, the mean age in the external comparator group was 10.5 years before weighting and 9.4 years after weighting, compared with 9.3 years in the SUNFISH Part 1 treated group. The proportion of patients in each age group also became more balanced between the two groups after weighting.

For the MMRM analysis, patients with a baseline and at least one post-baseline result at Month 12 or at Month 24 were included. In SUNFISH Part 1, all 48 patients had a result at baseline, Month 12 and Month 24. For the external comparator group, 109 patients had a result at baseline and at Month 12, and 79 patients had a result at Month 24. Under the MMRM analysis, for those with a result at Month 12 but not at

Table 6 Mean (LS Mean) change in motor function (as measured using the MFM) at month 12 and month 24 in patients with type 2 or 3 SMA in SUNFISH part 1 compared with an external comparator group

MFM-derived total score	Risdiplam (weighted $N = 48.0$)	External comparator (weighted $N = 49.3$)
Month 12		
Baseline, mean total score (SD)	47.8 (15.35)	47.8 (18.26)
Change from baseline at Month 12, mean (95% CI)	2.12 (0.61–3.62)	−0.56 (−2.08–0.95)
Difference from external comparator, LS mean (95% CI)	2.68 (1.44–3.93)	
p-value	$p < 0.0001$	
Month 24		
Change from baseline at Month 24, mean (95% CI)	1.99 (0.33–3.66)	−2.00 (−3.73 to 0.27)
Difference from external comparator, LS mean (95% CI)	3.99 (2.34–5.65)	
p-value	$p < 0.0001$	

Source: Mercuri et al. [22]

Data analyzed using an MMRM. The statistical model included treatment group (treated and untreated); age at enrolment; SMA type; SMN2 copy number; ambulatory status; scoliosis; MFM scale; MFM total score at baseline; time; treatment group*time interaction; and MFM total score at baseline*time interaction

CI confidence interval, *LS* least squares (a standard method for fitting a curve to a set of points), *MFM* Motor Function Measure, *MMRM* mixed model for repeated measures, *SD* standard deviation

Month 24, their missing results were assumed as missing at random, i.e., those with missing results behaved similarly to other patients with a similar covariate profile in the same treatment group. As shown in Table 6, at Month 12, the change from baseline in MFM total score was greater in the risdiplam group compared with the external comparator group, and this difference continued to increase at Month 24. The improvement in motor function in patients who received risdiplam treatment compared with the external comparator group was both clinically meaningful and highly statistically significant.

Figure 5 shows that risdiplam treatment in SUNFISH Part 1 led to an increase in mean MFM total score from baseline over 24 months, which was significantly different from the progressive decline observed in the untreated external comparator group [22].

After both 12 and 24 months of treatment, a significantly higher proportion of patients treated with risdiplam showed improvement (≥3-point change) in MFM total score compared with the untreated external comparator group (Fig. 6).

These results provided evidence of longer term efficacy of risdiplam in a broad population of patients with type 2 and 3 SMA compared with untreated patients. Supplemental analysis results from the weighted analysis comparing SUNFISH Part

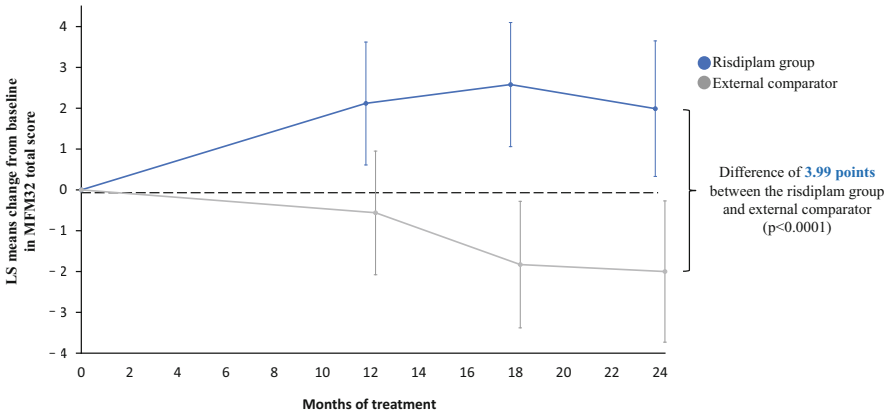

Fig. 5 Mean (LS means) change in motor function (as measured using the MFM) over 24 months in patients with Type 2 or 3 SMA in SUNFISH Part 1 compared with an external comparator group. Error bars represent the 95% CI. *CI* confidence interval, *LS* least squares, *MFM* motor function measure, *SMA* spinal muscular atrophy

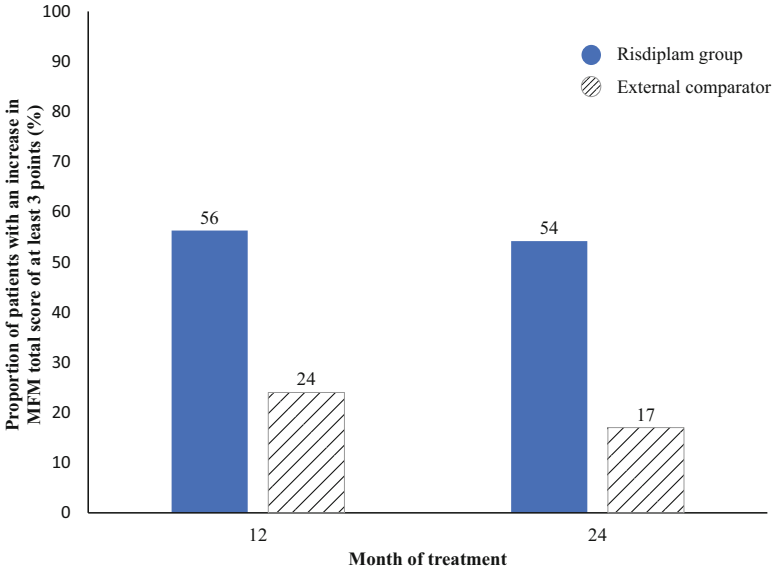

Fig. 6 Percentage of patients whose MFM score increased by at least 3 points compared with their score at the start of the study, over 12 and 24 months in patients with Type 2 or 3 SMA in SUNFISH Part 1 compared with an external comparator group. *MFM* motor function measure

1 with each of the two external comparator sources were generally in agreement with the above results, supporting the robustness of the conclusions. This retrospective weighted analysis comparing patients with type 2 and 3 SMA from SUNFISH Part 1 with two external comparator sources, the NatHis-SMA study and the placebo arm

of the olesoxime study, helped to further understand the benefits of risdiplam. In summary, the improvement in motor function with risdiplam treatment in SUNFISH Part 1 was markedly different from the untreated external comparator groups, where the expected decline in function inherent to this progressive disease was observed. The difference observed at Month 12 continued to increase at Month 24, supporting and confirming the benefits of prolonged exposure to risdiplam. This analysis also complements recent results from the placebo-controlled Part 2 of the SUNFISH study, in which the primary endpoint of the change from baseline in MFM32 total score at Month 12 was met. At month 12, the least squares mean (95% CI) change from baseline in MFM32 was 1.36 (0.61–2.11) in the risdiplam group and −0.19 (−1.22–0.84) in the placebo group, with a treatment difference of 1.55 (0.30–2.81, $p = 0.016$) in favor of risdiplam [29].

4 Discussion

The use of RWD and comparisons with external controls are a rapidly evolving area. In this section, we summarize the benefits of using RWD, especially in a rare disease setting. We also discuss some of the challenges we faced and lessons learned.

4.1 Benefits of Using RWD

The use of RWD was critical in clinical development planning, contextualizing the results and providing substantial evidence of efficacy of a disease-modifying therapy in a rare disease setting. Randomized, double-blind, placebo-controlled (or active controlled if standard of care exists) clinical trials are the gold standard. Randomization avoids systematic differences between groups with respect to known or unknown baseline variables that could affect outcomes. Blinding minimizes the bias due to subject and investigator expectations, and a placebo arm provides internal evidence of assay sensitivity. However, placebo-controlled trials with blinding are not always feasible or appropriate. When an effective therapy that is known to prevent death or irreversible morbidity exists for a particular patient population that population cannot usually be ethically studied in placebo-controlled trials; the particular conditions and populations for which this is true may be controversial (ICH E10 [30]).

Given the severity, rapid decline, and high mortality and morbidity in infants with type 1 SMA described in natural history, a comparator placebo group was not included in the FIREFISH study. In addition, in 2016 when the study started, there were no approved disease-modifying therapies for SMA, which precluded an active comparator arm. In the absence of a control arm, comparisons with external comparators or available natural history data are a valid approach in the development of new treatments for serious and rare diseases. Such comparisons are possible if

the natural history of the disease course is well understood, the external comparator group is similar to the treatment group (e.g., with regard to patient characteristics and endpoints), and a large treatment effect is expected to be seen with the study drug. In our case studies, the external comparators were carefully selected based on rigorous criteria described earlier.

RWD were also important in determining the anticipated treatment effect for the primary endpoint (MFM32) in the SUNFISH study as natural history data demonstrated that patients with type 2 and 3 SMA had a decline in motor function over time. In addition to clinical development planning, RWD were important in contextualizing study results for both FIREFISH and SUNFISH.

In FIREFISH, RWD from publications were used to define the performance criteria and benchmarks for success. Infants in FIREFISH attained motor milestones such as sitting without support that would never be achieved in infants with type 1 SMA without treatment. Infants in FIREFISH also achieved improved rates of event-free survival compared with those observed in natural history studies. These results confirmed that the disease course with risdiplam treatment substantially diverged from the natural history of the disease [11]. Infants treated with risdiplam also continued to benefit beyond Year 1. After 3 years of treatment in the FIREFISH study, event-free survival time was greatly improved compared with natural history [31].

In SUNFISH, the improvement in motor function with risdiplam treatment was markedly different from the untreated external comparator group, where the expected progressive decline in function inherent to SMA was observed. The comparison of SUNFISH data (Part 1 and later Part 2) with RWD also confirmed the longer term benefit of risdiplam over 24 months [32]. This was important because the placebo-controlled period in SUNFISH Part 2 was only for 1 year. RWD were used to provide substantial evidence of the efficacy of risdiplam, which was used to support registration and approval of risdiplam in different countries around the world.

The use of RWD for type 1 SMA was pivotal for regulatory decisions. The FDA stated, "... the study [*FIREFISH*] showed improvements in multiple clinical functional measures compared to the natural history of SMA, including motor function and developmental milestones as well as survival and ventilation free survival." The RWD used to contextualize the FIREFISH study results were included in the US prescribing information, "Of the patients who were treated with the recommended dosage of EVRYSDI 0.2mg/kg/day, 41% (7/17) were able to sit independently for ≥5 s (BSID-III, item 22). These results indicate a clinically meaningful deviation from the natural history of untreated infantile-onset [type 1] SMA. As described in the natural history of untreated infantile-onset SMA, patients would not be expected to attain the ability to sit independently, and no more than 25% of these patients would be expected to survive without permanent ventilation beyond 14 months of age." (EVRYSDI® prescribing information [21]).

The use of RWD for type 2 and 3 SMA was supplemental for regulatory decisions. In order to accelerate the regulatory review and approval timelines in the United States, the exploratory dose-finding SUNFISH Part 1 motor function

results were compared with RWD to contextualize the study results before the confirmatory SUNFISH Part 2 placebo-controlled results were available. The RWD used to contextualize the results in patients with type 2 and 3 SMA showed clear divergence between patients treated with risdiplam and untreated patients from natural history. The use of RWD also accelerated the filing and approval timelines for risdiplam, which was crucial given the high unmet need and the severe nature of the disease. Approval in the United States was expedited by at least 7 months.

4.2 Challenges

The role of RWD in providing substantial evidence of efficacy was different across regulatory regions. All regions accepted RWD as the benchmark for success for type 1 SMA. However, for patients with type 2 and 3 SMA, the acceptance of RWD as substantial evidence of efficacy differed across regions. In the United States, the FDA accepted an early filing based on FIREFISH Part 1 and SUNFISH Part 1, supplemented with placebo-controlled data from SUNFISH Part 2 during the review. Approval by the FDA was granted in August 2020. In contrast, in Europe, the EMA requested all data from SUNFISH Part 1 and Part 2 at the time of filing. Approval by the EMA was granted in March 2021. To overcome this challenge, a flexible filing strategy was implemented across regions.

At the time of our submission to the FDA in 2019, the recent fit-for-purpose framework for RWD to support regulatory decision making was not in place. Statistical considerations for fit-for-use RWD to support regulatory decision making in drug development can be found in a recent publication [33].

The requirement to provide individual patient-level data also varied across regions. The FDA required all individual patient data in Clinical Data Interchange Standards Consortium (CDISC) format including RWD, whereas the EMA did not. In order to file in the United States, the legacy data from the NatHis-SMA study were converted to CDISC standards for submission to the FDA. A good understanding of CDISC standards and regulatory requirements, technical skills, and upfront planning for this activity were critical to avoid a delay in submission timelines.

Although every effort was made to derive the performance criteria from infants who were as similar as possible to the FIREFISH study population, there were some limitations and challenges associated with this approach. These included potential differences in patient characteristics between the historical cohorts and study population; the limited number of studies available for some endpoints or small sample sizes of the historical cohorts; and studies being conducted in a limited number of countries or sites. For some endpoints, a performance criterion could not be derived as no suitable natural history studies were available. The results observed for motor function and motor milestone endpoints were consistent across natural history studies, but for other endpoints such as event-free survival, the results were more variable. These endpoints are more dependent on individual clinician practice

and caregiver preferences, in particular pulmonary and nutritional intervention strategies, and so may vary across countries and sites. Different definitions of permanent ventilation were also used in each study, including differences in the number of hours of ventilation per day, the number of consecutive days with this level of respiratory support, and the type of respiratory support provided. Despite these limitations, it should be noted that standard of care guidelines were considered during the selection of study sites, and the FIREFISH results were clearly differentiated from the natural course of type 1 SMA described in the literature.

Subtle differences in definitions of endpoints were also challenging for the SUNFISH and external comparator comparison. For example, MFM was administered differently across studies depending on a patient's age. To overcome this issue, an MFM total score was derived using either the MFM32 or MFM20 items depending on the patient's age. Missing data were also a challenge, with more missing data in the external comparator data sources. For example, for those included in the analysis, in the external comparator group, 73% of patients (i.e., 79 out of 109 patients) completed the Month 24 assessment, while in SUNFISH, all patients (i.e., all 48 patients included in the analysis) completed the Month 24 assessment. To deal with this, sensitivity analyses were performed to assess the robustness of the results using different methods for dealing with missing data.

4.3 Lessons Learned

The risdiplam clinical development program highlighted a number of key planning considerations that can be applied to other drug development programs in which RWD may play an important role, as follows:

1. *Incorporate RWD into the Clinical Development Plan*

We engaged with regulators early and pushed regulatory boundaries with robust arguments. The FDA eventually accepted a single-arm design for FIREFISH despite a preference for a placebo-controlled study. The overall study designs and choice of endpoints for both FIREFISH and SUNFISH incorporated Health Authority advice from both the EMA CHMP and the FDA. We pre-planned the statistical analysis for the confirmatory parts, documented this in a statistical analysis plan, shared the statistical analysis plan with Health Authorities, and asked for feedback in advance of filing.

Collaborating with healthcare providers and Patient Advisory Groups (PAGs) was also important to the success of the program. In particular, PAGs were actively involved in designing the NatHis-SMA study, advising on the interpretation of the clinical study results from both FIREFISH and SUNFISH, and validating the meaningfulness of the results from both a patient and caregiver perspective. This feedback was further supported by data collected by PAGs (e.g., a treatment

expectation survey conducted by SMA Europe [34]), which were incorporated into our regulatory dossiers to further contextualize our data.

2. *Take Steps to Reduce Bias When Using RWD*

We selected data sources that reflected considerations from ICH E10 to minimize bias when using external controls [30], including selecting a control population as similar as possible to the study population and selecting more than one external control. In FIREFISH, the identified RWD studies were documented and ranked based on the level of similarity of the patient population to the expected population in the FIREFISH study. The endpoints were also robust and objectively measured.

In SUNFISH, two independent studies, the NatHis-SMA study and olesoxime study, were selected based on their similarities to SUNFISH, such as the same efficacy outcome measure (change in MFM total score) and similar patient population (type 2 and 3 SMA and age range). In addition, some of the study centers collecting the external comparator data also enrolled patients in SUNFISH, and the first year of follow-up in the NatHis-SMA study occurred just before trial enrolment started for SUNFISH. These common features mitigated potential biases relating to different endpoint bias; selection bias; regional bias; different standards of care; and calendar time bias. The RWD in the external comparator group were weighted based on key prognostic factors to ensure the populations were as similar as possible. The important prognostic factors that were used to perform the weighting were defined a priori and described in the statistical analysis plan. It was also important to ensure that these data were available. For example, if severity of scoliosis was an important prognostic factor and was not collected in the RWD, it could not be used for the calculation of propensity scores.

It is also recommended to perform sensitivity analyses. For SUNFISH, the comparison was performed on the pooled external comparator data and separately for each study to support the robustness of the conclusions. A compilation of the different sources of biases from various types of external controls, such as those used in this chapter, and the potential mitigation steps can be found in a recent publication [35].

5 Conclusion

Risdiplam has now been approved in more than 90 countries worldwide, including the United States and the EU for the treatment of SMA in a broad patient population. In this case study, the use of RWD was pivotal in clinical development planning, contextualizing the results and providing substantial evidence of the efficacy of risdiplam, a disease-modifying therapy in a rare disease setting.

The results from the comparison of SMA patients treated with risdiplam from both the FIREFISH and SUNFISH studies versus RWD clearly diverged from the natural history of the disease and were clinically meaningful.

RWD were also critical in our filing strategy and led to significantly reduced approval timelines in the United States in a rare disease setting with a high unmet medical need. RWD were more widely accepted for objective endpoints with well-defined natural history (e.g., motor milestones and survival).

Acknowledgments The risdiplam clinical development program was funded by F. Hoffmann-La Roche Ltd in collaboration with the SMA Foundation and PTC Therapeutics. We thank all the patients who participated in the studies and the staff of clinical sites around the world. In addition, we thank Carol Reid who contributed to the statistical design and interpretation, Fani Petridis for valuable insights and Helena Bailes of Chrysalis Medical Communications for medical writing assistance.

References

1. Verhaart IE, Robertson A, Wilson I et al (2017) Prevalence, incidence and carrier frequency of 5q–linked spinal muscular atrophy–a literature review. *Orphanet J Rare Dis* 12(1): 1–15. https://doi.org/10.1186/s13023-017-0671-8.
2. Lefebvre S, Bürglen L, Reboullet S et al (1995) Identification and characterization of a spinal muscular atrophy-determining gene. Cell 80(1):155–165. https://doi.org/10.1016/0092-8674(95)90460-3.
3. Hamilton G & Gillingwater TH (2013). Spinal muscular atrophy: going beyond the motor neurons. Trends Mol Med 19(1):40–50. https://doi.org/10.1016/j.molmed.2012.11.002.
4. Mercuri E, Finkel RS, Muntoni F et al (2018) Diagnosis and management of spinal muscular atrophy: Part 1: Recommendations for diagnosis, rehabilitation, orthopedic and nutritional care. Neuromuscul Disord 28(2):103–115. https://doi.org/10.1016/j.nmd.2017.11.005.
5. Munsat TL, Davies KE (1992) International SMA consortium meeting. Neuromuscul Disord 2(5–6):423–8. https://doi.org/10.1016/s0960-8966(06)80015-5.
6. Zerres K, Rudnik-Schöneborn S (1995) Natural history in proximal spinal muscular atrophy. Clinical analysis of 445 patients and suggestions for a modification of existing classifications. Arch Neurol 52(5):518–523. https://doi.org/10.1001/archneur.1995.00540290108025.
7. D'Amico A, Mercuri E, Tiziano FD et al (2011) Spinal muscular atrophy. Orphanet J Rare Dis 6:71. https://doi.org/10.1186/1750-1172-6-71.
8. Darras BT (2015) Spinal muscular atrophies. Pediatr Clin North Am 62:743–766. https://doi.org/10.1016/j.pcl.2015.03.010.
9. Ratni H, Ebeling M, Baird J et al (2018) Discovery of Risdiplam, a Selective Survival of Motor Neuron-2 (SMN2) Gene Splicing Modifier for the Treatment of Spinal Muscular Atrophy (SMA). J Med Chem 61(15):6501–6517. https://doi.org/10.1021/acs.jmedchem.8b00741.
10. Baranello G, Darras BT, Day JW et al (2021) Risdiplam in Type 1 Spinal Muscular Atrophy. N Engl J Med 384(10):915–923. https://doi.org/10.1056/NEJMoa2009965.
11. Darras BT, Masson R, Mazurkiewicz-Beldzinska M et al. (2021) Risdiplam-treated patients with Type 1 Spinal Muscular Atrophy versus historical controls. N Engl J Med 385(5):427–435. https://doi.org/10.1056/NEJMoa2102047.
12. Bayley N. (2006). Bayley Scales of Infant and Toddler Development– Third Edition. Psychological Corporation.
13. Finkel RS, McDermott MP, Kaufmann, P et al (2014) Observational study of spinal muscular atrophy type I and implications for clinical trials. Neurology 83(9):810–817. https://doi.org/10.1212/WNL.0000000000000741.
14. Kolb SJ, Coffey CS, Yankey J et al (2017) Natural history of infantile-onset spinal muscular atrophy. Ann Neurol 82(6):883–891. https://doi.org/10.1002/ana.25101.

15. Finkel R, Bertini E, Muntoni F, Mercuri E; ENMC SMA Workshop Study Group. 209th ENMC International Workshop: Outcome Measures and Clinical Trial Readiness in Spinal Muscular Atrophy 7–9 November 2014, Heemskerk, The Netherlands. Neuromuscul Disord 2015 Jul;25(7):593–602.
16. Eichler HG, Bloechl-Daum B, Bauer P et al (2016) "Threshold-crossing": A useful way to establish the counterfactual in clinical trials?, Clinical Pharmacology and Therapeutics, 100(6):699–712. https://doi.org/10.1002/cpt.515.
17. Wang CH, Finkel RS, Bertini ES (2007) Consensus statement for standard of care in spinal muscular atrophy. J Child Neurol 22(8):1027–1049. https://doi.org/10.1177/0883073807305788.
18. Krosschell KJ, Kissel JT, Townsend EL et al (2018) Clinical trial of L-Carnitine and valproic acid in spinal muscular atrophy type I. Muscle Nerve 57(2):193–199. https://doi.org/10.1002/mus.25776.
19. Kolb SJ, Coffey CS, Yankey J et al (2016) Baseline results of the NeuroNEXT spinal muscular atrophy infant biomarker study. Ann Clin Transl Neurol 3(2):132–145. https://doi.org/10.1002/acn3.283.
20. De Sanctis R, Coratti G, Pasternak A et al (2016) Developmental milestones in type I spinal muscular atrophy. Neuromuscul Disord 26(11):754–759. https://doi.org/10.1016/j.nmd.2016.10.002.
21. EVRYSDI® prescribing information: https://www.gene.com/download/pdf/evrysdi_prescribing.pdf. Accessed July 2022.
22. Mercuri E, Baranello G, Boespflug-Tanguy O et al (2022) Risdiplam in Types 2 and 3 spinal muscular atrophy: a randomised, placebo-controlled, dose-finding trials followed by 24 months of treatment. Eur J Neurol [in press]. https://doi.org/10.1111/ene.15499.
23. Berard C, Payan C, Hodgkinson I and Fermanian J (2005) A motor function measure for neuromuscular diseases. Construction and validation study. Neuromuscul Disord 15(7):463–470. https://doi.org/10.1016/j.nmd.2005.03.004.
24. Vuillerot C, Payan C, Iwaz J et al (2013) Responsiveness of the motor function measure in patients with spinal muscular atrophy. Arch Phys Med Rehabil 94(8):1555–1561. https://doi.org/10.1016/j.apmr.2013.01.014.
25. Annoussamy M, Seferian AM, Daron A et al (2021) Natural history of Type 2 and 3 spinal muscular atrophy: 2-year NatHis-SMA study. Ann Clin Transl Neurol 8(2):359–373. https://doi.org/10.1002/acn3.51281.
26. Bertini E, Dessaud E, Mercuri E et al (2017) Safety and efficacy of olesoxime in patients with type 2 or non-ambulant type 3 spinal muscular atrophy: a randomised, double-blind, placebo-controlled phase 2 trial. Lancet Neurol 16(7):513–522. https://doi.org/10.1016/S1474-4422(17)30085-6.
27. Stürmer T, Webster-Clark M, Lund JL et al (2021) Propensity Score Weighting and Trimming Strategies for Reducing Variance and Bias of Treatment Effect Estimates: A Simulation Study. Am J Epidemiol 190(8):1659–1670. https://doi.org/10.1093/aje/kwab041.
28. Stuart EA (2010). Matching methods for causal inference: A review and a look forward. Stat Sci 25(1):1–21. https://doi.org/10.1214/09-STS313.
29. Mercuri E, Deconinck N, Mazzone E et al (2022) Safety and efficacy of once-daily risdiplam in type 2 and non-ambulant type 3 spinal muscular atrophy (SUNFISH part 2): a phase 3, double-blind, randomised, placebo-controlled trial. Lancet Neurol 21(1):42–52. https://doi.org/10.1016/S1474-4422(21)00367-7.
30. European Medicines Agency. ICH-E10 Choice of control group in clinical trials. January 20, CPMP/ICH/364/96. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-10-choice-control-group-clinical-trials-step-5_en.pdf
31. Servais L et al (2022) FIREFISH Parts 1 and 2: Safety and efficacy of risdiplam in Type 1 spinal muscular atrophy (3-year data). Abstract presented at 14th European Paediatric Neurology Society (EPNS) Congress, Glasgow, UK, 28 April – 2 May 2022.

32. Servais L et al (2022) SUNFISH Part 2: 24-month efficacy of risdiplam compared with external control comparators. Abstract presented at Muscular Dystrophy Association (MDA) Clinical and Scientific Conference, Nashville, TN, USA, 13–16 March 2022.
33. Levenson M, He W, Dharmarajan S et al (2022) Statistical consideration for fit-for-use real-world data to support regulatory decision making in drug development, Journal of Biopharmaceutical Statistics [in press].
34. Gusset N, Stalens C, Stumpe E et al (2021) Understanding European patient expectations towards current therapeutic development in spinal muscular atrophy. Neuromuscul Disord 31(5):419–430. https://doi.org/10.1016/j.nmd.2021.01.012.
35. Burger HU, Gerlinger, C, Harbron C et al (2021) The use of external controls: To what extent can it currently be recommended? Pharm Stat 20(6):1002–1016. https://doi.org/10.1002/pst.2120.

Index

W. He et al. (eds.), *Real-World Evidence in Medical Product Development*,
https://doi.org/10.1007/978-3-031-26328-6