Examining the Impact of Real-World Evidence on Medical Product Development: Proceedings of a Workshop Series (2019)

DETAILS

230 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-48829-7 | DOI 10.17226/25352

CONTRIBUTORS

Erin Hammers Forstag, Benjamin Kahn, Amanda Wagner Gee, and Carolyn Shore, Rapporteurs; Forum on Drug Discovery, Development, and Translation; Board on Health Sciences Policy; Health and Medicine Division; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine 2019. *Examining the Impact of Real-World Evidence on Medical Product Development: Proceedings of a Workshop Series*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25352.

Examining the Impact of Real-World Evidence on Medical Product Development

PROCEEDINGS OF A WORKSHOP SERIES

Erin Hammers Forstag, Benjamin Kahn, Amanda Wagner Gee, and Carolyn Shore, *Rapporteurs*

Forum on Drug Discovery, Development, and Translation

Board on Health Sciences Policy

Health and Medicine Division

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by contracts between the National Academy of Sciences and AbbVie Inc.; American Diabetes Association; Amgen Inc. (Contract No. GHC-COPS-CSARF-126570); Association of American Medical Colleges; AstraZeneca; Burroughs Wellcome Fund (Contract No. 1017664); Critical Path Institute; Eli Lilly and Company; FasterCures; Foundation for the National Institutes of Health; Friends of Cancer Research; GlaxoSmithKline; Johnson & Johnson; Merck & Co., Inc. (Contract No. CMO180308-0002499); National Institutes of Health (Contract No. HHSN263201200074I; Task Order No. HHSN26300093): National Cancer Institute, National Center for Advancing Translational Sciences, National Institute of Allergy and Infectious Diseases, National Institute of Mental Health, National Institute of Neurological Disorders and Stroke, Office of the Director; *New England Journal of Medicine*; Pfizer Inc.; Sanofi (Contract No. 33215505); Takeda Pharmaceuticals (Contract No. 53108); and the U.S. Food and Drug Administration (Grant No. 5R13FD005496-04 and Contract No. HHSP2233201400020B, Order No. HHSP23337055): Center for Drug Evaluation and Research, Office of the Commissioner. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-48829-7
International Standard Book Number-10: 0-309-48829-X
Digital Object Identifier: https://doi.org/10.17226/25352

Additional copies of this publication are available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2019. *Examining the impact of real-world evidence on medical product development: Proceedings of a workshop series*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/25352.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

PLANNING COMMITTEE ON EXAMINING THE IMPACT OF REAL-WORLD EVIDENCE ON MEDICAL PRODUCT DEVELOPMENT[1]

MARK MCCLELLAN (*Co-Chair*), Robert J. Margolis Professor of Business, Medicine, and Policy, and Director, Duke-Margolis Center for Health Policy, Duke University
GREGORY SIMON (*Co-Chair*), Senior Investigator, Kaiser Permanente Washington Health Research Institute
JEFF ALLEN, President and Chief Executive Officer, Friends of Cancer Research
ANDREW BINDMAN, Professor of Medicine, Epidemiology, and Biostatistics, University of California, San Francisco
JOHN GRAHAM, Senior Vice President, Medical Engagement and Value Evidence & Outcomes, GlaxoSmithKline (*from March 2018*)
ADAM HAIM, Chief, Treatment and Preventive Intervention Research Branch, National Institute of Mental Health, National Institutes of Health
MICHAEL HORBERG, Executive Director, Research, Community Benefit, and Medicaid Strategy, Mid-Atlantic Permanente Medical Group, and Director, Mid-Atlantic Permanente Research Institute
PETRA KAUFMANN, Director, Office of Rare Diseases Research and Division of Clinical Innovation, National Center for Advancing Translational Sciences, National Institutes of Health (*until May 2018*)
RICHARD KUNTZ, Senior Vice President and Chief Scientific, Clinical & Regulatory Officer, Medtronic, Inc.
ELLIOT LEVY, Senior Vice President, Global Development, Amgen Inc.
DAVID MADIGAN, Professor of Statistics, and Dean, Faculty of Arts and Sciences, Columbia University
DEVEN MCGRAW, Chief Regulatory Officer, Ciitizen
JOHN DAVID NOLEN, Physician Executive, Clinical Strategy, Cerner Corporation (*until October 2017*)
RICHARD PLATT, Professor and Chair, Department of Population Medicine, and Executive Director, Harvard Pilgrim Health Care Institute, Harvard Medical School
PATRICK VALLANCE, President, Research and Development, GlaxoSmithKline (*until March 2018*)

[1] The National Academies of Sciences, Engineering, and Medicine's planning committees are solely responsible for organizing the workshop, identifying topics, and choosing speakers. The responsibility for the published Proceedings of a Workshop Series rests with the workshop rapporteurs and the institution.

JOANNE WALDSTREICHER, Chief Medical Officer, Johnson & Johnson
MARCUS WILSON, Co-Founder and President, HealthCore, Inc.

Health and Medicine Division Staff

CAROLYN SHORE, Senior Program Officer (*from August 2017*)
ANNE CLAIBORNE, Senior Program Officer (*until August 2017*)
AMANDA WAGNER GEE, Program Officer
BEN KAHN, Associate Program Officer (*from October 2017*)
JONATHAN PHILLIPS, Associate Program Officer (*until October 2017*)
MELVIN JOPPY, Senior Program Assistant
ANDREW M. POPE, Director, Board on Health Sciences Policy

Consultant

ERIN HAMMERS FORSTAG, Science Writer

FORUM ON DRUG DISCOVERY, DEVELOPMENT, AND TRANSLATION[1]

RUSS B. ALTMAN (*Co-Chair*), Kenneth Fong Professor of Bioengineering, Genetics, Medicine, and (by courtesy) Computer Science, Stanford University
ROBERT M. CALIFF (*Co-Chair*), Vice Chancellor for Health Data Science, Duke University, and Scientific Advisor, Verily Life Sciences
CHRISTOPHER P. AUSTIN, Director, National Center for Advancing Translational Sciences, National Institutes of Health
LINDA BRADY, Director, Division of Neuroscience and Basic Behavioral Science, National Institute of Mental Health, National Institutes of Health
TANISHA CARINO, Executive Director, FasterCures
RICHARD DAVEY, Deputy Clinical Director, National Institute of Allergy and Infectious Diseases, National Institutes of Health
JAMES H. DOROSHOW, Deputy Director, Clinical and Translational Research, National Cancer Institute, National Institutes of Health
JEFFREY M. DRAZEN, Editor-in-Chief, *New England Journal of Medicine*
STEVEN K. GALSON, Senior Vice President, Global Regulatory Affairs and Safety, Amgen Inc.
CARLOS GARNER, Vice President, Global Regulatory Affairs, Eli Lilly and Company
JULIE L. GERBERDING, Chief Patient Officer and Executive Vice President, Strategic Communications, Global Public Policy, and Population Health, Merck & Co., Inc.
LYNN D. HUDSON, Chief Science Officer, Critical Path Institute
JEFF HURD, Regional Clinical Account Director, AstraZeneca
S. CLAIBORNE JOHNSTON, Dean, Dell Medical School, The University of Texas at Austin
KATHARINE KNOBIL, Chief Medical Officer, GlaxoSmithKline
FREDA C. LEWIS-HALL, Executive Vice President & Chief Medical Officer, Pfizer Inc.
ALLISON MCELVAINE, Vice President, Research & Scientific Programs, American Diabetes Association
ROSS MCKINNEY, JR., Chief Scientific Officer, Association of American Medical Colleges
JOSEPH P. MENETSKI, Deputy Director, Research Partnerships, Foundation for the National Institutes of Health

[1] The National Academies of Sciences, Engineering, and Medicine's forums and roundtables do not issue, review, or approve individual documents. The responsibility for the published Proceedings of a Workshop Series rests with the workshop rapporteurs and the institution.

BERNARD H. MUNOS, Founder, Innothink Center for Research in Biomedical Innovation
MICHAEL SEVERINO, Executive Vice President, Research & Development, and Chief Scientific Officer, AbbVie Inc.
RACHEL E. SHERMAN, Principal Deputy Commissioner, Office of the Commissioner, U.S. Food and Drug Administration
ELLEN V. SIGAL, Chair and Founder, Friends of Cancer Research
LANA R. SKIRBOLL, Vice President, Science Policy, Sanofi
BRIAN L. STROM, Chancellor, Rutgers Biomedical and Health Sciences, Rutgers University
AMIR TAMIZ, Director, Division of Translational Research, National Institute of Neurological Disorders and Stroke, National Institutes of Health
PAMELA TENAERTS, Executive Director, Clinical Trials Transformation Initiative, Duke University
JOHN WAGNER, Senior Vice President and Head of Clinical and Translational Sciences, Takeda Pharmaceuticals
JOANNE WALDSTREICHER, Chief Medical Officer, Johnson & Johnson
CARRIE WOLINETZ, Associate Director of Science Policy, Office of Science Policy, National Institutes of Health
JANET WOODCOCK, Director, Center for Drug Evaluation and Research, U.S. Food and Drug Administration

Forum Staff

CAROLYN SHORE, Forum Director
REBECCA ENGLISH, Program Officer
AMANDA WAGNER GEE, Program Officer
SYLVIA NCHA, Associate Program Officer
EESHAN KHANDEKAR, Research Associate
MELVIN JOPPY, Senior Program Assistant
ANDREW M. POPE, Director, Board on Health Sciences Policy

Reviewers

This Proceedings of a Workshop Series was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each published proceedings as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process.

We thank the following individuals for their review of this proceedings:

DEBORAH ASHBY, Imperial College London
JENNIFER GRAFF, National Pharmaceutical Council
PALL JONSSON, National Institute for Health and Care Excellence (United Kingdom)
SARAH LEATHERMAN, U.S. Department of Veterans Affairs

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the proceedings nor did they see the final draft before its release. The review of this proceedings was overseen by **DAN MASYS,** University of Washington. He was responsible for making certain that an independent examination of this proceedings was carried out in accordance with standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the rapporteurs and the National Academies.

Contents

Boxes, Figures, and Tables

BOXES

FIGURES

TABLES

Acronyms and Abbreviations

ACE	angiotensin converting enzyme
AE	adverse event
AFib	atrial fibrillation
aNSCLC	advanced non-small cell lung cancer
ARIA	Active Risk Identification and Analysis
ARISTOTLE	Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation
BEST	Biologics Effectiveness and Safety
BLA	biologics license application
Cal INDEX	California Integrated Data Exchange
CBER	Center for Biologics Evaluation and Research
CDER	Center for Drug Evaluation and Research
CDRH	Center for Devices and Radiological Health
CE	European conformity
CIOMS	Council for International Organizations of Medical Sciences
COPD	chronic obstructive pulmonary disease
CPT	Current Procedural Terminology
CRISP	Chesapeake Regional Information System for our Patients
CSR	Clinical Study Report
CTTI	Clinical Trials Transformation Initiative

EHR	electronic health record
EMA	European Medicines Agency
ENGAGE AF-TIMI 48	Effective Anticoagulation with Factor Xa Next Generation in Atrial Fibrillation Thrombolysis in Myocardial Infarction 48
FDA	U.S. Food and Drug Administration
GCP	good clinical practice
GSK	GlaxoSmithKline
ICD	*International Statistical Classification of Diseases and Related Health Problems*
ICH	International Conference on Harmonisation
IMDRF	International Medical Device Regulators Forum
IMI	Innovative Medicines Initiative
INR	international normalized ratio
InterSePT	International Suicide Prevention Trial
INVESTED	INfluenza Vaccine to Effectively Stop Cardio Thoracic Events and Decompensated heart failure
IRB	Institutional Review Board
JHHS	Johns Hopkins Health System
KP	Kaiser Permanente
MAPPs	Medicines Adaptive Pathways to Patients
MDEpiNet	Medical Device Epidemiology Network
MI	myocardial infarction
NEST	National Evaluation System for health Technology
NESTcc	NEST Coordinating Center
NHS	National Health Service
NICE	National Institute for Health and Care Excellence (United Kingdom)
NOAC	novel oral anticoagulant
NSAID	non-steroidal anti-inflammatory drug
OAC	oral anticoagulant
OHDSI	Observational Health Data Sciences and Informatics
OS	overall survival

PCORnet	National Patient-Centered Clinical Research Network
PEER	Platform for Engaging Everyone Responsibly
PFS	progression-free survival
PGHD	patient-generated health data
PMA	premarket approval
PPRN	patient-powered research network
PRO	patient-reported outcome
PROTECT	Pharmacoepidemiological Research Outcomes of Therapeutics by European Consortium
PTSD	posttraumatic stress disorder
PXE	pseudoxanthoma elasticum
RCT	randomized controlled trial
RE-LY	Randomized Evaluation of Long-term Anticoagulation Therapy
ROCKET AF	Rivaroxaban Once Daily Oral Direct Factor Xa Inhibition Compared with Vitamin K Antagonism for Prevention of Stroke and Embolism Trial in Atrial Fibrillation
RWD	real-world data
RWE	real-world evidence
SOC	standard of care
SOP	standard operating procedure
SSRI	selective serotonin reuptake inhibitor
TAVR	transcatheter aortic valve replacement
TTNT	time to next treatment
TTP	time to progression
TTTD	time to treatment discontinuation
VA	U.S. Department of Veterans Affairs

1

Introduction[1]

Randomized controlled trials (RCTs) have traditionally served as the gold standard for generating evidence about medical interventions. However, RCTs have inherent limitations and may not reflect the use of medical products in the real world (e.g., specific therapeutic interventions may perform differently within different patient cohorts based on age, gender, race, ethnicity, disease severity, comorbidities, or polypharmacy). Additionally, RCTs are expensive, time consuming, and cannot answer all questions about a product or intervention. Evidence generated from real-world use—based on sources such as patient registries, electronic health records (EHRs), and medical claims data—may provide valuable information, alongside RCTs, to inform medical product decision making. This supplemental (or complementary) information is generally based on analysis of information gathered in the "real world" of routine clinical practice outside of a tightly controlled RCT. As standards and rigor for the collection and analysis of real-world data (RWD) evolve and improve over time, stakeholders will have the opportunity to explore new areas for which RWD and real-world evidence (RWE) may be used to answer scientific questions and guide more effective and cost-efficient medical product decision making.

[1] These workshops were organized by an independent planning committee whose role was limited to planning the workshops, and the Proceedings of a Workshop Series was prepared by the workshop rapporteurs and staff as a factual summary of what occurred at the workshops. Statements, recommendations, and opinions expressed are those of individual presenters and participants, and are not necessarily endorsed or verified by the National Academies of Sciences, Engineering, and Medicine, and they should not be construed as reflecting any group consensus.

Regulatory agencies and medical product sponsors are increasingly interested in incorporating RWE into their programs. For example, the U.S. Food and Drug Administration (FDA) has committed, in both the Prescription Drug User Fee Act VI (for drugs and biologics) and the Medical Device User Fee Amendments IV (for devices), to developing policies that support the use of RWE in medical product evaluation (see remarks by Mark McClellan, director of the Duke-Margolis Center for Health Policy, in Chapter 1 and by Rachael Fleurence, executive director of the National Evaluation System for health Technology [NEST] Coordinating Center in Chapter 3). Medical product developers are already successfully using data sources such as registries and techniques such as historical comparator arms to earn marketing authorization (see remarks by Fleurence in Chapter 3; Scott Gottlieb, commissioner of FDA, in Chapter 1; Janet Woodcock, director of FDA's Center for Drug Evaluation and Research [CDER], in Chapter 5; Steven Anderson, director of the Office of Biostatistics and Epidemiology at FDA's Center for Biologics Evaluation and Research, in Chapter 10; Jacqueline Corrigan-Curay, director of the Office of Medical Policy at CDER, in Chapter 10; Jeff Shuren, director of FDA's Center for Devices and Radiological Health, in Chapter 10; and speakers throughout). Furthermore, FDA has supported projects that are designed to incorporate RWE into its decision-making and safety monitoring processes, such as Sentinel (see remarks by Richard Platt, professor and chair in the Harvard Medical School Department of Population Medicine at the Harvard Pilgrim Health Care Institute, in Chapter 3). Moreover, the recently enacted 21st Century Cures Act mandates that FDA develop a framework for using RWE in support of new indications and to satisfy postapproval studies. Europe is likewise engaged in developing policies and infrastructure to support additional use of RWE. The European Medicines Agency is developing the EudraVigilance network for safety monitoring and reporting and the Adaptive Licensing pathway for product marketing authorizations (see remarks by Alasdair Breckenridge, emeritus professor of Clinical Pharmacology at the University of Liverpool, in Chapter 10). The European Union and the regulated industry are supporting several RWE initiatives through the Innovative Medicines Initiative (see remarks by Breckenridge and Pall Jonsson, associate director of Research and Development at the United Kingdom's National Institute for Health and Care Excellence [NICE], in Chapter 10).

To explore the potential for using RWE in medical product decision making, an ad hoc workshop planning committee of the National Academies of Sciences, Engineering, and Medicine planned a three-part workshop series, sponsored by FDA and hosted by the National Academies' Forum on Drug Discovery, Development, and Translation (see Box 1-1 for the Statement of Task). The series was designed to examine the current system of

BOX 1-1
Workshop Series Statement of Task

An ad hoc committee of the National Academies of Sciences, Engineering, and Medicine will plan and conduct a three-part workshop series to be held over the course of a 2-year period. As part of the Food and Drug Administration's (FDA's) continued focus on building a national governance system for evidence generation, the proposed focus of these workshops will be on the generation and utilization of real-world evidence (RWE) to evaluate efficacy, effectiveness, tolerability, and safety for both review of new indications and postapproval studies. These workshops would include presentations and perspectives from thought and knowledge leaders representing a range of disciplines, including but not limited to federal regulatory and funding agencies, clinical and academic medicine and research, medical professional organizations, the regulated biopharmaceutical industry, patients and patient-focused and disease-advocacy organizations, payers, consumer organizations, health systems, and other interested stakeholders that represent the myriad views of those involved in drug, biologic, and device discovery, development, translation, and regulation. The workshop audiences are expected to be similarly diverse, and they will have opportunities to engage in discussion during the workshops. The objectives of the workshops are

- To advance discussions and common knowledge among key stakeholders (including FDA and the public) about complex issues relating to the generation and utilization of RWE; and
- To foster development and implementation of the science and technology of RWE generation and utilization.

Topics to be covered at the workshop include

- Aligning incentives and addressing barriers to support collection and use of high-quality evidence derived from real-world data sources in health product review, payment, and delivery;
- Definitions surrounding the core components of RWE;
- Sources of data that are curated, standardized, and analyzed to derive RWE, such as safety surveillance, observational studies, registries, claims, or patient-centered outcomes research;
- Gaps in data collection activities, and priority areas and pilot opportunities that RWE incorporation could address;
- Standards and methodologies for collecting and analyzing RWE in support of new indications or postapproval studies, and the circumstances under which that evidence could be applied;
- Applications for using RWE to supplement traditional clinical trials, pragmatic/effectiveness trials, or routine clinical applications;
- Mechanisms for determining which discrete types of RWE could support regulatory decisions; and
- Operational challenges and barriers for generating and incorporating RWE in the context of a learning health system and how clinicians can best be involved in the collection and utilization of RWE.

evidence generation and its limitations, to identify when and why RWE may be an appropriate type of evidence on which to base decisions, to learn from successful initiatives that have incorporated RWE, and to describe barriers that prevent RWE from being used to its full potential. Issues regarding RWE unrelated to evidence development, use, and application were not discussed in detail during the series.

At the first workshop, held in September 2017, participants heard from stakeholders—patients, providers, and payers—about what kind of data they need to make decisions, and explored how to generate this "fit-for-purpose" evidence. Researchers who have successfully used RWE in their work were invited to share their successes, and workshop speakers and participants discussed the challenges and misaligned incentives that prevent RWE from being used more widely. Finally, speakers presented their perspectives on the shortcomings and limitations of the current system of evidence generation, and discussed how integrating RWE into the system could improve decision making for medical product development and evaluation.

The second and third workshops, held in March and July 2018, focused on illuminating when it may be appropriate to use RWE and which questions RWE may help address, and identifying key questions that stakeholders might consider when collecting or using RWE. These questions were used to draft RWE study design decision aids; the aim of these decision aids was to help stakeholders make thoughtful choices about the development and design of studies involving RWE.

The decision aids (discussed in Chapters 6 through 9) served as a starting point for discussions at the third workshop, and were informed by discussions that took place during the first and second workshops in this series. The decision aids were drafted by some individual participants of the first and second workshops, with additional input by attendees of the third workshop. The decision aids were developed to guide discussion at the third workshop, and they may also be useful in helping workshop attendees and other stakeholders think about and evaluate opportunities to use RWD and RWE for medical product decision making and make informed choices about the design of prospective or retrospective studies, primarily for regulatory review (akin to clinical decision aids that are designed to help patients make informed decisions about treatment options). There are no right or wrong answers to a given question. Instead, the decision aids lay out key questions for stakeholders to consider early on to help make thoughtful choices about the development and design of rigorous, but manageable, RWE studies that relevant parties (e.g., patients, clinicians, researchers, sponsors, regulators, payers) agree in advance will generate reliable results.

The questions in the decision aids aimed at capturing relevant information about the potential risks associated with either the treatments themselves or trade-offs associated with particular decisions; costs in terms of monetary investment, time investment, and/or patient and clinician investment; and reporting and transparency expectations for showing study methods and results. The decision aids can be further modified and refined to be broadly applicable across study types (e.g., studies on medical products to treat prevalent chronic diseases), account for different types of data sources (e.g., EHRs, claims data, patient-generated data), and "future-proofed" to accommodate new sources of data going forward (e.g., data from mobile devices, the Internet of things). The decision aids were divided into four topic areas:

1. When a particular RWD element is fit to assess study eligibility, treatment exposure, or outcomes;
2. Considerations for obscuring intervention allocation in trials to generate RWE;
3. Considerations for controlling or restricting treatment quality in real-world trials; and
4. Assessing and minimizing bias in observational comparisons.

At the third workshop, speakers and participants used the decision aids to explore the four topic areas and used case studies to explore and illuminate the practical, ethical, technological, and scientific issues that arise when designing RWE studies for decision making. Throughout the workshop series, attention was paid to how evidence is used in regulatory decision making, and how RWE may be incorporated into this process.

This workshop proceedings was prepared by rapporteurs in accordance with National Academies guidelines, and is a summary of the discussions held during the three workshops. The proceedings is divided into 10 chapters. Chapter 1 introduces the topic and frames the workshop series; Chapter 2 presents various stakeholder perspectives on RWE; Chapter 3 discusses successful examples of the use of RWE in research; Chapter 4 discusses barriers and disincentives to the use of RWD and RWE; and Chapter 5 considers RWE in the context of the current evidence-generation system. Chapters 6 through 9 focus primarily on in-depth discussions from the second and third workshops. Chapter 6 discusses the use of RWD; Chapter 7 discusses treatment quality in real-world research; Chapter 8 focuses on obscuring intervention allocation; and Chapter 9 discusses observational data research. Chapter 10 concludes the proceedings and describes topics that were discussed throughout the workshop series.

DEFINING THE TERMS

For these proceedings, definitional terms used by other organizations and descriptions of RWD and RWE that emerged over the workshop series are outlined here. Gregory Simon, senior investigator at Kaiser Permanente Washington Health Research Institute, said it is essential to establish a common language to describe RWD and RWE in order to discuss and potentially change the paradigm of evidence generation. However, he said, RWD and RWE are not easy to define, and different institutions use varying definitions (see Box 1-2). First, said Simon, it is important to differentiate between RWD and RWE: RWD is a necessary, but not sufficient, condition to produce RWE. Simon said that while RWE is sometimes thought of as "everything but a randomized trial," this is not accurate: RWE can sometimes involve randomization, and not all non-RWE is randomized. A better way to think about RWE, said Simon, is to identify its core characteristics:

- RWE is generalizable: The answers available through RWE will generally be true if they are implemented in the future.
- RWE is relevant: It seeks to directly provide information that stakeholders need to make decisions. In other words, RWE is "fit for purpose," meaning that the evidence is capable of answering a research question, regardless of its source.
- RWE is adaptable: When evidence is generated in the real world, by necessity it incorporates the broad heterogeneity of real patients and real providers.
- RWE is efficient: Evidence can be produced more quickly and at a lower cost than traditional methods.

Mark McClellan, director of the Duke-Margolis Center for Health Policy, offered the definitions of RWD and RWE that were developed through the work of a collaborative agreement between FDA and the Duke-Margolis Center. RWD are "data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources." RWE, said McClellan, is "evidence derived from RWD through the application of research methods." In the specific context of RWE for regulatory applications, RWE can be further defined as "clinical evidence regarding the use and potential benefits or risks of a medical product derived from analysis of RWD."

McClellan said there is a common misconception that RWD simply means observational data. He stressed that the key characteristic of RWD is not about the research method used, but instead pertains to the provenance of the data. RWD are data that are "part of the routine delivery of care," such as clinical records or insurance claims. In addition to these clinical

BOX 1-2
Definitions of Real-World Data and Real-World Evidence from Various Stakeholders

There is no consensus on the definitions of real-world data (RWD) and real-world evidence (RWE), and different institutions describe and use these terms differently:

- **21st Century Cures Act:** RWE is "data regarding the usage, or the potential benefits or risks, of a drug derived from sources other than traditional clinical trials" (21 U.S. Code § 355g).
- **U.S. Food and Drug Administration:** RWD are the 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 analysis of RWD (https://www.fda.gov/scienceresearch/specialtopics/realworldevidence/default.htm, accessed December 17, 2018).
- **Eli Lilly and Company:** RWE is one form of evidence (along with randomized controlled trials [RCTs], health economics studies, etc.) derived from primary or secondary real-world data sources, with appropriate design/analyses, for the purpose of providing insights, on diseases, medicines, patient populations, and health care practices, that will inform customer and internal decision making (Yaist presentation, July 17, 2018).
- **Innovative Medicine Initiative's GetReal consortium:** "RWD" is an umbrella term for data regarding the effects of health interventions (e.g., benefit, risk, and resource use) that are not collected in the context of conventional RCTs. Instead, RWD are collected both prospectively and retrospectively from observations of routine clinical practice. Data collected include, but are not limited to, clinical and economic outcomes, patient-reported outcomes, and health-related quality of life. RWD can be obtained from many sources, including patient registries, electronic medical records, and claims databases (Goettsch and Makady, 2015).
- **International Society for Pharmacoeconomics and Outcomes Research:** RWD are data used for clinical, coverage, and payment decision making that are not collected in conventional RCTs (Berger et al., 2017).

data, RWD also can include patient-generated data that are not part of an encounter with the health system, for example, data from a smartphone. Turning RWD into RWE is not an automatic process, said McClellan; it requires working with the data and applying appropriate research methods to turn the data into quality, useful evidence. Research methods may include randomized trials, prospective or retrospective approaches, observational approaches, and cluster randomization. These methods are used to turn a

deluge of data—RWD—into evidence that is fit for purpose for a specific application, RWE.

Eleanor Perfetto, senior vice president of strategic initiatives at the National Health Council, added that the difficulty in defining RWD and RWE is contributing to confusion in the patient advocacy community. Patient advocates and individual patients are interested in using RWE to make medical decisions and to advocate for regulatory and payment decisions, but the confusion over terms is making it difficult for them to fully understand what RWE is (and is not), and to understand the benefits and limitations of RWE. Rachel Sherman, principal deputy commissioner of FDA, concluded that while using clear and understandable language is important, "The goal is not to define RWD and RWE. The goal is to get better information and to do it in a more sensible way."

THE U.S. FOOD AND DRUG ADMINISTRATION'S VISION FOR THE FUTURE

Simon opened the first workshop with an ode to *Fiddler on the Roof*, a 1960s musical about maintaining Jewish traditions in the face of change. Simon noted that in *Fiddler on the Roof*—and in scientific research—there are some traditions that are vital to the central purpose of the community, while other traditions are merely followed because they always have existed. Discerning which traditions are important and which can be let go, said Simon, is critical for moving scientific progress forward efficiently and effectively. FDA commissioner Scott Gottlieb concurred with this analogy in his keynote address, saying that expanding traditional notions of evidence generation to incorporate the use of RWE could help make the medical product development process more efficient and more cost-effective. In addition, Gottlieb said, RWE could help doctors and patients to be better informed and make better decisions, which will ultimately help achieve better health care outcomes.

Despite these potential advantages, there is uncertainty about the role that RWE should play in making regulatory decisions, said Gottlieb. RWE is already being used for decision making by many stakeholders in the medical community—payers in particular—and it is time to "close the evidence gap between the information we use to make FDA's decisions" and the information being used by others. As the use of RWE is increasing, so is the rigor with which it is collected and the reliability of the data. One of the scientific research community's longstanding traditions, said Gottlieb, is the hierarchy of evidence, in which randomized prospective placebo-controlled trials are at the top. Even as tools and technologies for collecting and using other forms of evidence, including RWE, have progressed, the hierarchy of evidence remains unchanged.

FDA, said Gottlieb, needs to find ways to leverage these new constructs to better inform decisions. To do so, FDA must "support the development of, and access to, appropriate forms of reliable evidence that meet our standard for approval." Just as clinical decisions are often made based on a "mosaic of information," regulatory decisions could likewise consider a broad range of informative sources of evidence such as RWE. Gottlieb noted that there are no statutory or regulatory barriers to incorporating RWE, and that using RWE to make decisions about product marketing would be consistent with FDA's practice of using RWE to make decisions about safety. Data from the real world, said Gottlieb, may even be more rich, diverse, and informative than data from RCTs that "speak to a limited and rigidly constructed circumstance."

When FDA approves a medical product, a line is by necessity drawn between safety and efficacy on one side, and risk and uncertainty on the other, said Gottlieb. However, FDA as a public health agency has a mandate to embrace the full continuum of evidence available from all sources along the entire life cycle of a product. "We can't allow our need for a point of regulatory accountability to prevent us from looking across the line we have to draw at practical information that's collected both before and after our point of demarcation when a product gains a license for initial market entry." RWE, said Gottlieb, could help FDA make more informed decisions along the continuum, from providing data about the benefit–risk profile of a product, allowing for early identification and a richer understanding of safety concerns.

To enable greater adoption of RWE for regulatory decisions, FDA will need to work with the health care system to change the way that clinical information is collected, said Gottlieb. Currently, structured data within EHRs are usually geared toward billing, and clinically relevant information is often hidden in unstructured notes that are inaccessible. While data from EHRs are not perfect—for example, data might be missing or nonsensical—there are new tools and technologies that allow for electronic audits of the integrity of the data that can give FDA and other stakeholders more confidence in the quality of the data they are using. In addition to data from EHRs, there is a need to collect more information directly from patients themselves. For example, a tool could collect information about gait and physical activity directly from an individual, rather than having a doctor do a walk test on a treadmill in the office. These tools will need to be validated, and because they are relatively new, both the product developer and the regulators will need to take a "leap of faith" to get these products on the market, he said.

In conclusion, Gottlieb stated that FDA "needs to do its part to advance the use of RWE." FDA is taking steps to provide more clarity about its approach to the use of RWE in regulatory decisions, said Gottlieb, including

final guidance that was issued in August 2017[2] about the use of RWE in the development of devices. Gottlieb also noted that in the past 3 years, FDA has approved or cleared at least eight new medical devices and expanded the use of at least six technologies based on evidence derived from RWD. This evidence, said Gottlieb, was generated in less time and at a lower cost than in the past, and using RWE saved up to 2 years of development time. The adoption of RWE in the context of medical device regulation is more straightforward than in the context of drug regulation, said Gottlieb, because the end user of a device is usually a clinician, who is in a position to collect information, whereas the end user of a drug is a patient, who may not be able to collect and share information as readily. However, despite these challenges, FDA is currently working on policies to support the use of RWE in the approval of new indications for already marketed drugs, which may be especially relevant for drugs for rare diseases or unmet medical needs. Although RWE is not likely to replace traditional clinical data in many cases, Gottlieb said, there is an opportunity to incorporate RWE into FDA's entire life cycle approach to medical product development.

FIT-FOR-PURPOSE EVIDENCE

McClellan followed Gottlieb's discussion about FDA's vision with a presentation about a cooperative agreement between FDA and Duke-Margolis. This partnership is driven by bipartisan legislative action by Congress[3] that directed FDA to further explore using RWE in the regulatory framework. The legislation specifically requires FDA to hold workshops, evaluate the potential use of RWE, and issue draft guidance by the end of 2021. As part of these activities, FDA partnered with Duke-Margolis for the purpose of exploring the use of RWE for regulatory purposes and sponsored the three-part workshop series hosted by the Forum on Drug Discovery, Development, and Translation described in this proceedings. The Forum workshop series provided an ongoing venue to discuss overarching governance issues associated with using RWE for evaluating drugs, biologics, and devices. The complementary work at Duke-Margolis focused primarily on drugs and discussed more technical issues related to RWE application. Duke-Margolis released a white paper in September 2017[4] that proposes a framework that can be used to guide sponsors and FDA in discussions about RWE, and seeks to clarify the current landscape of RWD/RWE for regulatory use.

[2] See https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm513027.pdf (accessed November 6, 2018).

[3] Prescription Drug User Fee Act VI and the 21st Century Cures Act.

[4] See https://healthpolicy.duke.edu/sites/default/files/atoms/files/rwe_white_paper_2017.09.06.pdf (accessed November 6, 2018).

In particular, the white paper discusses the various considerations that affect how researchers decide to structure a study that uses RWD, including the regulatory context, the clinical context, the availability of quality data, and the research methods for the evidence needed. McClellan said this framework can be used for thinking about what data and methods would be appropriate for a specific use of RWE. For example, many different methods can be used to turn RWD into RWE, including randomized methods, prospective or retrospective observational methods, or a hybrid of methods. The method that a researcher chooses depends on the intended use of the RWE, taking into consideration the clinical question, the type of data available, and the regulatory purpose for which the evidence is being generated. McClellan noted that regardless of the chosen method, it is unlikely that "one single real-world evidence study is going to be all the evidence that exists for most of these regulatory questions." FDA looks at the totality of the evidence when making regulatory decisions, and RWE can be a part of the overall picture.

Another area of focus of the Duke-Margolis partnership with FDA, said McClellan, has been on the challenges of data. Although there is a growing quantity of RWD, ensuring the quality of RWD is considerably more difficult, said McClellan. There are concerns about making sure the data are of good provenance, and could be traceable back to the source if questions arose. Patient-generated data are increasingly available, but turning these data into quality RWE requires patient and provider support and adoption as well as good governance and stewardship practices. McClellan said there are promising tools and technologies that may help mitigate the challenges of data, such as blockchain, which can enable more secure aggregation of patient data from a variety of devices and sources.

The partnership has also been discussing methods and how to ensure that the research methods used are appropriate for the regulatory purpose, said McClellan. Several basic good practices for RWD were identified by the partnership, including

- Developing analytic plans that are transparent and specified in advance;
- Using robust primary data sources;
- Ensuring there are enough quality data to address the endpoints and conclusions;
- Making sure that methods fit into a "totality of the evidence" approach; and
- Using randomization when possible and appropriate.

McClellan concluded that "having a clear path to regulatory acceptability" for RWE could be "a big driver toward getting more robust and

interpretable" RWD, including clinic- and patient-generated data. This move toward better data could be "synergistic with a move toward new payment models" that focus on value and patient outcome, and together these movements could generate support for infrastructure needed for RWE studies. However, said McClellan, further work is needed. The practical use of RWE is still limited, and it will take further investment by stakeholders to build a foundation for the use of RWE for regulatory purposes, as well as to improve clinical practice. McClellan noted that some stakeholders—health insurance companies in particular—are making major investments in data infrastructure, with a goal of integrating data to support better decisions for patients. These types of efforts will be highly relevant to advancing the generation of RWE for regulatory purposes, he said, and will help to integrate the regulatory framework with a framework based on patient outcomes.

2

Perspectives on Real-World Evidence

Key Messages Identified by Individual Workshop Participants

- One of the most critical factors in using real-world evidence (RWE) is making it fit for purpose and the needs of end users. (Bindman, Califf)
- Payers are constantly balancing concerns between patient access and affordability of therapeutics; value-based agreements between drug manufacturers and regulators, and subsequent data collections based in such agreements, could support ongoing real-world data collection and create a more stable drug pricing system. (M. Sherman)
- Delivery systems value medical practices that are supported by quality, relevant evidence that demonstrates value for patients; typical evidence generation practices for such evidence include traditional clinical trials, delivery system data collection, and reviews of current practices and literature. (Ford, Horberg)
- Patient-generated data and community-led registries can be an important source of evidence generation because they focus on patient priorities and lived experiences; these data sources require rigorous validation, but should be integrated into clinical decision making. (Terry)
- Patients are most concerned with evidence that will inform their clinical care decisions, and they may also tend to believe that data belong to them and should be treated that way. (Perfetto)

- There are promising opportunities to use real-world trials to generate evidence, such as randomized pragmatic trials, platform trials, and master protocols. However, the bar for quality in RWE studies should be similarly high to the bar for randomized controlled trials; it is important to have regulatory buy-in before embarking on an RWE study. (Waldstreicher)

As discussed in Chapter 1, one of the core characteristics of real-world evidence (RWE) is that it is "fit for purpose." That is, evidence should be capable of answering a research question, even if it was originally generated for a purpose other than research. Several workshop participants said that regardless of the source of the data or the research method employed, the most critical element of using RWE is making sure that the evidence is fit for purpose. Andrew Bindman, professor of medicine, epidemiology, and biostatistics at the University of California, San Francisco, said there are many ways to arrive at scientifically valid evidence, so the means by which evidence is generated should be adaptable to the needs of the end user. Robert Califf, vice chancellor of health data science at Duke University and scientific advisor at Verily Life Sciences, agreed, and said that different end users not only have different evidence needs, but they also need different degrees of certainty in the data. These end users could be payers, providers, patients, or others; each has his or her own unique perspective and needs for evidence. The workshop participants heard from four speakers representing three different groups: payers, delivery systems, and patients. The speakers discussed how they use evidence to make decisions, and were asked to identify how RWE does or could inform their decision-making process.

PAYER PERSPECTIVE

The main challenge for payers, said Michael Sherman, senior vice president and chief medical officer at Harvard Pilgrim Health Care, is balancing access and affordability while at the same time driving innovation. Sherman said one out of every four dollars that Harvard Pilgrim spent in 2016 was on drugs, and that Harvard Pilgrim members spent more than 20 percent of their out-of-pocket health care dollars on drugs. Sherman stressed that for drugs that keep people out of the hospital, extend life, and improve chronic disease, "cost should not be a barrier," but that drugs with less certain or less profound impacts may not have the same value.

With this tension between access and affordability in mind, Sherman moved on to the topic of generating evidence for medical interventions. Clinical trials, said Sherman, are not always relevant to the real world of medical practice. In clinical trials, he emphasized, the investigators are

experts, the patients are carefully selected, there is full compliance with all protocols, and patients are closely monitored to address concerns and ensure adherence. In the real world, by contrast, providers are not necessarily experts in every field, work with large patient loads, and may have their own biases and ways of practicing, Sherman said. In particular, new diagnostics offer challenges; while a test may give information to physicians and patients about what course of action to take, there are many other considerations that inform treatment decisions, so the value of the diagnostic is unclear. Sherman said that even for commonly accepted treatments, evidence-based medicine is not always followed. For treatments or diagnostics that are newer, more complicated, and more expensive, it is not always clear if the investment is worthwhile because decisions made by the physician and patient may not be fully aligned with the results of these diagnostics, he said.

Financial pressures on payers are increasing, said Sherman, particularly as innovation is resulting in new drugs and therapies such as gene therapy. Although these innovations are very exciting, he said, evidence must show they are worth the price. Unfortunately, data are limited for some of these new innovations, and some recent U.S. Food and Drug Administration (FDA) approvals have been seen as "overly broad" and based on limited evidence, said Sherman. When data are limited, pricing is uncertain, and clinical variability is unknown, payers are "understandably a little concerned" about paying for these new innovations. Sherman noted, however, that FDA has a difficult job, and that patients and families are reasonable to expect access to drugs that can give hope, even when evidence is limited. For some rare conditions, there will never be the possibility of a high-quality, randomized clinical trial of sufficient size; for other conditions, the time that it would take to generate proper data is significant, and during that time, patients are suffering. This is an ethical "tightrope" that patients, providers, regulators, and payers all must walk, said Sherman.

Sherman offered suggestions that could help to balance concerns about access and affordability for such approvals that may have been made based on limited evidence, while also encouraging innovation in research and development. As a caveat, Sherman noted that FDA is limited in what it can require of manufacturers. That said, he suggested that in cases where an approval may be based on limited evidence, FDA could consider

- Requiring manufacturers to enter into value-based agreements that tie reimbursement to success of the drug (tied to outcome measures used to gain approval);
- Requiring manufacturers to submit data to an objective third party (e.g., Institute for Clinical and Economic Review) and agree to pricing that aligns with findings; and/or

- Encouraging postmarketing payer–pharmaceutical company collaboration to use data generated by these value-based agreements.

Sherman said this proposal could have several benefits. First, it could create a structure for payers in which they could offer a drug, but would only be required to pay for it if it was shown to actually add value. Second, it could provide transparency and certainty to the pricing of drugs, and the value of the drug would be reflected in the price. Third, it could create RWE in the normal course of practice—as drugs are made available to patients, data would be collected and analyzed to study the patient outcomes and value of the drug. Finally, this process could reduce uncertainty for companies that are submitting drugs for approval. Currently, a drug may be approved, but there is still uncertainty about whether insurers will pay for it, which creates frustration for the companies as well as patients and providers. This type of proposal, said Sherman, could increase transparency and certainty, help to grow the evidence base for new innovations, and improve access to life-saving drugs for patients.

DELIVERY SYSTEMS PERSPECTIVE

Concurring with many of Sherman's remarks was Michael Horberg, executive director of research, community benefit, and Medicaid strategy and the Mid-Atlantic Permanente Research Institute at the Mid-Atlantic Permanente Medical Group, Kaiser Permanente (KP). In particular, he agreed that a "clinical trial is often an idealized version of what we would all hope the care would be." In the real world, there are issues with patient adherence and care delivery that can impact the effects of a medical intervention. Because KP both delivers care and pays for the care, said Horberg, medical practices must be backed up with quality, relevant evidence that shows a benefit for patients. When assessing the evidence base for a new intervention, KP looks at a variety of considerations, including

- Who conducted the studies (e.g., KP, industry, or government funded)?
- Who is the population at risk? Do the data reflect this population, or can data be generated to reflect this population?
- What is the current medical practice in this area? Is the new intervention an improvement?
- What will the new intervention cost?
- How will implementation of the new intervention be operationalized in clinical care?

In assessing the data, said Horberg, KP heavily relies on its own data collection and analysis. KP is "swimming in data," but the challenge lies in

curating the quality of the data and using appropriate statistical methodologies. KP also convenes internal guideline panels and performs systematic reviews, both of which use data from inside and outside the system. In addition to the evidence assessment, KP also performs a financial analysis of how the system would integrate the new costs of care if an intervention were to be adopted.

During the assessment process, many points of tension must be balanced. First, KP is driven by its mission to "do the right thing the first time." However, this desire to provide members with the best possible care is tempered by the fact that the system must remain financially viable to provide care. Second, KP believes in "prevention first . . . if possible." While providing drugs to treat disease is important and necessary, it is better to prevent the disease in the first place—both for the patients' quality of life and for the cost savings to the health system. Third, KP considers whether the new intervention is actually better than the current treatment, and how much of an improvement it represents. Fourth, KP looks not just at the availability of evidence, but the credibility and relevance of the evidence. For example, KP has two very different populations of members with HIV. More KP members on the East Coast with HIV tend to be female, heterosexual, and African American, whereas on the West Coast, members with HIV are mostly white men who have sex with men. Evidence for an HIV intervention would need to be relevant to both of these groups, if it were to be adopted. KP also looks for gaps in the data, said Horberg, particularly if there are gaps in the data on certain vulnerable populations.

The process for deciding which interventions to adopt, said Horberg, is both bottom-up and top-down. The impetus for assessing an intervention may come from new information in the literature, a provider request to review the literature, patient demand, or new regulatory or statutory requirements. There are a variety of interregional groups that are convened to make decisions about adoption, including formulary committees, new technology committees, guideline committees, specialty groups (e.g., gastroenterology chiefs), and special ad hoc groups. However, Horberg noted that the groups may come up with different decisions and these are not always in alignment. For example, the guideline committee may recommend a new drug as a good treatment, and the formulary committee may decide to approve it to the formulary, but the benefits committee may not approve it for payment. In addition, despite the emphasis on evidence-based medicine, individual care decisions are often based on discussions and experiences of providers and patients. Overall, KP aims for "collaboration" within the organization, with the goal of "getting what's best for the patient," concluded Horberg.

Daniel Ford, director of the Institute for Clinical and Translational Research at the Johns Hopkins University School of Medicine, described

the Johns Hopkins Health System (JHHS) as both a generator of evidence and a consumer of evidence. Although one might assume that this dual role would lead to systematic synergies, in reality, there are sometimes tensions and inconsistencies between the two parts of the system. JHHS still relies heavily on traditional clinical research, said Ford, but recently has branched out into conducting clinical research at community hospitals. In the system's three community hospitals, there are about 450 patients in a clinical trial at any one time, with 23 research coordinators supporting them. This project marks a transition to collecting evidence outside of the traditional academic health center, said Ford. In addition to the community hospital research, JHHS also conducts clinical research with its patients; about 10 percent of JHHS patients (300,000 out of 2.5 million) have been involved in a clinical trial over the past 8 years, said Ford.

Ford said the JHHS process for making coverage decisions is similar to KP's. JHHS uses internal or external data summaries, consults experts about their views of the available data, and looks at the patients who have received the drug. Ford said that while JHHS has a fair amount of internal data, it would be a "stretch" to rely solely on these data to judge the clinical effectiveness of a drug. He noted that because physicians use the electronic health record (EHR) daily, and know its drawbacks and limitations, they may be less persuaded by a conclusion based on EHR data versus data from a regulated clinical trial. However, he noted that the EHR data are consistently improving, and that there are new ways to integrate other data sources, such as data from other hospitals or death records. One database, called the Chesapeake Regional Information System for our Patients (CRISP), contains data from nearly all hospitals in Maryland, and hospitals from Delaware; Washington, DC; and West Virginia are joining as well. This integrated database allows researchers to track all hospitalizations, emergency room visits, and deaths, and serves as an important tool for both research and clinical practice.

Ford discussed the roles and expectations of patients and providers in the generation of RWE. Patients often desire information about their treatment plans in order to inform their personal decision making, but this information is not always accessible in the current environment. Academic researchers, meanwhile, frequently express interest in researching drug effects on off-label indications and applying their findings to usage recommendations, but providers still rely on traditional RCT evidence rather than other potential sources of evidence. Ultimately, Ford said, altering the evidence generation system will require changes on the part of multiple stakeholders. The funding for RCTs remains steady, providers use data from RCTs, and patients understand the RCT design. Integrating EHR data into research will require a shift in perspective and an effort to ensure that EHR data are as valid as data from traditional RCTs.

PATIENT PERSPECTIVE

Sharon Terry, president and chief executive officer at Genetic Alliance, started by questioning the term "patient" itself. To Terry, the word "patient" conjures up an image of a person sitting quietly in a gown on an exam table, with a "tremendous information asymmetry and power asymmetry." Terry told workshop participants that they are all patients first and professionals second, and that "we make very different decisions [as patients] than we do when we sit here primarily as professionals."

Terry briefly told the story of how she transitioned from a mother of two with a background in religious studies to a researcher who is involved in clinical trials with four different therapies. Terry's children were born with a rare genetic condition called pseudoxanthoma elasticum (PXE). Upon their diagnosis, Terry and her husband founded PXE International, started a patient registry, discovered the gene responsible for PXE, patented it, and developed a diagnostic test. Terry now serves as the chief executive officer of Genetic Alliance, which is a network of more than 1,000 organizations and patient advocacy groups, representing millions of people with genetic diseases. Terry stressed that people like her—patients, families, and communities—have gathered data for years. Abundant data have been gathered through disease advocacy organizations, community-based participatory research, and activist and citizen science contributions, she said. However, these communities are fighting an "uphill battle" to collect data, and the data are often not used or integrated with data from other sources to impact the medical system. Terry questioned, "When are we going to start to pay attention to that information, and not keep comparing it to other sources of information? We need it all." In other industries, data from consumers are highly valued, such as reviews and ratings on Angie's List. In health care, these data may be more difficult to collect and manage, but they are equally—if not more—useful.

One example of community-led evidence generation is patient registries. These registries, which are often created and managed by community and advocacy groups, capture information about the lived experience of patients. The validity and accuracy of the data collected by these registries has long been questioned—even though, said Terry, there are also issues with validity and accuracy when data are collected in a clinical trial or in the course of clinical practice (e.g., EHRs). There are now registries for a wide variety of communities, from individuals working in homeless communities to people using a specific medical device to parents of autistic children. Technological advances have enabled better communication among these communities, and have facilitated the collection of real-world data (RWD) such as data collected on smartphones and other devices.

The difference between a community-led registry and an industry-led registry, said Terry, is that the community-led registry is focused on the priorities and the lived experiences of the community, rather than on the financial bottom line. Focusing on community priorities enables the registries to answer the questions most relevant to the community, and to give a realistic view of the opportunities and risks of taking a certain path (e.g., using a certain therapeutic), said Terry. Another benefit of community registries is the opportunity for education. Unlike a clinical or trial setting, in which a patient comes in for brief visits, community registries often have opportunities for daily interactions (e.g., through Facebook or chats), and patients can educate and communicate with each other. However, despite the benefits of community-led registries, there is a need for "rigorous and accessible methods for validation" of the data. Terry noted that working toward validation should be a communal effort, and that the methods used to validate should be made accessible to all.

Terry introduced workshop participants to a platform that Genetic Alliance developed called Platform for Engaging Everyone Responsibly (PEER) (see Figure 2-1). This platform allows organizations and communities to create a custom registry, and to offer individuals control over the data they share. Terry explained the process. First, an organization creates a registry and puts a link on its website for patients to register. Individuals who register can choose their own personal privacy settings and the purposes for which their data may be used. The health data, the contact information, and the privacy preferences are held in three separate databases, and the registrant has control over who may access the data. The process of choosing privacy settings is guided by people from the same community as the registrant—for example, the same socioeconomic status, the same race, and/or the same experience with the disease, said Terry. She noted that of the tens of thousands of people who have input their data into the system, about 95 percent say, "Share my data with everybody." This customizable platform that can be embedded into the communities' existing website allows organizations to create registries that are responsive to the needs of the community and that "look like" the community. However, they all share the same underlying data structures and are therefore interoperable across diseases. There are currently 45 communities using PEER to build the database they need to get industry attention or to start clinical trials themselves, said Terry.

Terry closed with a quote about keeping the patient at the center of decisions about health care: "Nothing about us without us." As stakeholders move forward with collecting and using RWE, she said they must keep in mind that patients' lived experiences are a valuable source of information, and that patients are experts on themselves and their experiences. Of course, the data collected must be aggregated in a way that is rigorous and clear, said Terry, but there is an existing plethora of communi-

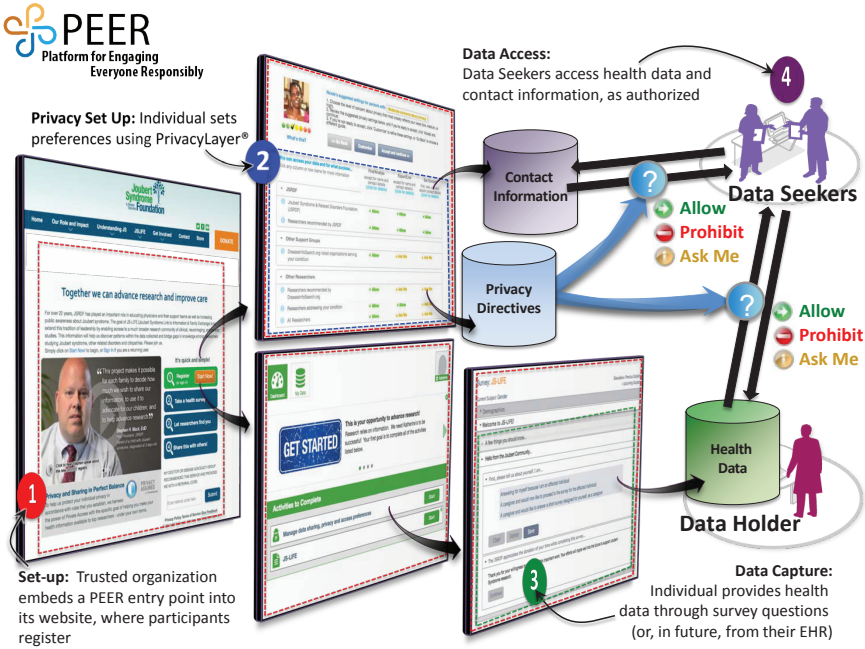


FIGURE 2-1 Platform for Engaging Everyone Responsibly (PEER).
NOTE: EHR = electronic health record.
SOURCE: Terry presentation, September 19, 2017.

ties who are ready, willing, and able to generate evidence that can be used for decision making. Stakeholders could build on the successes of these communities, and help to facilitate the rigorous collection and analysis of patient-generated data.

DISCUSSION

Joanne Waldstreicher, chief medical officer at Johnson & Johnson, and Eleanor Perfetto offered some reflections on the presentations by Sherman, Horberg, Ford, and Terry, as well as their own perspectives on the issue of RWE generation. Their input, as well as discussion from the audience, has been divided by topic area.

Patient Perspective: "Will This Work for Me?"

To follow on Terry's presentation about the patient perspective, Perfetto told participants about a National Health Council roundtable that was held

to gather patient views of real-world evidence. One of the predominant findings, said Perfetto, was that patients do not particularly care about the source of evidence. Rather, Perfetto said that what patients most want to know is, "Will this work for me?" Whether the evidence is generated in the real world or in a clinical trial, patients want evidence that will help them make a good decision about their care. In addition to this finding, many patients at the roundtable exercise conveyed a belief that data should belong to patients, and patients should have an opportunity to understand how and by whom their information is being used and to actively opt in to the use of their information in research. Ross McKinney, chief scientific officer at the Association of American Medical Colleges, added that a regulatory and ethical framework to deal constructively with RWD is lacking. In a prospective study, researchers obtain active consent from patients, whereas in studies involving RWD, patient consent and privacy are much less central, he said.

Roundtable contributors also said that EHRs and claims are "not authentic sources of real patient data" because these sources do not include information about patient preferences or experiences. It was proposed, said Perfetto, that clinical data be integrated with the data that matter to patients, such as patient-reported outcomes. This integration would not only present a fuller picture, but would also help with appropriately interpreting the data that come from clinical sources. Patients also revealed frustration with the difficulty of assessing the quality of different types of studies and data, and proposed that patient advocacy groups have access to a scientific board or consultants.

Perfetto echoed Terry's point that patient communities have long been a source of RWD, and have produced quality data that have made major contributions and led to changes in care. The challenge, said Perfetto, is to integrate all of the data, including community-generated evidence and other types of RWD. Only by looking at the weight of all of the evidence—regardless of where it comes from—can health professionals "help patients make the best decisions and the best choices."

Data and Analysis Considerations

Califf noted that two different dimensions are involved in generating RWE: the source of the data and the method of analysis. These dimensions are often conflated and discussed as if RWE equates to observational studies. However, studies on RWD can include observation, a variety of randomization methods, and prospective or retrospective analyses, said Califf. Bindman added that while RCTs are often held up as the "gold standard," using a method other than RCT does not mean that scientific principles are abandoned. Waldstreicher said that study designs are already

changing, and that real-world trials that use randomization are becoming more popular. For example, randomized pragmatic studies have made a major difference in some areas of medicine, such as in the field of statins. Designs and tools such as platform trials and master protocols are increasingly important, said Waldstreicher, but it is critical to have regulatory buy-in for these types of trials. Before embarking on a trial, she emphasized that industry needs to know that the evidence generated will be acceptable from a regulatory perspective.

When observational study designs are used, the bar for rigor and quality should be as high as it is for clinical trials, said Waldstreicher: Clinical trials have a number of requirements that grant collective confidence in the data, and observational trials could have similar requirements. Observational trials should be reported with transparency about the sponsor of the trial, the study design, the methodology, the protocol, and the analysis plan, and observational trials should be registered just like clinical trials, said Waldstreicher. When an observational trial is based on analysis of one database, it is valuable to test whether the results can be replicated using a different database.

Rory Collins, head of the Nuffield Department of Population Health at the University of Oxford, offered a slightly different perspective on observational studies. Although observational studies on large databases are expedient and "seductive," Collins worried that "we're planning on a very large scale to repeat the errors of the past." The drawbacks of randomized controlled trials (RCTs) have been discussed at length at this workshop and others, said Collins, but observational studies also have serious limitations. Perhaps the solution is not to do more observational studies, but instead to fix the issues with RCTs, he suggested. Waldstreicher clarified that while she believes there is a role for observational studies, the evidence from these studies should be looked at as part of the totality of the evidence, in combination with data from sources such as randomized pragmatic trials, safety data, and predictive modeling. "We should use all of the tools in our tool chest," said Waldstreicher, and use knowledge from all sources in an iterative and synergetic fashion. To this end, stakeholders could work to break down compartmentalization, share data, and collaborate in order to build a learning health care system, she said.

Role of Health Systems

Califf addressed the three presenters whose organizations provide and pay for health care in one way or another: Sherman, Horberg, and Ford. He noted that these organizations are making decisions with imperfect evidence, but they are also in a position to improve the evidence base by collecting and sharing RWD from their patients. Califf asked, "What is your

obligation to fix [the system]?" Sherman agreed that health systems are in a unique position to encourage behavior change: "Because we control the dollar and the policies, we are in a unique position to . . . encourage certain behaviors or certain types of activities." Sherman said that Harvard Pilgrim is working closely with other stakeholders to generate better evidence, to collaborate, and to incorporate RWE into decision making. Horberg concurred that payers have a unique role to play; he said that KP feels "a strong sense of obligation to contribute to medical knowledge." For Horberg, changing the system comes down to improving research practices. Research must be based on a sound scientific question that is the "right" question for the community and the patients, he said.

3

Learning from Success

Key Messages Identified by Individual Workshop Participants

- The Salford Lung Studies—the first studies to evaluate the effectiveness of a prelicense drug in a real-world setting—required wide stakeholder engagement to ensure their success, including practitioners, pharmacists, information technology, regulators, and payers; the real-time patient management allowed for close monitoring of safety concerns. (Gibson, Kane)
- One of the most important and challenging components of the Salford Lung Studies was their data platform, which was built specifically for the studies and has since been developed into a cloud-based platform that can be used to investigate other clinical questions; the platform includes a structured data reporting, collection, management, and validation system. (Gibson, Kane)
- Sentinel, a U.S. Food and Drug Administration medical product monitoring system that uses electronic health data to support postmarketing medical product evaluation, is a distributed system that allows data partners to retain private data prior to data curation; Sentinel allows for datasets to be studied using different methods, but still systematically evaluated under known conditions. (Platt)
- Sentinel has been used on its own and linked to other data sources, and several successful cases may be scalable; examples

include Sentinel-linked registries, electronic health records (EHRs), patient-reported data, and chart review, as well as use in randomized controlled trials. (Platt)

- Real-world data (RWD) and real-world evidence (RWE) are crucial for identifying problems with devices early in their use; FDA generally requires a lower evidence threshold for devices; reliable RWE would be useful for shifting pre/post-market approval time lines for devices as well as the Medical Device Reporting System. (Fleurence)
- Registries are already used widely in the device space, and increased use of RWD and RWE for devices could push registries toward Coordinated Registries Networks[a] that link existing registries to other existing data sources such as claims data and EHR data. (Fleurence)

[a] See https://mdepinet.org/wp-content/uploads/Recommendations-for-a-National-Medical-Device-Evaluation-System_24-Aug-2015.pdf (accessed January 4, 2019).

In this session of the first workshop, participants heard about successfully completed and ongoing initiatives that generated, collected, and/or analyzed real-world data (RWD) and real-world evidence (RWE). Speakers were asked to describe the features that led to success in their particular program, and to consider how these successes could be generalized and scaled for future projects.

SALFORD LUNG STUDIES

The Salford Lung Studies were two late-phase randomized controlled trials (RCTs) conducted in Salford, United Kingdom, and the surrounding areas, said Martin Gibson, chief executive officer of Northwest EHealth. Sponsored by GlaxoSmithKline (GSK), the Salford Lung Studies were the first studies in the world to evaluate the effectiveness of a prelicense medication in a real-world setting, said Gibson. There were two separate studies—one for chronic obstructive pulmonary disease (COPD) and one for asthma—evaluating the same drug (Relvar Ellipta); the studies enrolled more than 7,200 patients. Enrolled patients were monitored in near real time for safety and outcomes, using city-wide linked electronic health records (EHRs). The studies showed that the drug was effective at improving outcomes for both COPD and asthma.[1] Gibson said that tradi-

[1] See https://www.nejm.org/doi/full/10.1056/NEJMoa1608033 (accessed January 4, 2019); https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(17)32397-8/fulltext (accessed January 4, 2019).

tionally, premarket studies look at efficacy, and postmarket studies look at effectiveness. The Salford Lung Studies, however, were late Phase III trials of a premarket medication that were designed to gather information on effectiveness in real-world conditions. Gibson noted that new technologies are allowing researchers to blur the lines among traditional types of studies, and to look at real-world effectiveness earlier in the process.

Gibson explained the reasons why the Salford Lung Studies were conducted in this particular region. First, because the National Health Service (NHS) serves all UK residents, the results generated from NHS data are likely to be generalizable. Second, the general practitioners in NHS have been using EHRs since the mid-1990s, and the EHR system is a good system with high-quality data, said Gibson. Third, Salford itself is served by a single large university hospital, and the primary and secondary care data in this region have been integrated since 2002. The Salford Hospital is "regarded as the most digitally mature organization in the NHS," which greatly facilitated the lung studies. Finally, said Gibson, there was an existing close relationship between the community of Salford and the health system. Northwest EHealth, which was established in 2008 to improve research using EHRs, had already worked in the community and there was a "connected community of care." Gibson stressed that this element was critical to the success of the Salford studies.

Gibson gave a brief overview of the design of the COPD study (see Figure 3-1). He noted that the criteria for inclusion were much more open than they would be for a standard trial. In the trial itself, participants were randomized during an initial visit, and then had an end-of-study visit 12 months later. If assigned to the active group, patients received their drug through the usual community pharmacies, and all patients were monitored "behind the scenes" through the EHR system. Care was taken to ensure that there was minimal intervention, and that patients were truly receiving normal care during the 12 months between the initial and end-of-study visits.

When GSK first suggested these studies, nothing similar had ever been done before, said Gibson. Many questions needed to be answered to proceed:

- What study design would best achieve internal and external validity?
- What data existed?
- How good were the data?
- How could the data be accessed, managed, and validated?
- What were the evidence needs of research authorities, regulators, and payers?
- Was there a large enough population to power such a study?

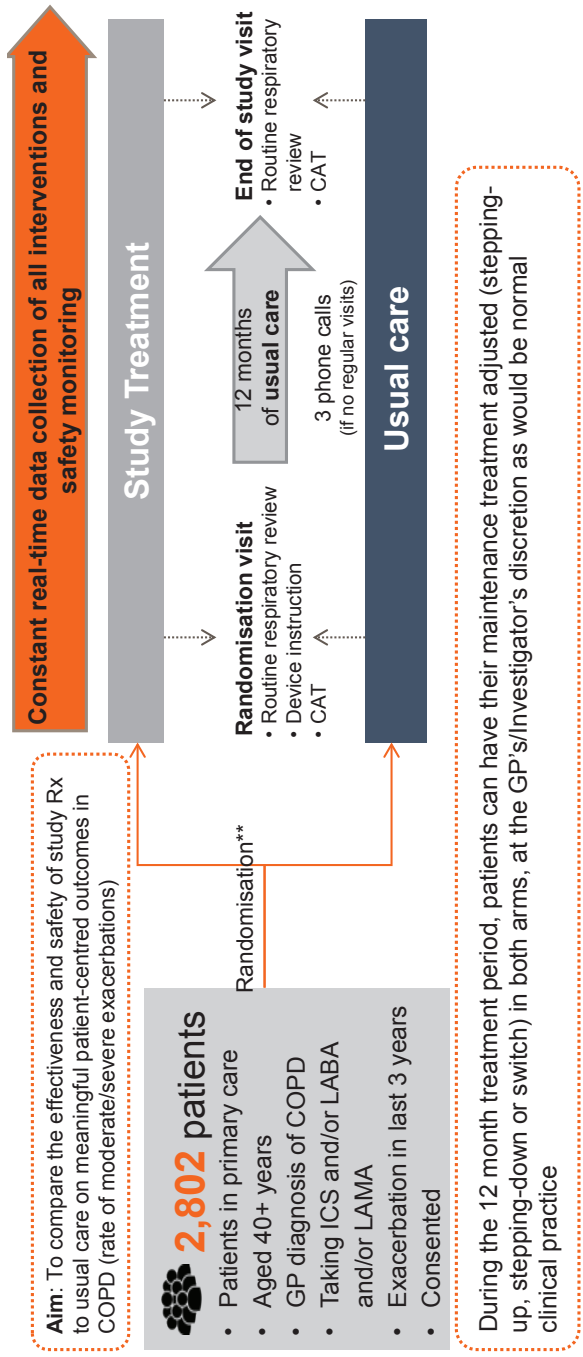


FIGURE 3-1 Salford Lung Studies design.
NOTE: CAT = computerized axial tomography; COPD = chronic obstructive pulmonary disease; GP = general practitioner; ICS = inhaled corticosteroids; LABA = long-acting beta agonist; LAMA = long-acting muscarinic antagonist.
SOURCES: Gibson presentation, September 19, 2017; concept/data from Bakerly et al., 2015.

Answering these questions, however, was just the "tip of the iceberg," said Gibson. To carry out the studies, every element had to be operationalized, which required extensive forethought. Marie Kane, chief operating officer of Northwest EHealth, explained the process. Kane said that although Northwest EHealth is primarily a technology company, the biggest part of carrying out the Sanford Lung Studies was engaging with people. General practitioners, pharmacists, and specialists were critical for supporting the study day to day, while people who worked at every level of the health care organization had to buy in as well. For example, said Kane, the information technology (IT) department was not accustomed to being involved in clinical research, so care had to be taken to enlist their cooperation and support. Regulators and payers were engaged in the process in an attempt to ensure that the evidence that came out of the studies would be acceptable to these stakeholders. More than 3,000 people—physicians, nurses, pharmacists, data managers, and administrators—were trained as part of the research delivery team.

In addition to engaging with people, Northwest EHealth had to build a system to collect, manage, and validate the data, as well as a system for safety reporting. Kane presented a simplified diagram of the platform that was built for the studies (see Figure 3-2). Only structured and coded data were collected, said Kane. Data were collected from a number of sources, including Salford general practitioners, out-of-area patient episodes, 140 retail pharmacies, and data from community services. Kane said these studies required far more investment in data processing and error management than traditional clinical studies, due in part to variability in the data as they were collected as well as the scale and complexity of the data linkages required to determine patient outcomes. One of the challenges, said Kane, was that these sources used different IT systems, collected different data, or had data that were not integrated. However, by going into the "back tables" of the systems, most of the data were able to be retrieved in a usable format.

One benefit of monitoring patients in near real time, said Kane, was that serious adverse events could be detected and reported almost immediately. The safety monitoring system was set up with criteria for alerting researchers about potential safety issues, and a safety team reviewed these alerts on a daily basis. Alerts were investigated using the patient's EHR, and when the cause of the adverse event could not be determined, the sponsor was alerted of the adverse event. The alert system, said Kane, can be "tuned up or tuned down" depending on the specific safety concerns and the needs of the sponsor.

In retrospect, said Kane, they underestimated the scale and complexity of the task of recruiting patients and staff, and of developing data systems that could collect and manage data from multiple unaligned sys-

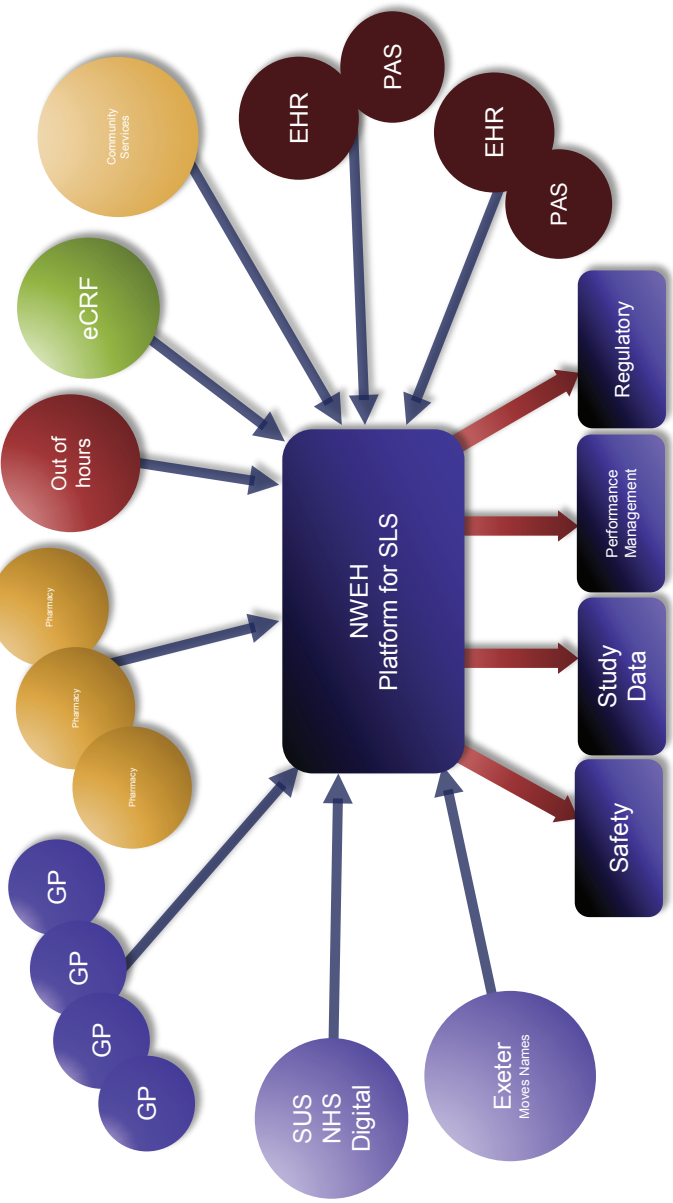


FIGURE 3-2 Salford Lung Studies platform diagram.
NOTE: eCRF = electronic case report form; EHR = electronic health record; GP = general practitioner; NHS = National Health Service (England); NWEH = Northwest EHealth; PAS = patient administration system; SLS = Salford Lung Studies; SUS = Secondary Uses Service.
SOURCE: Gibson and Kane presentation, September 19, 2017.

tems. Some of the challenges were small, but consequential. For example, because the drug was prelicense, the drug name was not in the dropdown list in the EHR system, so providers had to type in the name. Mistakes or typos would have resulted in missing these data, so an algorithm was built to identify and aggregate the disparate terms. Other challenges were large. For example, pharmacy data had never been used for these purposes before, and how to best access and use these data is still an open question. Another data-related challenge, said Kane, was ensuring that they got the *right* data rather than just *more* data. Collecting too much data costs money and can create inefficiencies. "It needs to be the right data with the right provenance, with the right frequency," said Kane. However, said Kane, the most critical component of the study was not building data systems or recruiting patients; it was building relationships among people by recruiting the right partners and ensuring that everyone understood how to fulfill their role in the study.

Based on the lessons learned from this experience, Northwest EHealth has converted the design developed for Salford into a cloud-based platform. The platform has been fully reengineered and revalidated, and has configurable and modular applications, said Kane. Depending on the needs of the researcher, he or she can use different components of the applications. There is also a master data management system, which allows a researcher to develop configurable data schema and data dictionaries. Other tools that Northwest EHealth has developed, said Gibson, include tools for recruiting and engaging patients. The system for recruitment allows researchers to screen the population and identify patients who may be appropriate subjects (using anonymized information); the system then contacts the patients' providers to recruit them into the trial. A patient engagement tool that has been built allows patients to sign up directly to give consent for the use of their EHR and to be recruited for studies. The benefits of performing research using EHR data, said Gibson, are numerous. First, a researcher can collect information on a patient throughout her life, rather than just for a set period of time in a study. Second, using EHR data reduces the influence of research on patient's care—patients and providers can go about their normal course of care, while data are being captured for research. Finally, data collection can be streamlined and designed for the specific protocol, saving money and increasing efficiency. Once the system for one protocol is up and running, it can be used for other things, such as safety monitoring, said Gibson.

During the discussion that followed presentations, participants discussed whether and how the Salford model could function as a "franchise" that could be exported to other health systems for other disease questions. Kane said the "franchise" part of the studies would consist of the methods developed and the lessons learned from each variation—not all studies

would need to be conducted using the same exact approach. John Graham, senior vice president, medical engagement and value evidence and outcomes of GSK, agreed that the studies had generated both a reusable infrastructure and lessons learned that could be applied in other disease areas; GSK is already applying a similar model in cardiovascular and renal disease studies in the United States. Califf added that finding a way to franchise these models—that is, to identify and scale up the common elements of successful programs—is critical for making research more efficient, but that each new project would likely have its own "regional flavor." Gibson said that recruiting health systems to participate became easier with each study because of the positive experiences of the study participants and the investments made in building relationships as well as the desire of health systems to participate in exciting, new, and relevant studies. These experiences help to alleviate fears of systems that do not normally conduct research: "We have the tools and the capability to give them the confidence that this is something they can take part in."

SENTINEL

The Sentinel Initiative is a national medical product monitoring system that was launched in 2008 by the U.S. Food and Drug Administration (FDA) in response to legislation that required FDA to use electronic health data to support postmarketing medical product evaluation, said Richard Platt, professor and chair, Harvard Medical School, and executive director, Harvard Pilgrim Health Care Institute. While the mandate asked FDA to assess the use, safety, and effectiveness of regulatory medical products, FDA has focused Sentinel primarily on safety. However, FDA has always intended that Sentinel would be used to support a variety of activities, including clinical research, randomized trials, and public health surveillance.

Sentinel is the product of a collaboration between FDA and a large number of organizations that bring both data and scientific expertise. In essence, Sentinel is a curated, distributed dataset that adheres to a common data model, said Platt. Sentinel is a fairly simple system, with a set of linked flat file records. The data that can be accommodated by Sentinel are wide ranging, including administrative data (e.g., age, sex, zip code, enrollment, medical and pharmacy benefits, encounter diagnoses and procedures, ambulatory pharmacy dispensing), laboratory test results, vital signs, death data, immunization records, and in-patient data (see Figure 3-3). Platt said "hundreds of millions of person years and billions of encounters" are represented in the Sentinel data, and about 10 percent of the people have some sort of laboratory test results.

Sentinel is a distributed system, which means that each of the data partners retains its own data, and the data are interrogated by the exchange of

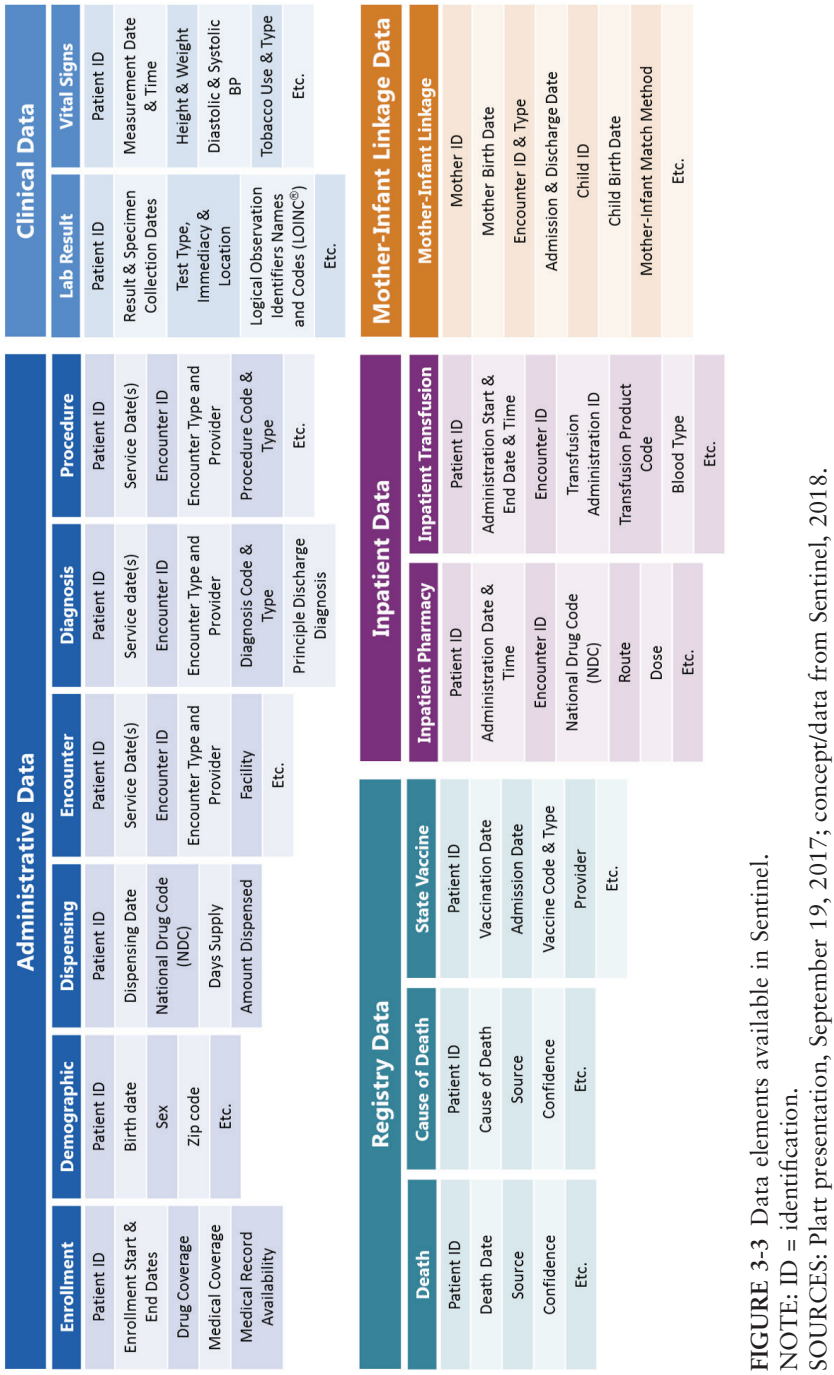


FIGURE 3-3 Data elements available in Sentinel.
NOTE: ID = identification.
SOURCES: Platt presentation, September 19, 2017; concept/data from Sentinel, 2018.

executable programs. Each data partner can opt in or out of any individual query. Platt noted that while Sentinel is a public enterprise and a public resource, the participation of private organizations is critical to its success: Sentinel "is nothing if the data partners aren't there." One of the things that makes Sentinel data particularly useful, said Platt, is the fact that the data are extensively curated before use, which is not necessarily true of other large medical datasets. This curation, while it takes time and energy, means that the users of the data do not need to spend time assessing the quality of the data.

One particular advantage of Sentinel, said Platt, is that it allows for different studies to be conducted on the same research question and the same data, using different methods. Simon noted that one of the issues with observational studies is that when researchers choose different methods and get different results, it can be difficult to know which method is correct. With Sentinel, different methods can be systematically evaluated, under known conditions, said Platt.

Platt gave workshop participants six examples of how Sentinel has been used, either by itself or linked to other data sources; he noted that these examples were chosen because they have the potential to be scalable.

Sentinel Alone: Prospective Surveillance Pilot of Rivaroxaban Safety (Bai et al., 2017; Chrischilles et al., 2018; Patel et al., 2011)

This project tested the ability to use Sentinel to do prospective surveillance, said Platt. The researchers compared data from patients on warfarin and rivaroxaban, while controlling for more than 70 confounders, including age, sex, comorbidities, usage, and diagnosis. Outcomes that were examined included gastrointestinal bleed, intracranial hemorrhage, and ischemic stroke. Platt noted that this observational study correlated well with randomized trials, which "should give us some confidence that we can be attentive to how the product is working in actual practice." Furthermore, using Sentinel data allows researchers to explore populations that are not well represented in the clinical trial. Platt noted that "others may not agree" with using observational data in this way, but that for situations or populations where a clinical trial will not be done, there may be ways "to decide whether the observational data are credible enough."

Sentinel with Chart Review: Risk of Intussusception After Rotavirus Vaccination (Yih et al., 2014)

The first vaccine to prevent rotavirus infection in infants was licensed in 1998, but withdrawn in 1999 due to risk of intussusception, a form of bowel obstruction, said Platt. Alternative rotavirus vaccines showed no

increased risk in clinical trials, but postlicensure studies in other countries suggested an increased risk. In 2010, FDA began a study to quantify the possible risk among U.S. infants. Researchers used data from three Sentinel partners, and gathered data using the Current Procedural Terminology (CPT) codes for immunization and the *International Classification of Diseases, Ninth Revision* (ICD-9) codes for intussusception and related issues. Because CPT and ICD-9 codes are "not sufficiently specific enough to do high-quality epidemiology," said Platt, the researchers actually reviewed the full-text medical records of patients after redacting direct identifiers. Using data on 500,000 patients, the algorithm identified potential cases, and researchers obtained full medical records for 80 percent of these. Pediatricians adjudicated the records, and they found a risk of intussusception of about 1.5 per 100,000 children immunized. The clinical trials that had been performed for these vaccines, said Platt, had enrolled only 60,000 children, which may not have been sufficient to find the actual risk. This study demonstrates the value of being able to link administrative data to full-text records, said Platt.

Sentinel Linked to Registries: Linking Mother–Infant Pairs[2]

Platt gave a brief overview of the difficulties of linking maternal and infant data; he noted that this gap has "bedeviled health services research" for decades. Sentinel data partners have data about linked mother–infant pairs, unlinked mothers, and unlinked infants, and state departments of health have birth certificate data. Researchers attempted to link moms and infants using these data, and were able to link more than 80 percent of the records from four data partners. Platt observed that while researchers thought linking mothers and infants would require the information from the registries (i.e., the birth certificates), nearly all of the pairs were able to be linked through information already in Sentinel (i.e., subscriber ID and last names and addresses). However, he noted, the birth certificates did include a lot of rich information that is useful for researchers, such as gestational age and smoking history.

Sentinel Linked to EHRs: PCORnet ADAPTABLE Trial[3]

Platt used the National Patient-Centered Clinical Research Network (PCORnet) ADAPTABLE trial as an example of how a Sentinel-like system has been used to link with data from EHRs. The ADAPTABLE trial, said

[2] See https://www.sentinelinitiative.org/communications/publications/2017-icpe-presentation-developing-mother-infant-cohort-sentinels-prism (accessed January 4, 2019).

[3] See http://theaspirinstudy.org (accessed November 6, 2018).

Platt, is simple, with the goal of randomizing 20,000 people with coronary artery disease to either low-dose or high-dose aspirin for prevention. Follow-up of patients will occur largely through EHRs and payer data. This trial is an example of using available data to enable pragmatic research, and demonstrates how Sentinel could similarly be used.

Sentinel Linked with Patient-Reported Data: A Mobile App for Studies of Medication Safety[4]

Sentinel researchers built a mobile app to enable individuals who have data in the Sentinel system to report information that can easily be merged with the system, said Platt. This app, now known as the MyStudies smartphone app,[5] is currently being field tested. Pregnant women with the app are asked a variety of questions, including smoking history, level of nausea, over-the-counter drug use, and other areas of interest. These patient-reported data can easily be merged with the individual's information in Sentinel, which should make for an even richer dataset.

Sentinel for Randomized Trials: Sentinel IMPACT-AFib (Cocoros et al., 2018)

The IMPACT-AFib trial is a pragmatic clinical trial that uses Sentinel data to test methods to improve the use of oral anticoagulants (OACs) in patients with atrial fibrillation. Working with five data partners, direct mailers were sent to 40,000 health plan members with AFib, at a high risk for stroke, and not currently taking an OAC. Mailers also went to the patients' providers to encourage the consideration of prescribing an OAC. Outcomes will be assessed at 12 and 24 months, using Sentinel data. The primary outcome is initiation of OAC, with secondary outcomes including rates of stroke hospitalization and bleeding events. Eligibility for the trial was also determined using Sentinel data; an algorithm identified patients who were at risk and not currently being treated with an OAC. This trial, said Platt, demonstrates that it is possible to use Sentinel to identify individuals who are eligible for intervention, and also to support the implementation and follow-up of the trial.

[4] See https://www.fda.gov/downloads/Drugs/ScienceResearch/UCM625206.pdf (accessed November 6, 2018).

[5] See https://www.fda.gov/Drugs/ScienceResearch/ucm624785.htm (accessed January 4, 2019); https://www.fda.gov/Drugs/ScienceResearch/ucm624785.htm (accessed January 4, 2019).

DEVICE REGISTRIES

Rachael Fleurence, National Evaluation System for health Technology (NEST), told workshop participants about applying lessons learned from device registries to other treatment types. Fleurence noted that devices are sometimes "forgotten" in a conversation that focuses primarily on drugs, but that RWD and RWE are essential for early identification of problems with devices. Fleurence gave several examples of devices for which RWE could be useful: metal-on-metal hip implants, the contraceptive device Essure, and power morcellators for uterine fibroids.

Fleurence noted that there are major differences in the ways that devices and drugs are regulated by FDA, and that these differences impact the generation and use of RWE in each arena. For drugs, said Fleurence, approval is based on substantial evidence from well-controlled investigations. Approval requires disclosure of adverse events, and manufacturers must conduct postapproval safety studies. The National Drug Code ensures that the drugs that patients take can be accurately identified in claims and EHR data. On the device side, there are two pathways for approval: premarket approval (PMA) and 5-10K. For PMA, generally a single small study is required, said Fleurence. Postapproval requirements are variable, and only apply to PMAs. While a system to accurately track the use of devices (the Universal Device Identifier) was put in place in 2015, it is not compulsory for payers or health systems. As a consequence, it is difficult, but not impossible, to identify brand-specific devices in EHR and claims data, said Fleurence.

Fleurence sees a particular benefit of RWE and RWD for devices in two ways. First, RWE can help support a "pre/postmarket shift," in which devices can receive approval more quickly, when there is confidence that there will be robust postapproval data collected on safety and effectiveness. Second, RWE could help to improve the Medical Device Reporting system, with implementation of automated surveillance methods. These methods could be used to quickly identify problems with a device so that action could be taken (e.g., pull device from the market and/or conduct further safety research). These two benefits of RWE would help patients gain access to innovative, safe, and effective devices more quickly, said Fleurence. To explore and capitalize on these potential benefits of RWE, the NEST Coordinating Center (NESTcc) was established in 2017 as a catalyst to "support timely and reliable development of high-quality real-world evidence." To do so, NESTcc will establish partnerships with a range of stakeholders that provide data and analytics solutions, will set data quality and methods standards, and will offer products and services of value to stakeholders, said Fleurence.

Registries have historically played an important role in the regulatory space for medical devices, said Fleurence. There are some very high-quality

registries in the device space, including the Vascular Quality Initiative, the International Consortium of Orthopedic Registries, and the Transcatheter Valve Therapy Registry. In addition to these individual registries, there is a movement toward a Coordinated Registries Network, which links existing registries with other sources of data such as claims and EHR data. These registries, she said, "provide high-quality and fit-for-purpose data, and support both observational and randomized interventions, possibly at a lower cost." In addition, algorithms can be used to assess registry data for automated safety surveillance. Registries are currently the main source of RWE decisions by FDA's Center for Devices and Radiological Health, said Fleurence.

However, despite the utility of registries, there are challenges. First, developing and maintaining a registry is expensive, and inputting data into registries is time consuming. Second, registries cannot be developed for all devices, disease areas, and patient populations. Third, there is limited use of the Unique Device Identifier in claims and EHR data. Finally, a device outcome can depend not just on the device itself, but on the skill of the provider who is implanting the device. Differentiating among issues linked to the device itself and issues due to the experience of the person performing the intervention must be accounted for in studies of registry data, she said. In addition to these device-specific challenges, there are a number of ecosystem-wide challenges—that is, challenges with using any type of RWD and RWE. These challenges include the difficulty of ensuring the quality of data, missing data, data linkage issues, data privacy and security concerns, and issues with appropriate analytic methods. Fleurence noted in particular that there are challenges with administrative issues—for example, the length of time it takes for research studies to obtain legal review and Institutional Review Board approval. She hypothesized that the single device registries model is likely to evolve soon. Rather, there will be a new model and an expanded definition of registries; for example, a "registry" might actually be a tool for linking preexisting data such as EHRs and claims data, rather than a repository for new input of data.

In conclusion, Fleurence said tremendous progress has been made in the past decade. Technology and adoption of new tools has accelerated, and stakeholders are collaborating to change culture and to increase the involvement of patients. However, the current barriers are still real, and time, resources, and leadership will be needed to overcome them. "There is no question that the future lies in the use of RWD and generating robust RWE," said Fleurence, but "how far off that future lies is the question."

4

Barriers and Disincentives to the Use of Real-World Evidence and Real-World Data

Key Messages Identified by Individual Workshop Participants

- Real-world data (RWD) and real-world evidence (RWE) could prove useful in product development, but there are several barriers to successful implementation; barriers may include a lack of knowledge about RWE and non-interventional research, systems that are not built to use RWE, and mistrust and misunderstanding of RWD. (Bradbury, Levy)
- Randomized controlled trials (RCTs) are the traditional design for evidence-generating clinical research, while observational database studies are seen as riskier due in part to less precise data; hybrid approaches such as pragmatic trials and cluster randomized designs require some real-world considerations, but combine some advantages of both RCTs and RWE studies. Clinical researchers require support and training to choose appropriate research methodologies. (Ford)
- Defragmenting data sources from different stakeholders to create integrated "deep data" provides a more complete picture of a medical product; defragmentation relies on data sharing while remaining cognizant of patient privacy, data security, and the protection of business interests. (Wilson)
- Evidence hierarchies that exist in medical product research could be revisited with the emergence of RWE as a way to assess products outside of classic RCTs. (Cao)

- Integrating RWD into clinical research can lead to stronger and more sustainable research designs that align with the needs of multiple stakeholders; potential approaches for RWD and RWE integration include extension, augmentation, enrichment, and pragmatic design. (Doyle)
- A lack of urgency to adopt RWE is problematic, and including non-traditional stakeholders in research—patients or researchers from different fields of study—could break down the barrier of RWD and RWE use in clinical research. (McCollister-Slipp)

In this session of the first workshop, presenters and workshop participants discussed the structures in various institutions that incentivize the maintenance of the current data generation process, identified disincentives and barriers to the incorporation of real-world evidence (RWE), and considered ways in which incentives could be better aligned. Presenters included researchers, product developers, and data aggregator and analytics companies.

THE USE OF REAL-WORLD EVIDENCE AND REAL-WORLD DATA IN PRODUCT DEVELOPMENT

Elliott Levy, senior vice president of global development at Amgen Inc., began by challenging the notion that there are perverse incentives that encourage companies to develop drugs in ways that are unnecessarily time consuming and costly. Levy said there is not "resistance from above" to the use of real-world data (RWD) and RWE; to the contrary, the pharmaceutical industry is acutely aware of the need to transform the drug development system in ways that acknowledge the revolution of big data and the potential for RWE and RWD. Drug developers, said Levy, are already making extensive use of RWD and big data for internal decision making along the entire process of drug development and marketing. Although traditional clinical trials still play a big role, the evidence from these trials and from RWE are "complementary sources of insight." RWD can help to improve the pragmatism and relevance of clinical trials, for example, by adding patient-reported outcomes, by making the intervention more similar to real-life clinical interventions, or by removing unnecessary exclusion criteria. RWE and clinical trials are currently playing complementary roles, he said, but the next step would be for RWE to actually replace clinical trial evidence. This would be a dramatic change to the process of product development, and would address the fundamental challenge of drug development today—the extraordinary cost.

There are some very significant barriers to replacing randomized controlled trials (RCTs) with RWE, said Levy, as described in the following paragraphs.

Lack of Knowledge and Awareness

Levy noted that many of the people working in product development are those "who are steeped in the randomized clinical trial model, who have benefited from the prestige and scientific cache of the randomized clinical trial, and who have deep faith in it and, to some degree, distrust for non-interventional methods." The physician scientists who lead product teams are accustomed to and comfortable with working with traditional forms of evidence generation, and simply are not aware of the potential uses and benefits of RWE.

Talents and Capabilities

Few people working in product development have experience conducting observational or non-interventional research, said Levy. Those who do, he said, are usually situated in groups (e.g., safety surveillance) that do not contribute to drug development.

Systems and Processes

Drug development organizations have been optimized to generate evidence from randomized interventional trials, and there are simply no systems or processes in place to efficiently and effectively implement alternative trial designs.

These barriers present a significant challenge for replacing RCTs with RWE, said Levy. However, the "good news" is that all of these barriers can be addressed. While the topic area—RWE and RWD—is new, the barriers are not. "These are classic challenges that organizations face as they try to develop new ways of working or launching new products," he said. Addressing these barriers will require senior leaders to demonstrate commitment to the adoption of RWE, investment in training of team members, and a willingness to examine and change organizational structures.

Brian Bradbury, executive director in the Center for Observational Research at Amgen Inc., offered his perspective as an epidemiologist within a product development organization. Bradbury said that although he and his team have conducted RWE studies on the effectiveness and/or safety of a therapeutic intervention, motivating colleagues within his organization who are less familiar with non-interventional studies to embrace RWE-based approaches is a challenge. Bradbury identified a number of existing

obstacles to the adoption of RWE-based approaches in drug development organizations:

- Lack of understanding of non-interventional research methodology;
- Inability to distinguish between higher and lower quality RWE;
- Mistrust of data that were captured outside of traditional RCTs; and
- No clear regulatory pathway for RWE-based approaches.

These barriers, said Bradbury, discourage product teams from pursuing RWE-based approaches. In particular, he said, the lack of a clear regulatory pathway means there is less willingness to mobilize resources and develop capabilities for an approach that may not be acceptable to regulators.

MAKING CHOICES ABOUT RESEARCH DESIGN

Clinical researchers have choices about how to design their research, said Daniel Ford. Different research designs have advantages and disadvantages, and these must be weighed against each other when determining what path to take. Johns Hopkins has a resource called the "Research Studio," said Ford, which provides researchers with expert advice on how to choose the appropriate methodology to answer a research question. Ford explained some of the considerations that might come into play when a researcher is deciding between a database observational study and a traditional RCT.

Traditional RCTs, said Ford, are what people have "grown up" with and make them feel more comfortable. Although RCTs require a commitment to recruitment and data collection, and are highly regulated, the analysis is fairly straightforward and the conclusions can have a high impact. If a researcher wishes to publish in a high-impact journal, a standard RCT is the "short ticket" to get there, Ford said. Database observational studies, on the other hand, are seen as riskier by many researchers, he said. It is more difficult to investigate biological pathways and causation with database studies, and analysis of the research depends on highly complex statistics. Ford said the complexity of analysis often means that researchers will need to find and rely on a colleague who is an expert in the area. In addition, the data in real-world sources tend to be less precise than data from clinical trials, and sometimes do not meet the needs of researchers. For example, electronic health records (EHRs) often have incomplete or missing data points, or the information is not readily accessible (e.g., test results are in PDF format). However, database observational studies are considerably less regulated and take a relatively short time to complete.

These challenges of observational database studies, said Ford, can dissuade researchers from choosing this pathway. Ford noted that younger

investigators can sometimes be hesitant to "rock the boat" by choosing the riskier path before establishing themselves with more traditional research methods. One particular challenge to using RWD for research is that phenomena are limited to what are available (e.g., the data in the EHR); this can be difficult for some researchers to accept, said Ford. To breach the divide between RCTs and RWE, said Ford, researchers need to consider hybrid research approaches such as pragmatic trials or cluster randomized designs. These approaches, too, have challenges—they require engagement of the health system and health providers, incorporation of disparate real-world data sources, and analysis that is usually more complicated than an RCT. However, these approaches can offer the rigor of an RCT with the cost and time savings of RWE. Clinical researchers need support, training, and tools to choose appropriate methodologies and to feel comfortable working with RWE and hybrid methods, concluded Ford.

ACCELERATING EVIDENCE GENERATION THROUGH DEFRAGMENTATION

One of the biggest barriers to the adoption of RWE and RWD, said Marcus Wilson, president of HealthCore, Inc., is the fact that there is a "natural gravitational pull" back to the old, traditional ways of doing things—in this case, RCTs. Unfortunately, RCTs have limitations, said Wilson. RCTs cannot generate clarity about how the product will perform in the real world. Differences between the patients studied in the trials and the patients using the product in the real world may include age, race, ethnicity, gender, comorbidities, concomitant drugs, lifestyle variances, and varying levels of compliance. RCTs also leave us with a lack of precision, said Wilson, which is a hindrance to effective clinical support. Clinical trials can indicate if a product works, and if a product is safe, but they cannot explain precisely for whom the product is effective or safe, he said. Individual patient decisions are still based on inferences and intuition, rather than on precise evidence. Data on how a product performs in the real world—and on how a patient's phenotype matters for efficacy and safety—are not collected until the product is already on the market, which means that many decisions are being made without this evidence.

Collecting these types of RWE that can support decision making is challenging, said Wilson, because patient data are highly fragmented. Most RWD—no matter the sources—are fairly limited in value on their own. Once the fragmented data are integrated together, however, one can "begin to see things you couldn't see before" (see Figure 4-1). Wilson said that while there is a great deal of attention on "big data," the availability of "deep data" is more relevant. Unfortunately, stakeholders are often reluctant to share or integrate their data due to a long history of a fragmented

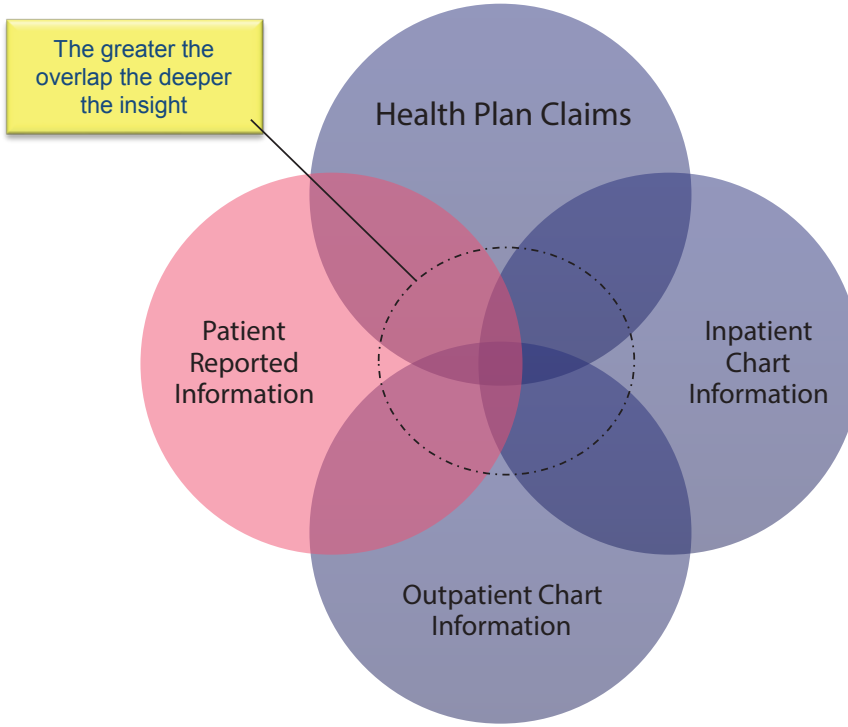


FIGURE 4-1 The importance of integrated data.
SOURCE: Wilson presentation, September 19, 2017.

system, competing priorities and agendas of institutions, and a lack of trust among stakeholders.

Wilson gave an example of his experience attempting to integrate data from different stakeholders. The California Integrated Data Exchange (Cal INDEX),[1] which launched in 2014, was a nonprofit organization seeking to develop a next-generation health information exchange. The idea, said Wilson, was to collect data from multiple health providers and health insurers across California, and these data could be accessed by providers in order to administer the highest quality care possible. Wilson said that in initial talks with California providers, 32 of the largest health care systems expressed a strong interest in participating. However, in the following 3 years, only one provider actually signed up. The biggest challenge, he said, was trust. Data sharing among stakeholders—even if it is beneficial

[1] Cal INDEX eventually merged with another health information exchange organization and the project is now known as Manifest MedEx.

to all parties—requires trust and collaboration, and this remains a huge barrier, he said. Wilson closed by identifying some of the core principles of defragmentation:

- Patient privacy and data security: This is essential to defragmentation because patients will only share data if they are confident that their information will remain private and secure.
- Avoid data misadventuring: RWD are often generated for a specific purpose, and can only be understood in context. Because of this, it is risky to simply pull RWD into research without understanding where they came from and how they can be used.
- Protect business interests of data sources: Much of the distrust among business stakeholders stems from a fear that the data they share will be used to their disadvantage. These stakeholders may be incentivized to share data if they are convinced of the benefits of an active learning health system. For these stakeholders, "it becomes a wonderful trade-off" to share their data in exchange for a richer and more useful system of data.
- Accelerate progress from RWD to RWE: RWD are not generally created with research in mind, so more work needs to be done to understand how and for what purpose RWD should be used, and stakeholders should work collaboratively to improve the data sources.

EVIDENCE HIERARCHIES

A major barrier to the adoption of RWE, said Hui Cao, executive director, Center of Excellence for RWE, Global Medical Affairs, Novartis, is a "fixed mindset" introduced in medical school that evidence from RCTs is the best possible evidence. There are "evidence hierarchies," said Cao, that place RCTs at the top, and put RWE at level 2 or even lower. Researchers and providers are comfortable with RCTs, and they understand the benefits and limitations. However, Cao suggested that in order to truly make change, these hierarchies need to be revisited. Many years ago, the tools and methods for generating RWE did not exist, said Cao, but now there is a better understanding of what can be derived from the real world. The hierarchies of evidence should be adjusted accordingly, she said. Wilson said some of the reticence of researchers to embrace RWE stems from a lack of trust in the data source and the methods used to assess the data. Different types of RWE—and different sources of RWD—are variable in quality, and should be judged accordingly. Not every data source, he said, is fit for every purpose: there is a need for more rigor in vetting of RWD sources, as well as the methods used to assess RWD. Cao responded that while she agreed

that RWD should be scrutinized carefully, evidence from RCTs needs to be scrutinized as well. Some other workshop participants discussed additional barriers to using RWD and RWE in research (see Box 4-1).

OPPORTUNITIES TO INTEGRATE REAL-WORLD DATA AND REAL-WORLD EVIDENCE IN RESEARCH

There is an ongoing transformation in the health research world, said John Doyle, senior vice president and general manager, Real-World and Analytics Solutions at IQVIA (formerly QuintilesIMS), with trends toward growing costs, shrinking reimbursements, and more personalized medicine. This transformation is leading to an increased demand and need for real-world data and evidence, he said. For example, patients and payers are demanding proof of value of new interventions compared with standard of care, and the move to precision medicine requires generating evidence for smaller and more diverse subgroups. The traditional approach to health research, said Doyle, involved a systematic, methodical process of solving a problem for a single, isolated stakeholder. The new approach, by contrast, seeks to design studies in ways that can align the needs and requirements of multiple stakeholders and solve problems in a more integrated, evidence-based way with the use of RWD and RWE (see Figure 4-2).

BOX 4-1
Barriers to Adoption of Real-World Evidence and Real-World Data in Research as Discussed by Individual Workshop Participants

- Gravitational pull-back to the traditional ways. (McCollister-Slipp and Wilson)
- Lack of knowledge about real-world evidence (RWE). (Levy)
- Lack of understanding of non-interventional research methods. (Bradbury and Levy)
- Systems and processes are built for traditional evidence generation. (Levy and Wilson)
- Inability to distinguish between high- and low-quality RWE. (Bradbury)
- Researchers' mistrust of data that were not collected in a randomized controlled trial (RCT). (Bradbury and Wilson)
- No clear regulatory path for RWE-based approaches. (Bradbury)
- RWE-based research is dependent on complex statistical analysis. (Ford)
- Data from real-world sources are less precise or less accessible. (Ford)
- Fragmented patient data. (Wilson)
- Lack of trust among stakeholders prevents data sharing. (Wilson)
- Evidence hierarchies favor evidence from RCTs. (Cao)

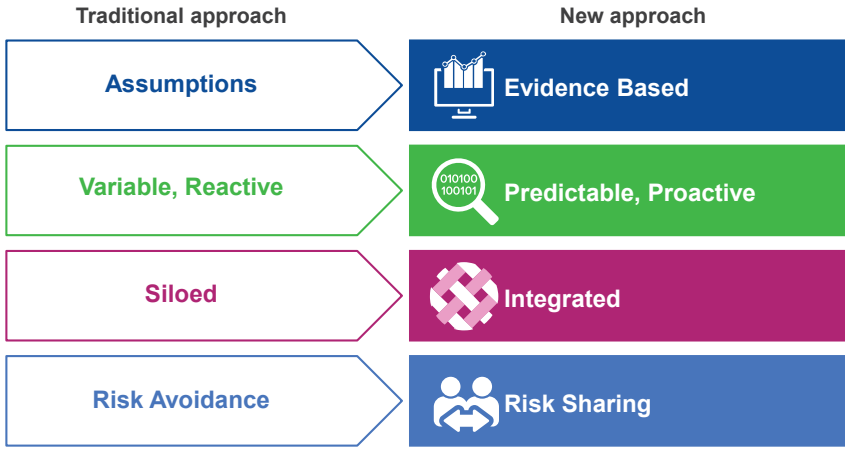


FIGURE 4-2 Transformation from traditional to new approach of evidence generation.
SOURCE: Doyle presentation, September 19, 2017.

All stages of clinical research have challenges and barriers, said Doyle, but the use of RWD can begin to address some of these barriers. For example, in the study design and planning phase, RWD can be used to validate protocol feasibility. Alternative and more efficient pathways to patient recruitment and enrollment can be achieved through the use of RWD. During the data collection and analysis phase, automated tools can be used for real-time tracking, analysis, and reporting. Researchers have already realized some of the advantages of the use of RWD—anecdotally, Doyle reported that recruitment times are being reduced and recruitment rates are increasing, start-up time lines are compressed, and the cost of evidence generation has been reduced.

Using RWD and RWE in clinical research, said Doyle, addresses one simple but often overlooked fact: "Real-world patients are fundamentally different than clinical trial patients." No matter how well a clinical trial is conducted, questions remain about how the findings will extend to more diverse patient populations, how the lack of perfect adherence will affect outcomes, or what the longer term outcomes will be. "We need to bridge that gap with real-world evidence," said Doyle.

Doyle discussed several "mosaic methodologies" in which traditional RCT components are blended with newer RWE components (see Figure 4-3). The first methodology is called "extension." This approach starts with a traditional RCT and patients consent to link their data from other sources, such as EHRs or Fitbits. This allows a researcher to conduct an initial RCT

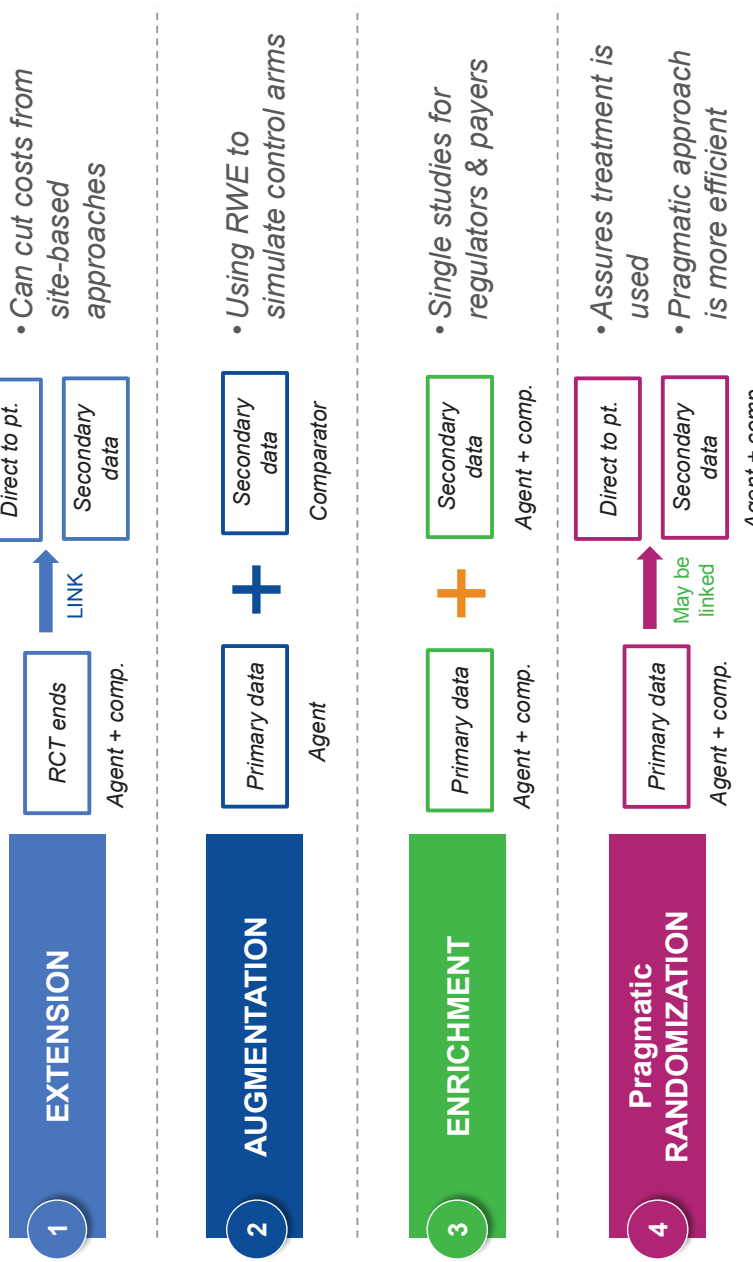


FIGURE 4-3 Mosaic methodologies to blend randomized controlled trial (RCT) and real-world evidence (RWE) approaches.
SOURCE: Doyle presentation, September 19, 2017.

and then collect follow-up data that can be linked. The second approach is "augmentation," in which RWE is used as control data for a single-arm trial. "Enrichment," the third approach, is similar, but combines primary data collected by patients and practitioners with secondary data from EHRs and other sources (e.g., registries). Finally, pragmatic design is a method in which study participants are randomized, initial data are collected, and follow-up is conducted through the collection of RWD. This method is "the best of both worlds," said Doyle, because it uses randomization to address the threats to internal validity, while also providing RWD for generalizability.

Doyle closed with a story about how integrating RWD into clinical research can lead to stronger, more sustainable research design. QuintilesIMS was building an evidence platform for lung cancer, and had designed it with input from multiple experts, including oncologists, epidemiologists, and patient advocates. After developing a dozen protocols, the designers decided to "hit the pause button" and test the design with patients and physicians in the real world. The designers found that although there was a lot of agreement with what they had already developed, there were a few places where they had prioritized outcomes that did not matter to patients while completely missing other outcomes that did matter. This was an "eye-opening" experience, said Doyle.

DISCUSSION

One of the biggest barriers to adopting RWE and RWD, said Anna McCollister-Slipp, chief advocate for participatory research at Scripps Translational Institute, founder of VitalCrowd, and co-founder of Galileo Analytics, is a "lack of a sense of urgency." Researchers, funders, reviewers, institutions, and other stakeholders are "stuck in a paradigm" that they are accustomed to, and they are unable or unwilling to break away from traditions and explore new and alternative sources of evidence. McCollister-Slipp said there are real consequences to overreliance on RCTs, and it is past time to change the traditional paradigm. One way to break down this barrier, she said, would be to invite other perspectives into the decision-making process from people who are not usually involved, for example, patients and people from other fields of study. Currently, the involvement of patients and communities in medical research is limited mostly to participating as subjects and consulting on Institutional Review Board requirements and ethical guidelines. Instead, patients and communities should be involved from the beginning stages of research, she said, and help to guide the design process. McCollister-Slipp urged workshop participants to think big: "We have got to stop tinkering at the edges of the way we do things. . . . We need to think very holistically about how we can . . . truly disrupt this process and create change."

Levy noted that randomization is still vitally important for determining if a product works, particularly in disease areas with small effect sizes. Rory Collins agreed and said the method of randomization in general is often conflated with the way in which RCTs are actually conducted. Rather than changing the methodology—that is, not using randomization—"we should use the methodology differently," said Collins. Michael Horberg added that not every question can be answered with an RCT, just as not every question can be answered with RWE. The research question, the clinical context, and the decision to be made all influence how a trial is designed, and what the sources of data are. RWD are not simple and cheap alternatives to RCTs, said Bradbury; the data are complex and need to be appropriate for the research question.

A primary goal for the research community, said John Graham, is to understand and accept that there is a role for RCTs, prospective observational studies, retrospective observational studies, and many other methodologies that span the spectrum. Together, the wealth of different data sources and methodologies can come together to give the full answer to the research question. Researchers need to start, he said, with the patients' needs, and develop the appropriate methodologies and data sources to find the answers.

5

Getting Unstuck: Mythbusting the Current System

Key Messages Identified by Individual Workshop Participants

- The clinical evidence-generating system should move past precision to reliability; the system should build reusable elements embedded in practice (learning system), use quality by design, use automation for repetitive tasks, and operate from basic principles of scientific research. (Califf)
- Real-world evidence (RWE) could address multiple challenges along the drug development pathway; possible uses for RWE throughout the development cycle include the development of clinical pathways, the optimization of trial design and price, the study of comparative effectiveness, and the study of compliance and adherence patterns. (Graham)
- Rather than focus on replacing randomized controlled trials (RCTs) with observational studies, clinical researchers should focus on improving RCTs. Using randomization, RCTs are good at exposing moderate effects on treatments, but they are also costly; innovation could focus on new principles for randomized trial designs that do not put unnecessary emphasis on data verification. (Collins)
- One approach to the use of observational data networks could be the implementation of "all by all analyses," which reflect the full amount of data on a particular set of medical products. (Ryan)

- The U.S. Food and Drug Administration acknowledges that the current evidence generation system needs to be fixed, and while RWE traditionally has been used to monitor product safety, there are opportunities to use it to test product effectiveness, too. Thoughtful research protocol is important for any use of RWE, and in particular, master protocol could be a promising platform for its use in drug approval. (Woodcock)

While the majority of the first workshop focused on exploring the future potential of real-world data (RWD) and real-world evidence (RWE), and identifying the challenges that need to be addressed, this final session instead focused on dissecting the current system of evidence generation. Evidence for decision making, particularly regulatory decision making, is traditionally generated through randomized controlled trials (RCTs). The first part of this chapter explores the drawbacks and misconceptions about RCTs and other current methods, while the second part of the chapter discusses ways in which the system could be improved, including through the incorporation of RWD and RWE.

FROM PRECISION TO RELIABILITY

The traditional system of evidence generation, said Robert Califf, has done enormous good. It has delivered evaluations of the benefits and risks of medical products and interventions that have enabled these technologies to have a dramatic impact on life expectancy, physical function, and the ability to enjoy life. However, the traditional evidence-generation system has become "bloated and burdened" with practices that massively increase the cost of research without necessarily improving the quality, he said. Califf said that while the old system has not failed, the current time is an important inflection point with the opportunity to refocus efforts and dramatically accelerate the generation of evidence while also improving quality.

Science is in an "explosive phase," said Califf, in which there is a proliferation of technologies and new approaches for research, prevention, diagnosis, and treatment. At the same time, the costs of doing research are increasing. In some cases, this has resulted in "putting things on the shelf because we can't afford to do the development," said Califf. In addition, the cost of health care has been rising, which is leading health systems to try to assess the comparative value of old and new therapies. The result of all of these changes is a dramatic need to generate more high-quality evidence about diagnostic and therapeutic technologies and clinical strategies. However, the current evidence-generation system is well intentioned, but flawed: it is expensive, slow, not always reliable, unattractive to clini-

cians and administrators, and does not answer the questions that matter most to patients. This old, unsustainable system, said Califf, was built at a time when technology was limited, clinical notes in health records were handwritten, and there were few electronic data in the context of routine clinical care. As computing advanced faster in non-medical sectors than in the practice of medicine, research experts developed parallel systems to record clinical findings entirely separately from clinical practice; the perpetuation of this "parallel universe" of data and antiquated systems led to some "bizarre" inefficiencies. For example, said Califf, as electronic health records (EHRs) developed, research coordinators were instructed to print out notes in order to produce a written record, which would then be checked against the electronic system. These types of "arcane practices" were codified and amplified through the development of good clinical practices (GCPs) and standard operating procedures (SOPs), he said.

One practice that is particularly troublesome, said Califf, is the idea that recording each data point with as much precision as possible will result in a more reliable estimate of treatment effect. However, this "patently incorrect" belief results in wasting millions of dollars, without an appreciable increase in the quality or utility of the evidence generated. As an example, Califf pointed to a case in which thousands of patients were studied to determine the dosing regimen of a certain drug. A U.S. Food and Drug Administration (FDA) inspector expressed a lack of confidence in the results because the exact time the drug was ingested was not recorded. Recording the time of ingestion would have cost "probably on the order of $10 million," and would have likely contributed little information in a drug administered twice per day. What is needed now, said Califf, is moving away from this narrow focus on precision to a broader focus on reliability. Clinical trials should be designed and conducted in order to produce reliable results that meet the needs of patients, providers, payers, and policy makers, said Califf.

Califf drew a distinction between a system focused on precision and a system focused on reliability by offering the definitions of each word:

- Precision: (1) The quality, condition, or fact of being exact and accurate. (2) Refinement in a measurement, calculation, or specification, especially as represented by the number of digits given.
- Reliable: Consistently good in quality or performance; able to be trusted.

Califf explained that for some types of research, precision is critically important. For example, if small samples and measurements are expensive to take, precision is essential. Precision is also important in early phases of research when little is known, or for research where the administration

of a drug must be carefully timed. However, for many studies, a focus on precision limits the potential of research by creating budget requirements that severely limit the size of the study that can be conducted, the duration of follow-up, or the number of endpoints that can be examined. For these studies, the more important characteristic is that the results are dependable, sound, and able to be trusted, he said. In essence, these studies should focus on "providing the answers to the questions that matter to the patients."

An evidence-generating system that focused on reliability, said Califf, would have four key principles:

1. Build a reusable system embedded in practice;
2. Use quality by design;
3. Use automation for repetitive tasks; and
4. Operate from basic principles.

First, an evidence-generating system focused on reliability would be a learning system that is embedded in clinical practice, and would enable learning from every encounter (see Figure 5-1). This is an old concept in health care and is a fundamental concept in business. In addition, the lessons learned would return to the point of care and be used to improve care. Califf observed that this capability is being developed by public–private partnerships and integrated health systems, and noted that "if these various systems can work together in a federated way, I think we are getting close to having a national system that can be reused at a very low cost for different kinds of questions." This vision of a new national system would collect data during routine care, would use active surveillance to protect patients, would leverage RWE to support regulatory decisions, and could be used to inform decisions by all stakeholders in the ecosystem, said Califf.

Second, the system would use a quality-by-design approach in order to focus on and eliminate errors that bias the results, while ignoring errors that do not affect the outcome. Trying to eliminate *all* errors, said Califf, is costly, inefficient, and unnecessary. Califf said the quality-by-design process requires researchers to think through the objectives of the trial, identify the factors that are critical to meeting those objectives, and work to mitigate the risks that are likely to lead to errors that matter. Califf directed workshop participants to the quality-by-design toolkit for further information.[1]

The third principle of a reliability-focused system, said Califf, would be to capitalize on the rapidly expanding technologies and infrastructure that are available. Examples would be using automation for repetitive tasks, performing real-time analysis of data that are routinely collected, and

[1] The Clinical Trials Transformation Initiative's quality-by-design toolkit can be found at http://www.ctti-clinicaltrials.org/toolkit/QbD (accessed November 2, 2018).

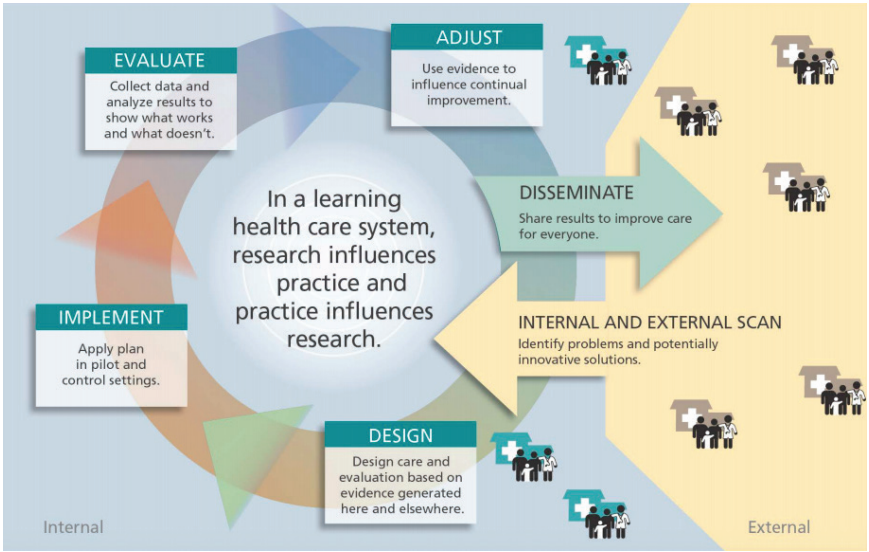


FIGURE 5-1 Learning health care system.
SOURCES: Califf presentation, September 20, 2017; Greene et al., 2012.

developing infrastructure to share the results with practitioners to support a constantly learning system. Califf noted that automated analysis of data could help fill in the evidence gaps on endpoints that are less well understood, at a considerably lower cost.

Finally, a reliability-focused system should operate from basic principles of good scientific research, rather than simply creating another "bureaucratic entanglement" of new and different SOPs (see Box 5-1). Some basic principles, said Califf, would include

- Focusing on errors that matter;
- Enrolling study participants who are likely to inform the question;
- Randomizing;
- Masking;
- Measuring outcomes in a manner that is fit for purpose;
- Considering strengths of different designs for different purposes; and
- Designing operations that yield an answer to the question in an efficient manner.

BOX 5-1
A Patient's Perspective on Randomized Controlled Trials (RCTs) as Presented by McCollister-Slipp

Anna McCollister-Slipp, chief advocate for participatory research at Scripps Translational Science Institute, founder of VitalCrowd, and co-founder of Galileo Analytics, talked about her eye-opening experience as a patient involved in a clinical trial. McCollister-Slipp has been managing her type 1 diabetes for decades with multiple medications and devices, many of which require monitoring specific outcomes that are seemingly unrelated to her actual health, she said. As a result of being a "frustrated patient," she co-founded Galileo Analytics, which takes real-world evidence and puts it in a platform that is easy, fast, and agile. Through Galileo and other ventures, McCollister-Slipp became involved in discussions about research policy and clinical research design, but had never been a participant in a clinical trial herself. A few years ago, McCollister-Slipp was selected for a trial, and the experience was "incredibly informative" as someone involved in the research space, she said. The questions that were asked of participants were "completely irrelevant" to what was actually being studied; she was required to fill out handwritten diaries and answer questionnaires; and she was asked to write down data that were already structured and machine readable. This experience led her to have doubts about the validity of the data that are collected in clinical trials. She wondered if the other participants, like her, were "just rushing through their diary to get it done," and whether participants had enough vested interest in the trial to enter the data correctly and accurately. While she had already known about the challenges of generalizability that are inherent to RCTs, her experience led her to question the validity of the data itself.

INTEGRATING THE NEW WITH THE OLD

The traditional evidence-generation system, said Califf, is not necessarily broken, but needs dramatic improvement. Several speakers addressed the issue of how to integrate RWD and RWE with the traditional evidence-generation system, and more generally, how to generate better answers for questions, no matter what method or source of data.

Real-World Evidence to Address Challenges Along the Drug Development Pathway

John Graham, head of value, evidence, and outcomes at GlaxoSmithKline (GSK), echoed other speakers in his opening remarks: "We need to have the right answers to the right questions at the right time." In getting these answers, said Graham, RWE is a must-have, but it is not a replacement for traditional research. Using both traditional and new methods of research

and sources of data will lead to a richer and more informative body of knowledge to inform decisions. GSK has been trying to move from a "study-by-study" assessment process to a "challenge-based thinking process," said Graham. The challenge-based process starts with clarifying the end goal: What does the patient need to have an improved outcome? The second step is identifying the challenges to getting the patient to that improved outcome. Finally, "We look for a book of work that can resolve that challenge," he said. The book of work can include multiple sources of data and types of studies, and can include evidence from traditional RCTs as well as RWE. Graham stressed that trials that include randomization are essential for understanding causation and laying a base of knowledge. However, RWE can be a useful adjunct to RCTs in order to expand understanding of disease and the patient experience.

Graham said that there are broad uses for RWE along the drug discovery and development pathway (see Figure 5-2). Much of the focus is often on using RWE in prediscovery or in postmarketing evaluation. However, said Graham, there are unique capabilities of RWE that can be used in other phases as well. For example, social media can be used to understand the perspectives and needs of patients. Graham noted that patient advocacy groups often give the perspective of the "professional patient," but social media allows GSK to tap into the "real-world" patients. He said that as researchers are planning RCTs, they can design components such as outcomes, measures, and tools in ways that are most appropriate for the patients. Another unique use of RWE is to understand the thresholds of effect that are most important for patients; for example, a drug could be produced that results in a small reduction in blood glucose in diabetes patients, but this effect might be too small for patients to want to take the drug.

Each step of drug discovery and development has challenges and tough decisions; RWE can help focus drug development along the entire pathway, said Graham. RWE can be used during discovery to estimate unmet needs, or to characterize patient heterogeneity. RWE can be used during development to form clinical pathways, optimize trial design, and optimize price. RWE can be used after approval to study comparative effectiveness, learn about compliance and adherence patterns, and investigate effectiveness in subpopulations. RWE, concluded Graham, is not about one method or one source of data, but about using information to overcome challenges and improve patient outcomes.

The Need to Streamline and Continue the Use of Randomized Controlled Trials

Several speakers at the workshop identified problems with the traditional reliance on RCTs, and proposed using non-randomized observational

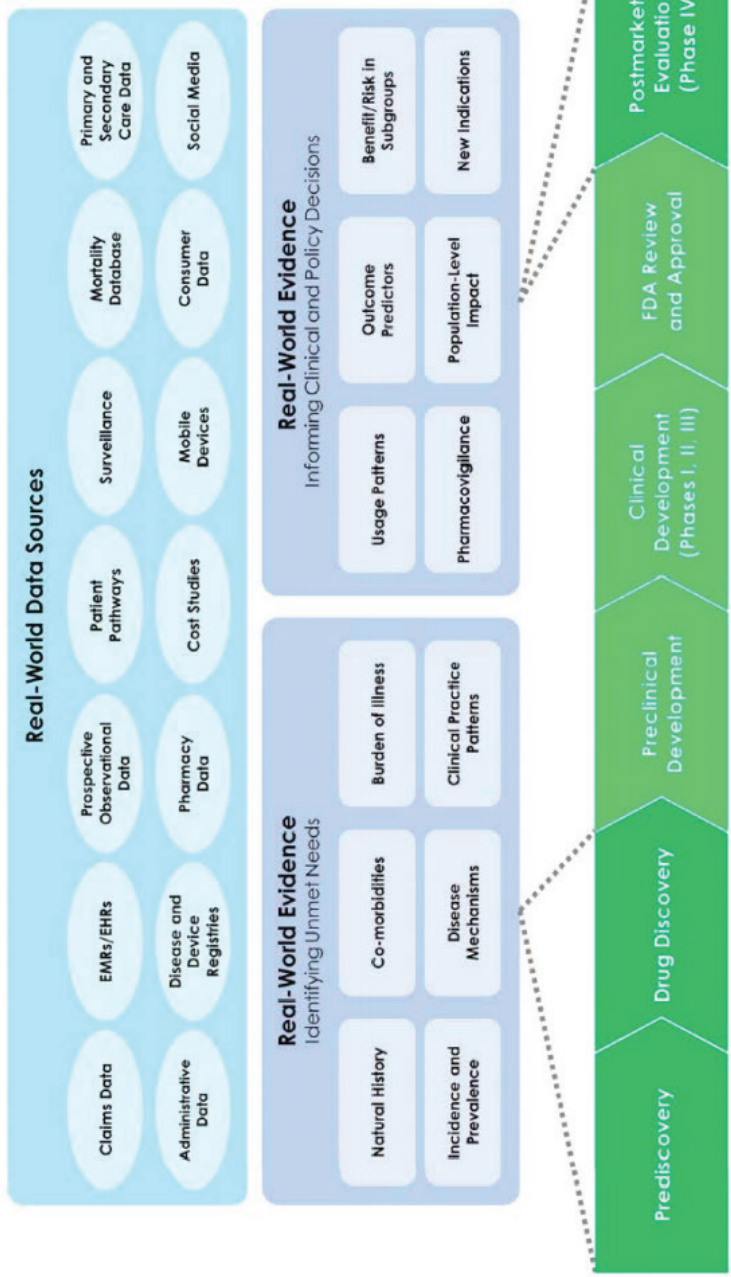


FIGURE 5-2 Broad sources of real-world data and broad uses of real-world evidence.
NOTE: EHR = electronic health record; EMR = electronic medical record; FDA = U.S. Food and Drug Administration.
SOURCES: Graham presentation, September 20, 2017; Galson and Simon, 2016.

studies of RWD as an alternative to RCTs. Rory Collins presented a different perspective: RCTs are overly regulated, unnecessarily expensive, and focused on rules that are not based on scientific principles. In his opinion, the solution is not to replace RCTs with non-randomized observational studies, but instead to make it easier to do RCTs.

While non-randomized observational studies may be useful for detecting large effects of treatments on health outcomes that are rare, RCTs are necessary for detecting moderate effects of treatments on common health outcomes reliably, said Collins. Non-randomized observational studies, he said, are limited in detecting moderate treatment effects and causal associations. When based on large databases, such studies may find associations of health outcomes with treatments that are highly statistically significant and precise, but that does not mean they are causally related, Collins said. This is because the underlying risk of people who take the treatment and those who do not may differ *systematically*, even after statistical adjustment. By contrast, randomization allows differences in outcomes to be causally attributed to treatment, because the randomized patient groups differ only *randomly* from each other in terms of their underlying risk of events. Randomization also allows use of a blinded control group, which can help ensure events are ascertained similarly in the randomized treatment groups, yielding unbiased treatment comparisons.

The current challenges with RCTs, said Collins, are in large part due to the widespread misapplication of the GCP guidelines for clinical trials issued by the International Conference on Harmonisation (ICH). Collins said these guidelines are not based on key scientific principles that are critical for the generation of reliable results in RCTs, and that the complexity and costs of adhering to them are unsustainable. In addition, ICH-GCP is applied far more widely than its original purpose: It was developed only for registration trials of new drugs, but compliance with it is now also required by governments (e.g., the European Union Regulation for Clinical Trials) and non-commercial funders (e.g., the Gates Foundation). Collins gave several examples of ICH-GCP–related practices that are wasteful, inefficient, and ineffective:

- Requirement to record all adverse events (AEs), not just serious AEs;
- Requirement to record narratives for all serious AEs in case there is an excess of a particular AE;
- Demands for unblinded results for AEs (including even primary outcomes) during ongoing trials; and
- Annual reports required by regulatory authorities that are so long that safety signals risk being lost.

In short, the ICH-GCP guidelines put undue emphasis on the quality of the *data* in RCTs, said Collins, rather than on the generation by RCTs of reliable *results* about the safety and efficacy of the treatment being studied. In his opinion, focus on compliance with rules due to overregulation and related bureaucracy, rather than on innovative designs and good results in RCTs, has resulted in obstacles, delays, and high costs. As a consequence, he contended, it has led some researchers to pursue the alternative of using non-randomized observational studies—what he called the misuse of RWE—to assess treatments, despite their potential for biases. Noting that the ICH-GCP guidelines require specific qualifications for investors, source data verification, and regulatory documentation, Collins argued there is an urgent need to improve RCT methodology through the development of comprehensive new RCT guidelines based on key scientific principles required to generate evidence about the safety and efficacy of treatments that can be trusted.

Other individual workshop participants noted that a purpose of ICH-GCP was to give providers guidelines for how to conduct research well. Janet Woodcock, director of FDA's Center for Drug Evaluation and Research, noted that regulations are designed to protect people, as well as the scientific enterprise, from bad actors. However, with this caveat, Woodcock said that regulations for the 21st century need to have a more "flexible understanding of quality" and a structure that allows for consistent monitoring and course correction, rather than rigid and unyielding rules.

From "One Study at a Time" to "All by All" Analyses

Patrick Ryan, senior director and head of epidemiology analytics at Janssen Research & Development, spoke to workshop participants about the Observational Health Data Sciences and Informatics (OHDSI) program. OHDSI is an open science community, said Ryan, where anyone can participate in conducting research on observational databases. The goal of OHDSI is to "improve health by empowering the community to collaboratively generate evidence," he said. OHDSI operates across 20 different countries, with more than 200 researchers. Similar to Sentinel, OHDSI has a distributed data network with open community standards. Collectively, there are more than 60 databases that contain patient records for 660 million patients. OHDSI's strategy, said Ryan, includes methodological research in order to establish and evaluate scientific best practices before applying them to observational data. The results of this research are codified into open-source tools that the entire community can use, with all code shared on GitHub (a Web-based repository for code).[2] This open-source approach,

[2] For more information about GitHub, see https://github.com (accessed November 2, 2018).

said Ryan, is part of OHDSI's "moral obligation to . . . generate the evidence and get it out to patients as quickly as possible." Ryan described the three focal points of OHDSI's research:

1. Clinical characterization: The diagnoses, treatments, and outcomes for a population;
2. Patient-level predictions: The probability of an individual patient developing the disease or experiencing an outcome; and
3. Population-level effect estimation: What are the causal effects between treatments and outcomes?

Focusing on population-level effect estimation, Ryan conducted a live demonstration of evidence analysis for the workshop participants. Ryan started with a paper about antidepressant medication use and the risk of preeclampsia in pregnant women with depression (Avalos et al., 2015). The study found an observed association between antidepressants and preeclampsia, and found that the association was stronger for selective serotonin reuptake inhibitors (SSRIs) in particular, with a statistically significant relative risk of 1.4, said Ryan. Another observational study, said Ryan, looked at the same question, using data from the Medicaid population (Palmsten et al., 2013). This study, in contrast with the first, found that other types of depression medications (serotonin and norepinephrine reuptake inhibitors and tricyclics) were associated with a higher risk of preeclampsia than SSRIs. In this study, SSRIs had a non-statistically significant relative risk of 1.00. A third paper that Ryan presented was a meta-analysis of research that looked at the link between antidepressants and preeclampsia. This meta-analysis concluded that "while some studies have suggested a moderately increased risk, the current data do not allow for a definitive conclusion." The meta-analysis pointed out the methodological limitations of many of the studies, and the fact that untreated depression and anxiety could not be disentangled from the results.

Ryan showed the audience a funnel plot that contained data from these studies, as well as other studies that had been mined from the published literature on the topic of antidepressants and preeclampsia (see Figure 5-3). The pattern on the funnel plot, said Ryan, "should be alarming." The plot showed that the evidence is skewed toward the right, that is, more results are positive than negative. In addition, 70 percent of the results are statistically significant and many of these are hovering right at the dashed line that represents a p-value[3] of 0.05. This pattern, said Ryan, suggests that "some-

[3] A p-value represents the probability of finding the observed results if the null hypothesis were true. A p-value of less than 0.05 is often used to determine whether a result is statistically significant.

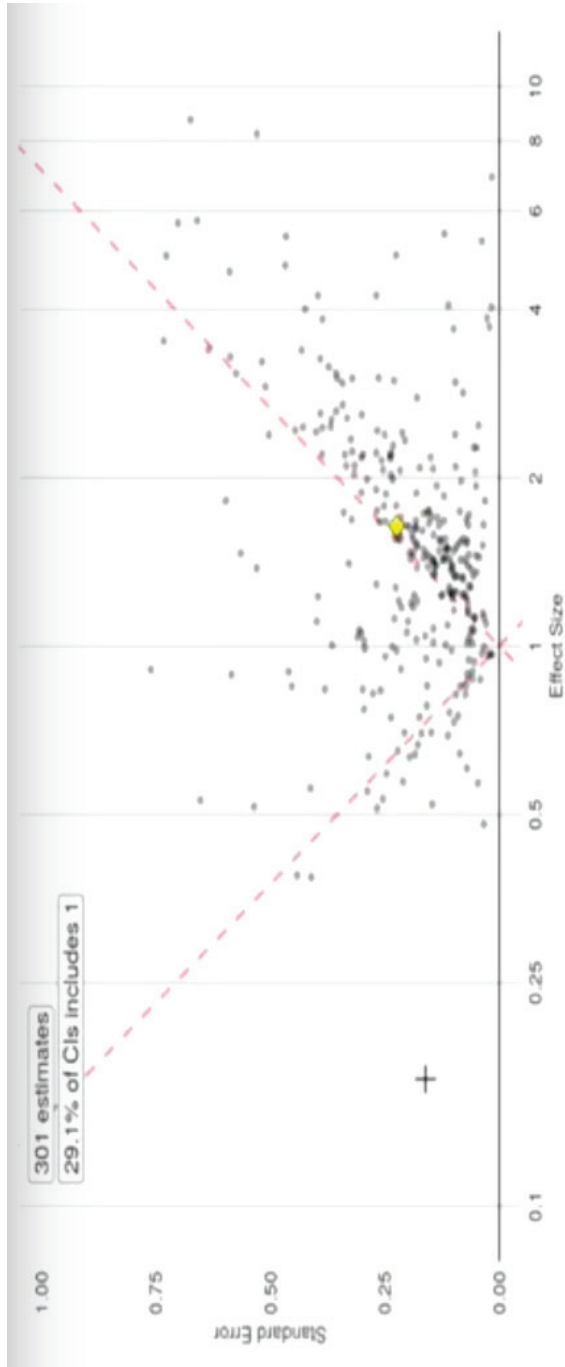


FIGURE 5-3 Effect estimates from published observational studies on antidepressants and the risk of preeclampsia.
NOTE: CI = confidence interval.
SOURCE: Ryan presentation, September 20, 2017.

one kept working at the data until they got p less than 0.05 and then quit." Ryan showed the workshop participants another funnel plot that plotted 60,000 published observational studies on multiple disease states (see Figure 5-4). This plot showed that, again, the studies were skewed toward positive results, and 80 percent of the published studies were statistically significant, with many studies hovering right at the 0.05 line.

This exercise, said Ryan, demonstrates that "we can't necessarily trust the process that we are using to generate evidence as a community." Our current process, Ryan said, is to conduct one observational study at a time, with one hypothesis, one dataset, and one method. Each of these studies is viewed individually, but given the pattern on the funnel plot, "it can't possibly be the case that all of these studies are totally correct." The process of generating evidence in RCTs, said Ryan, is not much better. He pointed to a meta-analysis of multiple clinical trials on depression. The meta-analysis concluded that there was not sufficient evidence to draw conclusions about the comparative risk of side effects, including suicidality, cardiovascular events, and seizures. If "we still don't know the answer, despite decades of research and hundreds of millions spent on this question," he asked, what could be done differently?

Ryan suggested that obtaining the answers we need is best done by considering the patient perspective. An individual patient with depression, said Ryan, wants to know which of many available treatments would be best for him or her. Different patients may prioritize different factors; for example, one patient may want to know about suicidality while another is more concerned about hepatotoxicity. Ideally, every patient would have access to every personally relevant data point. The way to do this, said Ryan, is through an observational data network, like OHDSI, of multiple standardized data sources that can answer questions one at a time. For example, a person could "ask" the network about whether one specific antidepressant increases the risk of diarrhea more than another. The observational data in the network may or may not be statistically significant, and may have large or small effect estimates, but the person asking the question can see all of this information and make a decision accordingly.

The question that remains, however, is how do researchers know these data are reliable? Ryan showed workshop participants another funnel plot. This figure graphed every observational data point within the OHDSI system, comparing all antidepressants against each other on all outcomes—an "all by all analysis." This funnel plot (see Figure 5-5), unlike the plots from RCTs, does not have a preponderance of data hovering around the line of statistical significance, and the huge majority of the dots are not statistically significant. Because these data are not subject to a researcher or a publication deciding what to publish, they reflect the true breadth of data and not a subjective selection. This pattern, said Ryan, suggests that the evidence

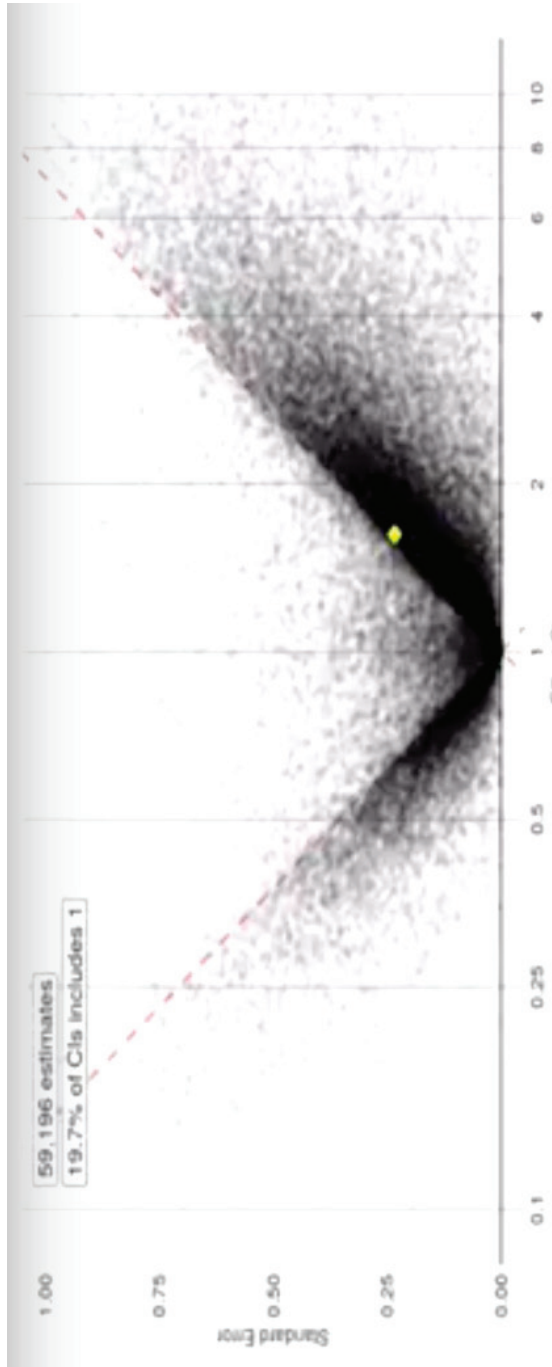


FIGURE 5-4 Effect estimates from published observational studies on all disease states, all treatments, and all causal effects.
NOTE: CI = confidence interval.
SOURCE: Ryan presentation, September 20, 2017.

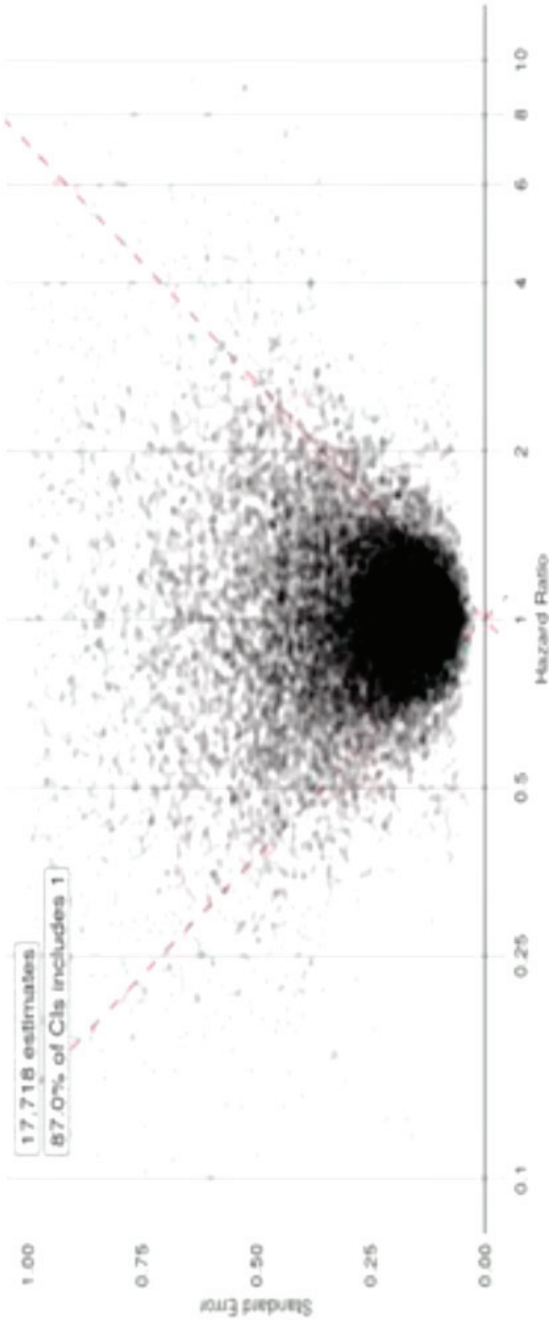


FIGURE 5-5 Systematically generated evidence from observational data, comparing all depression treatments on all outcomes of interest.
NOTE: CI = confidence interval.
SOURCE: Ryan presentation, September 20, 2017.

in these databases is less biased than the evidence from the published trials, and therefore more reliable for making decisions. There is still variability within these data, said Ryan; however, the systematic approach helps to remove the variability that is introduced and leaves the variability that is inherent to patient heterogeneity and health system bias and other factors.

Randomized trials, said Ryan, are still appropriate for many purposes. However, in the current system, there are many clinical areas in which trials have not occurred and practitioners operate without evidence. Using this systematic approach to examine the totality of observational evidence in order to generate answers is one solution to this evidence gap, he said.

REGULATORY PERSPECTIVE

Woodcock agreed with the speakers on several points. The current evidence-generation system for medical products is very costly and time consuming, and leaves many questions about product use unanswered. One consequence of this situation, Woodcock said, is that many clinical decisions are not evidence based because generating the answers is too expensive. As one potential solution to this problem, Woodcock said that "FDA is committed to exploring the use of real-world evidence in regulatory decisions."

FDA is exploring the use of RWE in several ways, Woodcock said. In the drug space, FDA is involved in a demonstration project called IMPACT-AFib, a randomized educational intervention that uses Sentinel to examine outcomes (see Richard Platt's presentation in Chapter 3). In the device space, FDA's Center for Devices and Radiological Health and Center for Biologics Evaluation and Research issued guidance in mid-2017 about the use of RWE for device decisions. The device guidance, titled "Use of Real-World Evidence to Support Regulatory Decision-Making,"[4] discusses the challenges with current device evidence development, and proposes potential uses of RWE for device regulation. These uses include using RWE to examine outcomes, but also as historical or concurrent controls, to expand the label for an approved device, or for safety surveillance. This guidance, said Woodcock, should serve as an incentive for the device industry to invest in making RWE generation more robust. Similar considerations apply to the use of RWE for drug approval, said Woodcock. RWE has long been used for the evaluation of safety in the postmarketing of products, but there is little historical use of RWE for decisions about effectiveness. However, she noted that "there are no hard and fast rules" about how evidence must be generated for drug approval, with the exception of rules about

[4] See https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm513027.pdf (accessed November 7, 2018).

informed consent and patient privacy. FDA is open to a wide spectrum of evidence, from standard clinical trials to pragmatic trials conducted in the health care system. However, she noted, there are trade-offs involved in the choice of trial design, including data reliability, pragmatism, control of errors, safety, and other factors, and FDA would consider these trade-offs when evaluating RWE. For example, it would be inappropriate to run a first-in-human trial in a real-world setting, she said.

Woodcock gave several examples of ways in which FDA has used or is considering using RWE for drug approval. Drugs for rare diseases have been approved using data from registry-like case series, she said. For example, Lumizyme for Pompe disease was approved using survival data from an international registry of infantile-onset disease. Registry data have also been used for external controls for uncontrolled experience data, she said. FDA is exploring how randomization would work in registry or health care settings, and they are collaborating with other stakeholders to improve the validity of key data elements that are collected during the course of health care. Woodcock referred to the quality-by-design approach that Califf had mentioned, and said using this approach ensures that "you get it right the first time"—data are put into the EHR correctly and do not have to be adjudicated and curated later.

There are several potential uses for RWE during the drug development process, said Woodcock, including

- Natural history information: RWE is valuable for learning about patients' experiences with a disease, and what their burdens and needs are. RWE can help develop appropriate outcomes for a study, based on patient progression and self-reported outcomes. This is particularly true for rare diseases and/or diseases that are very heterogeneous. Rare disease experts, she said, are often wrong, because their opinions are based on the few patients that they have seen, so it is essential to get RWE from as many patients as possible.
- Biomarker development: Understanding biomarkers and choosing appropriate markers is important for developing drugs in an efficient way. Biomarkers that are critical to development should be explored in humans as thoroughly as possible before initiating a study, and RWE approaches could be essential to gathering this information.
- Hybrid model for investigational drugs: There are ways to combine traditional study approaches with RWE. For example, an investigational drug could be evaluated in a hybrid model that uses traditional randomization for initial assignment of patients and uses RWE to measure outcomes. This approach requires integrat-

ing the trial into the health care process, and collaborating with caregivers as research partners. A good example of this approach is the National Institutes of Health Collaboratory.

- Add new indications to an approved drug: When extending the label of an approved drug, RWE may be compelling enough that an RCT is not needed. For example, ivacaftor, a drug used for cystic fibrosis, could have been approved for additional mutations based on registry data, combined with trial and mechanistic data.

Woodcock said that for any use of RWE, the important part is ensuring that the research protocol is designed thoughtfully and appropriately. If the design is excellent and takes into consideration potential errors and how to manage them, FDA or any other regulatory body could agree on the alternative design and agree to accept the evidence.

One design approach that is particularly promising, said Woodcock, is the use of master protocols. Master protocols are continuous, ongoing trials that can study multiple interventions and outcomes, with the goal of having "continuous improvement in the disease outcome." The use of master protocols, said Woodcock, saves time, offers an opportunity to include community practitioners and integrate research and practice, can answer multiple questions, is patient-centric, and can use adaptive designs creatively. However, there are also challenges involved with this approach: It is a novel approach that is difficult to set up at the beginning, and it does not comport with the traditional models of pharmaceutical development, academic rewards, or grant funding. Master protocols offer an opportunity to incorporate RWD as extensively as possible, said Woodcock, although this will require additional work in standardization, data verification, training, and curation. These initial investments, however, will likely pay off in terms of lower costs, greater efficiency, the engagement of first-line practitioners, and the ability to answer more questions.

Science and medical care are rapidly changing, said Woodcock, and these changes mean "that we are going to have to change our traditions." More rare and orphan diseases are being studied, and even in common diseases, there are targeted therapies with companion diagnostics. These changes are narrowing the target population for a medical product in such a way that traditional trials do not work very well, she said. The current inability to efficiently generate needed evidence for drug development and for clinical practice, she said, will continue to be a major barrier to innovation and the quality of care. Drug developers and regulators will have to adapt to this new world with innovative designs and the use of RWE to get the answers that patients need. Woodcock stressed the need for pragmatism in research and for improving the current situation. She noted that what clinicians do now is often "based on observational studies or, even worse,

individual experience." The research community could instead focus on making evidence generation easier and more efficient, while still emphasizing reliability, in order to get the answers that patients, providers, and regulators need to make decisions.

6

When Is a Real-World Data Element Fit for Assessment of Eligibility, Treatment Exposure, or Outcomes?

Key Messages Identified by Individual Workshop Participants

- The usefulness of a real-world data (RWD) source for a particular question depends on whether it has information about the correct population, exposures, and outcomes. (Altan, Yaist)
- Accuracy and confidence in RWD vary predictably, depending on factors such as treatment administration method or the outcome being measured. (Cao)
- RWD of different quality for different purposes may be acceptable. (Platt, Temple, Yaist)
- RWD collected by providers are affected by both the experience of the provider and the incentives they face. (Hernandez, Simon)
- Challenges in RWD analysis are often also challenges to improving patient care, such as fragmentation. (Berliner)
- Organizations interested in conducting or assessing RWD studies often use specific, sequenced questions to determine study feasibility, identify appropriate data sources, and assess methods and tools for use in the study. (Altan, Ball, Yaist)
- Patient-generated health data can be collected nearly continuously, come from many sources, answer research questions that were not previously answerable, and potentially facilitate access and participation from otherwise unrepresented patient populations. It can present difficulties in analysis or data stor-

age, and is subject to different biases than data collected within the health care system. (Foschini)
- RWD can be affected by systemic bias as well as random bias, and is unique from other data sources because of its dynamic nature. These factors can be compensated through various techniques, but are important for researchers to be mindful of as they are using the data. (Altan, Berger, Berlin, Foschini, Graff, Hernandez, Izem, Simon, Yaist)
- Data sharing and transparency in data curation and analysis techniques could be improved to encourage broader use of reliable real-world evidence. (Berger, Lieberman, McGraw)

While the first workshop explored the general issues concerning the use of real-world data (RWD) and real-world evidence (RWE), many individual participants at the second and third workshops drilled down into these issues in an attempt to identify specific questions to consider before using RWD and RWE in a study design. During the second workshop, individual participants suggested sets of questions organized by topic based on the workshop's three sessions. These questions, the discussions at the second workshop, and additional work by several individual workshop participants between the second and third workshops informed further refinement of the questions into several "decision aids." The decision aids were discussed at the third workshop and were intended to prompt discussion among the participants and inform them and potentially other stakeholders about topics in study design. Several sessions at the third workshop focused on these decision aids and explored the "sticking points" that individual workshop participants had identified at the second workshop. The presentations and discussions in the second and third workshops are covered in Chapters 6 through 9 of these proceedings, divided by the topic areas of the decision aids.

REAL-WORLD DATA ELEMENTS

As discussed in Chapter 1, there is no consensus on a definition of RWD. However, Brande Yaist, senior director of global patient outcomes and RWE at Eli Lilly and Company, said there are common themes among definitions. In particular, RWD are data that are derived from a variety of real-world sources, such as electronic health records (EHRs) and claims data, pragmatic trials, registries, social media, and directly from patients (see Figure 6-1). There are also hybrid approaches, where RWD are used in combination with primary clinical trial data collection.

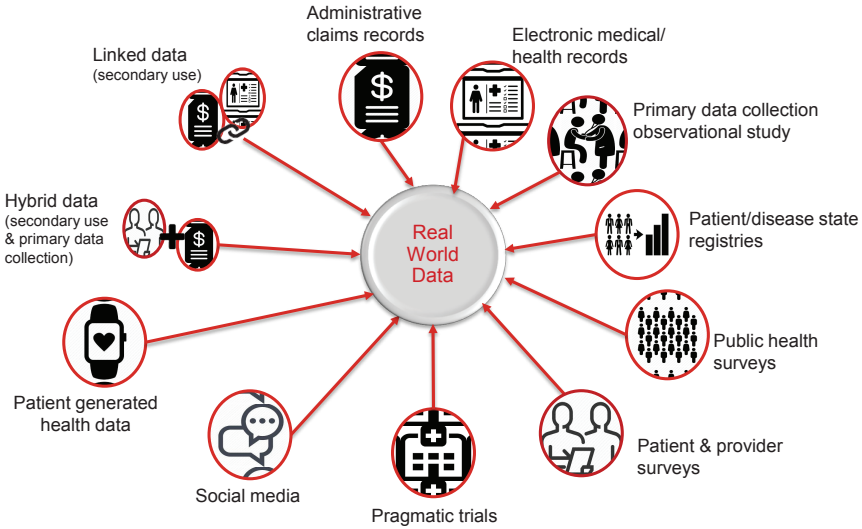


FIGURE 6-1 Possible sources of real-world data.
SOURCE: Yaist presentation, July 17, 2018.

With this wide definition of RWD, there is an almost infinite amount of data. Gregory Daniel, deputy director of the Duke-Margolis Center for Health Policy, asked when can we rely on these data? That is, when can RWD be used to assess characteristics of participants in a trial, such as eligibility, baseline health state, or key prognostic factors? When can RWD be relied on to assess patient outcomes? When can data that are generated by patients or by their devices be considered reliable? Perhaps most importantly, asked Daniel, how can the reliability of RWD be assessed *before* time and money are spent to conduct a study?

To answer these questions, workshop presenters and participants discussed the reliability of RWD, using illustrative examples to elucidate some of the main challenges, and referring to the draft decision aid (see Figure 6-3 later in this chapter).

ILLUSTRATIVE EXAMPLES

To explore the issues surrounding the use of RWD, speakers at the second and third workshops presented case studies as illustrative examples of the considerations that go into designing and conducting a real-world study.

NOACs Versus Warfarin

At the second workshop, Adrian Hernandez, vice dean for clinical research at the Duke University School of Medicine, presented a suite of trials that compared novel oral anticoagulants (NOACs) with warfarin. Oral anticoagulants like warfarin have long been used in patients for a number of indications, including reducing the risk of stroke and embolism in patients with atrial fibrillation. However, there are challenges involved with warfarin use, such as requirements for patient monitoring, said Hernandez, and many patients do not receive effective management (Go et al., 1999). The use of warfarin is suboptimal, even among high-risk patients, he said (Waldo et al., 2005). Researchers have been developing novel anticoagulants as an alternative to warfarin, and there were four pivotal trials used for approval of these drugs for the treatment of atrial fibrillation and risk reduction for stroke. Together, these trials enrolled more than 70,000 patients to compare NOACs with warfarin; each used slightly different methods and targeted different types of patients, but all involved some use of RWD. The trials presented were the following:

- Randomized Evaluation of Long-Term Anticoagulation Therapy (RE-LY) (Connolly et al., 2009);
- Rivaroxaban Once Daily Oral Direct Factor Xa Inhibition Compared with Vitamin K Antagonism for Prevention of Stroke and Embolism Trial in Atrial Fibrillation (ROCKET AF) (Patel et al., 2011);
- Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation (ARISTOTLE) (Granger et al., 2011); and
- Effective Anticoagulation with Factor Xa Next Generation in Atrial Fibrillation-Thrombolysis in Myocardial Infarction 48 (ENGAGE AF-TIMI 48) (Giugliano et al., 2013).

The outcomes of these trials, said Hernandez, were largely consistent and showed that NOACs were non-inferior to warfarin. A 2014 meta-analysis showed that all four trials favored NOACs over warfarin for the risk of stroke and systemic embolic events, as well as secondary outcomes, such as ischemic stroke, hemorrhagic stroke, myocardial infarction (MI), and all-cause mortality (Ruff et al., 2014).

Hernandez outlined some of the major difficulties in conducting these studies. Three of the trials—which all enrolled thousands of patients—were double blinded. This led to, said Hernandez, enormous challenges for the investigators in terms of monitoring, ensuring standard of care, and adjudicating the outcomes (see Chapter 8 for more details on blinding).

Hernandez suggested that in order to assess the quality of these trials, one could use time in therapeutic range as a surrogate for quality. Warfarin has a narrow therapeutic window—for most indications, the international normalized ratio (INR) should be between 2 and 3. If the INR is lower, the risk of ischemic stroke is higher; if the INR is higher, the risk of intracranial hemorrhage is higher. For this reason, patients on warfarin must be monitored frequently to ensure that they are in the appropriate therapeutic window. In the ROCKET-AF study, a large majority of patients were in the therapeutic range (with an INR target of 2.5, inclusive from 2 to 3).

Friends of Cancer Research Pilot Project

At the third workshop, Jeff Allen, president and chief executive officer of Friends of Cancer Research, talked about a pilot project that investigated the performance of real-world endpoints among patients with advanced non-small cell lung cancer (aNSCLC) who were treated with immune checkpoint inhibitors. The goal of the project, said Allen, was to explore potential endpoints that may be fit for regulatory purposes as well as to assess the long-term benefits of a product. The project used a retrospective observational analysis design with patient-level data. The data were derived from EHRs and claims databases from six data partners. The study was conducted in a distributed manner, said Allen, which meant that each partner maintained and analyzed its own data; the partners collaborated to develop common data elements and methodological approaches so that the analyses would be as similar as possible. Allen described the three research objectives:

1. Characterize the demographic and clinical characteristics of aNSCLC patients treated with immune checkpoint inhibitors.
2. Assess ability to generate real-world endpoints in aNSCLC patients treated with immune checkpoint inhibitors, and segmented by clinical and demographic characteristics.
3. Assess performance of real-world endpoints as surrogate endpoints for overall survival.

Investigators defined and assessed five endpoints of interest for this population, all of which could be gleaned from the RWD sources. Allen noted that it was remarkably challenging to define and standardize these endpoints among different data sources to ensure that the study was comparing "apples to apples." Allen also noted that these endpoints were not necessarily indicative of the full potential of the partners' datasets, but rather represented the common denominator among them all. The endpoints defined were

- Overall survival (OS): The length of time from the date that treatment was initiated to the date of death;
- Time to next treatment: The length of time between initiation of treatment and initiation of the next systemic treatment;
- Time to treatment discontinuation: The length of time between initiation and discontinuation of treatment;
- Progression-free survival: The length of time between initiation of treatment and a progression event (as evidenced in the patient's chart) or death; and
- Time to progression: The length of time between initiation of treatment and a progression event, but excludes death as an event.

Allen noted that the intention of this study was not to compare different drugs or readjudicate clinical trials, but rather to look at what evidence could be extracted from diverse data sources and what the strength of that evidence would be. For the first research objective of characterizing the patients, Allen said they found "really great consistency [among] the different characteristics." While the patient demographics and characteristics were not identical across the six databases, they were relatively consistent in terms of age, histology, sex, and treatment.

On the next objective—assessing the ability to generate real-world endpoints—there were some challenges, said Allen. Even a simple endpoint such as death of the patient can be challenging to collect, due to the limited availability of accurate and timely death records. Despite these challenges, said Allen, there was relative consistency among the datasets on many of the endpoints (see Table 6-1). When segmented by patient demographics, there was again relative consistency, despite some outliers. This consistency among datasets suggests that these types of data could be used to assess patient populations when randomized controlled trials (RCTs) are not feasible or desirable.

The third research objective was to evaluate the correlations between these real-world endpoints and overall survival (see Table 6-2). These correlations, while not "overwhelmingly strong," were consistent, said Allen. This suggests that these real-world endpoints—readily accessible in EHRs—could potentially serve as surrogate endpoints for overall survival.

Finally, the investigators wanted to examine how closely the overall survival rate from the real-world datasets would align with what had been observed in clinical trials for these drugs. Using data for patients on any of three different immune checkpoint inhibitors, the investigators compared the real-world overall survival rates with the ranges that had been observed in pivotal clinical trials for each drug. The overall survival rates from the databases were generally in line with the rates from the clinical trials, said Allen (see Figure 6-2). This came as a bit of a surprise, he said, because

TABLE 6-1 Real-World Endpoints in Six Datasets

Data Set	rwOS	rwTTD	rwTTNT
Data Set A	13.50 [12.80, 14.50][a]	7.03 [6.27, 9.97]	22.50 [NA]
Data Set B	15.78 [12.20, 24.59]; 8.58 [7.56, 10.26][b]	3.25 [2.76, 3.75]	
Data Set C	8.67 [6.83, 10.02]	4.70 [3.68, 5.52]	11.60 [8.80, 16.10]
Data Set D	9.15 [8.82, 9.51]	3.21 [3.21, 3.44]	14.03 [12.89, 15.15]
Data Set E	12.69 [11.70, 13.87]	3.63 [3.40, 3.87]	12.07 [11.24, 13.48]
Data Set F	12.30 [9.61, 16.94]	4.60 [3.71, 6.32]	12.50 [9.29, NA]

[a] Overall survival was calculated as days between I/O initiation and disenrollment.
[b] Sites with social security or state death data, censored at estimated earliest date such data should be available if no death was observed.
NOTE: NA = not applicable; rwOS = real-world overall survival; rwTTD = real-world time to treatment discontinuation; rwTTNT = real-world time to next treatment.
SOURCE: Allen presentation, July 17, 2018.

TABLE 6-2 Correlation Between Real-World Overall Survival and Real-World Extracted Endpoints

	rwOS Versus rwTTNT		rwOS Versus rwTTD	
Data Set	N	Correlation [95% CI]	N	Correlation [95% CI]
Data Set A	83	0.36	254	0.63
Data Set B			225	0.62 [0.54, 0.69]
Data Set C	96	0.70 [0.58, 0.79]	295	0.89 [0.86, 0.91]
Data Set D	1,203	0.61 [0.57, 0.64]	4,337	0.80 [0.79, 0.81]
Data Set E	358	0.62 [0.54, 0.68]	1,456	0.77 [0.75, 0.79]
Data Set F	39	0.46 [0.33, 0.81]	142	0.80 [0.66, 0.85]

NOTE: CI = confidence interval; rwOS = real-world overall survival; rwTTD = real-world time to treatment discontinuation; rwTTNT = real-world time to next treatment.
SOURCE: Allen presentation, July 17, 2018.

there had long been speculation that the survival rates of the homogeneous clinical trial populations might be lower once the treatment was applied to a more diverse real-world population.

Allen concluded with what he sees as the main takeaways from this pilot project. First, he said, there is a high level of shared patient characteristics among the datasets, despite the fact that the datasets have variable sample sizes and data capture processes. This similarity among sources

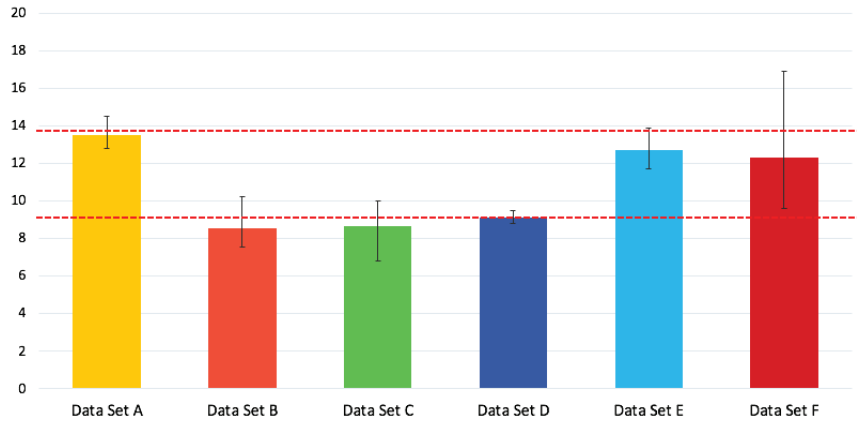


FIGURE 6-2 Real-world overall survival.
NOTE: Each bar represents a combination of information based on three different products in the class from each data partner. The y-axis shows the median overall survival in months of advanced non-small cell lung cancer (aNSCLC) patients treated with a checkpoint inhibitor. The red dashed lines represent the observed range in overall survival levels in the pivotal randomized controlled trials for the same products.
SOURCES: Allen presentation, July 17, 2018; concept/data from Huang et al., 2018.

demonstrates the feasibility of identifying aNSCLC patients from diverse RWD sources. Second, the study demonstrated that several real-world endpoints correlate well with OS. However, more research is needed to determine whether the endpoints could be reliable surrogates for OS, and whether these endpoints could support decision making by regulators and payers. Finally, the overall survival rates assessed from EHR and claims data were quite consistent with the rates observed in clinical trials, he said, suggesting a need for additional research on the association between data from real-world sources and data from clinical trials.

DISCUSSION: CHARACTERIZING REAL-WORLD DATA AND REAL-WORLD EVIDENCE

Following the presentation of the illustrative examples, the workshop participants discussed some of the general issues and overarching considerations with using RWD, in particular how one could characterize the utility of RWD before a study is performed. Several participants highlighted challenges in defining the population, exposure, and outcomes; concerns about

data collection by providers; considerations of whether and when expert adjudication might be necessary; and safety issues that could be addressed.

Defining the Population, Exposure, and Outcomes

One of the pivotal parts of assessing the quality and relevance of data is determining if the source has information about the right population, the right exposure, and the right outcomes, said Yaist and Aylin Altan, senior vice president of research at OptumLabs, during the third workshop. However, defining these elements is not as straightforward as it might seem. During the second workshop, Jesse Berlin, vice president and global head, Epidemiology, Johnson & Johnson, gave an example of trying to determine—based on RWD—which patients have diabetes. He said there are multiple codes, drugs, and other data points that could indicate that a patient has diabetes, but there is no obvious way to determine this with 100 percent certainty. He noted that a colleague had developed a predictive model that would classify the probability of patients being diabetic (which could mean that a patient would be a 0.8 diabetic, he noted). Gregory Simon concurred with the idea of a probabilistic model, noting that while we commonly use dichotomous classifications for medical conditions, many medical phenomena are "fuzzy." The line between an MI and not an MI, or between depression and not depression, is not "completely crisp," he said. A probabilistic approach would help to better capture the fuzziness of medical conditions, but could also be challenging for researchers and regulators to understand.

Exposure, said Berlin, can also be difficult to determine using RWD. For example, exposure to a drug is usually indicated through a prescription for the drug in the EHR, or a record of payment for the drug in the claims data. However, neither of these data points can prove that the patient is taking the drug as prescribed. Hernandez added that when using RWD to capture population, exposure, and outcome, it is possible to use sensitivity and specificity analyses to assess the robustness of the evidence. In other words, even though there is variability in the data, this variability can be accounted for in the analysis.

Hui Cao said that although it can be challenging to assess RWD sources for population, exposure, and outcomes, they are fairly straightforward in some situations. For example, for certain diseases (e.g., diabetes), established algorithms can identify the population with the disease. These algorithms can be employed before a study is conducted in order to understand how well the population can be identified and with what level of confidence, she said. For exposure, there is generally high confidence about accurately capturing drugs given by injection or intravenously because of the administration method. However, capturing oral medications or inhaled products can

be trickier, she said. In terms of outcomes, the data for certain events, such as hospitalization for MI, are fairly accurate. For other outcomes, such as laboratory measures or continuous variables, the data may be less reliable.

Robert Temple, deputy director for clinical science at the U.S. Food and Drug Administration's (FDA's) Center for Drug Evaluation and Research (CDER), added that while much of the discussion had been on assessing the quality of data for RWE studies, RWD can also be used to facilitate clinical trials. For example, RWD can be used to identify patients who may be appropriate subjects for a trial. For this use of RWD, the quality of the data does not matter as much, he said, because the participants will undergo further evaluation for enrollment.

Data Collection by Providers

One concern with RWD collection, said Simon, is whether providers can accurately assess the condition or event of interest. In RCTs, data are collected by providers specifically trained in the trial protocol, whereas in RWD, providers in real-world settings assess patients and must accurately and completely record necessary information. This accurate data collection is "foundational" to the idea of RWD, he said. If the assessment requires special training, technology, or tools, Simon suggested it would be better suited to an RCT. Hernandez emphasized the importance of examining the incentives (or disincentives) to accurate data collection that affect the providers. For example, said Simon, when Medicare changed its payment structure in a way that incentivized the diagnosis of "major depressive disorder" instead of "depression not otherwise specified," the "ratio of these two diagnoses in most large health systems flipped overnight." The epidemiology of depressive disorders did not change, said Simon, but rather the incentives that governed their recording.

Relatedly, said Hernandez, different EHR systems vary considerably, and these variations can impact how providers record information. For example, a colleague of Hernandez "teaches his Fellows to never code for diabetes in their EHR system, because it pulls up a laundry list of choices that you have to make which don't quite fit." As a consequence, RWD from this particular EHR system are not likely to have accurate data about diabetes, which could result in systematic bias, he said.

One workshop participant noted that while retrospective studies depend on how the provider put data into the system in the past, prospective pragmatic trials have the opportunity to improve this initial data collection. For example, researchers in the Salford Lung Studies (see Chapter 3) embedded prompts in the EHR to improve data collection, said Simon.

While much of the attention around data is often on assessing data quality once they are collected, said one workshop participant, the start-

ing point should be ensuring that data are collected in a standardized and accurate way, both for research purposes and clinical purposes. EHRs, the workshop participant said, are designed primarily for patient care; the primary goal should be for the EHR to capture data that are meaningful and useful for patient care, including facilitating, rather than complicating, patient care by providers. Elise Berliner, director, Technology Assessment Program, Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, concurred, noting that some of the challenges in data analysis are also challenges to patient care—for example, patients seeking care outside their usual health network, or missing or fragmented health data. Fundamentally, said Berliner, the health care system is about patients and their providers. She asked, "How do we work together with the other stakeholders to make the data infrastructure better and more reliable?" Simon noted that although different stakeholders use different terms, the data needs of all stakeholders are essentially the same: knowing whether the care that a patient received worked and whether it was safe.

Expert Adjudication

Expert adjudication, said Daniel, can sometimes be necessary to confirm that the recorded data are reliable or reasonably complete. Simon said adjudication is not for issues such as missing data or technical problems, but rather for validating that the source clinician correctly assessed the patient and accurately recorded the data. Unfortunately, he said, "We cannot put ourselves in a time machine and go back . . . and interview that patient ourselves." Adjudication generally means using low-quality text notes in order to validate the data against the record, said Simon. Hernandez said that not all clinical data need to be adjudicated, because ultimately they may not matter if they are correct. For example, if errors in assessment or recording are random and not systematic, this random error should not affect the results. Joanne Waldstreicher noted that there is empirical literature about expert adjudication and when it makes a difference.

Safety Issues

In clinical trials, said Simon, adverse safety events can be detected because trial participants' baseline health status is measured before the trial begins, so any adverse events that occur after the exposure may be attributable to the intervention. For example, a participant's blood pressure will be measured at the beginning of a trial and then measured again after exposure to the intervention. However, in a real-world setting, "things are measured when they're measured," said Simon. In this scenario, it may be more difficult to differentiate between comorbidities (i.e., a health condition

that was preexisting) and adverse events (i.e., a health condition that was due to the intervention). This may make it more difficult to assess safety issues using RWD, said Simon.

Another safety issue raised by Simon was the issue of misclassification of data. Normally, random misclassification of data biases a study toward the null (i.e., random error may result in a finding of no effect). In a study examining effectiveness, the result of this bias would be a finding that the intervention had no effect on outcomes, he said. However, random misclassification may also result in missing safety events, which could "lead to conclusions that would damage the public's health or be unsafe."

DECISION AID

The general issues discussed by individual workshop participants in the first and second workshops were used to identify topics that could benefit from further exploration in the third workshop. Draft "decision aids" were developed by some individual workshop series participants on discrete aspects of study design to organize the topics that could benefit from further exploration and to facilitate deeper discussions at the workshop. A decision aid (like that presented at the workshop; see Figure 6-3), said session moderator Pall Jonsson of the National Institute for Health and Care Excellence in the United Kingdom, is "intended to lay out key questions for stakeholders to consider early on" in order to "make thoughtful choices around the development and design" of rigorous studies that use RWD. Many workshop participants reflected on the concepts highlighted in the decision aid over the course of their discussions, and some workshop participants offered direct feedback on the decision aid itself (see Box 6-1).

"FIT FOR PURPOSE" AND RELEVANCE OF DATA

At the third workshop, a panel of speakers representing different RWE stakeholders shared their perspectives and experiences using RWD. The speakers were asked to discuss how a decision aid such as the one in Figure 6-3 could help guide the use of RWD. When considering using RWD to answer a research question, said Daniel, the essential question is the following: Is the accuracy of the data good enough to reasonably and consistently identify the right population, the exposure or the intervention, and the right outcome? That is, are the data relevant and fit for purpose for the research question at hand? Several speakers discussed the processes they use to assess the relevance of RWD.

Researchers Using RWD Primarily from External Sources

Yaist said that when considering the use of RWD, it is imperative to start with the research question and the context of the decision. Only with this information in hand can one determine what data elements might answer the question and provide the information necessary to meet a certain need. Once the context of the decision is clarified, said Yaist, one can begin to identify possible data sources, and to evaluate these sources for relevance. Yaist said the first step is to see if there are already validated ways to get information; for example, major adverse cardiac events in claims data have been extensively studied. Next, the researcher would look to see where the needed data elements could be found—are there existing data sources, either from clinical care or from patients? Or do the data need to be collected? (See Figure 6-4 for this decision tree.) Once data sources are identified, the researcher would look at a number of factors to assess the relevance of the data for the research question:

- Availability of key data elements (e.g., exposure, outcome, and covariate variables);
- Representativeness of population;
- Sufficient number of subjects;
- Availability of complete exposure window;
- Longitudinality of data; and
- Availability of elements for patient linking.

Organizations with Existing Databases of Real-World Data

Altan presented a similar approach as Yaist for assessing the relevance of an existing source of RWD. OptumLabs is an aggregator of data that stores de-identified data for use by 30 partners that can access these data. The data at OptumLabs originate from a wide variety of sources, including data derived from EHRs, administrative claims, laboratory results, consumer data, and socioeconomic status (see Figure 6-5).

Altan outlined a simplified version of the decision process that she uses when an organization approaches OptumLabs with a research question that potentially can be answered with existing RWD (see Figure 6-6). The first step, said Altan, is asking if the required sample and variables exist. That is, are there data on the right population, the exposure and outcome under study, and covariates or confounders of interest? If there are data, is there a sufficient sample size to power the study? Can the available data be used for any special needs, for example, repeated measurements or measurements of exposure around a specific event?

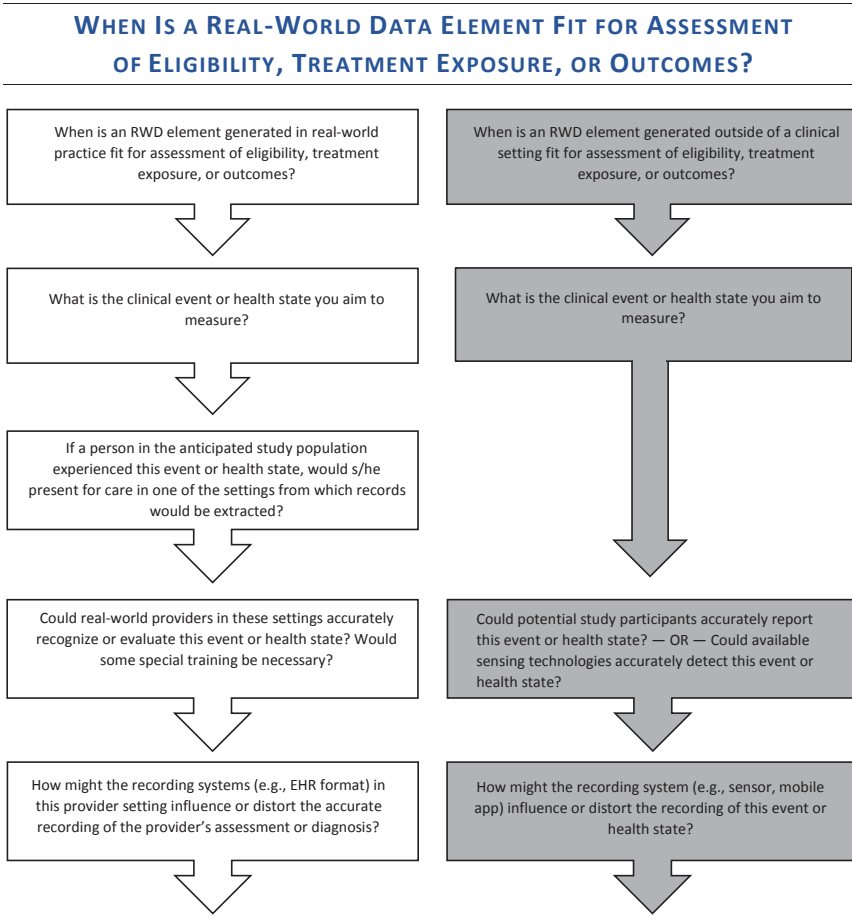


FIGURE 6-3 Real-world data (RWD) decision aid.
NOTES: This decision aid reflects questions about development and design for studies using RWD, and served to assist conversation and frame workshop discussions. The boxes with white backgrounds (on the left) show questions relevant to data generated within a health care system, such as electronic health record (EHR) or claims data. The boxes with the gray background (on the right) show questions relevant to data generated directly by patients and that would not necessarily be seen by a provider. This decision aid was drafted by some individual workshop participants based on the discussions of individual workshop participants at the first and second workshops in the real-world evidence series. The questions raised are those of the individual participants and do not necessarily represent the views of all workshop participants; the planning committee; or the National Academies of Sciences, Engineering, and Medicine, and the figure should not be construed as reflecting any group consensus.
SOURCE: Jonsson presentation [Session 2], July 17, 2018.

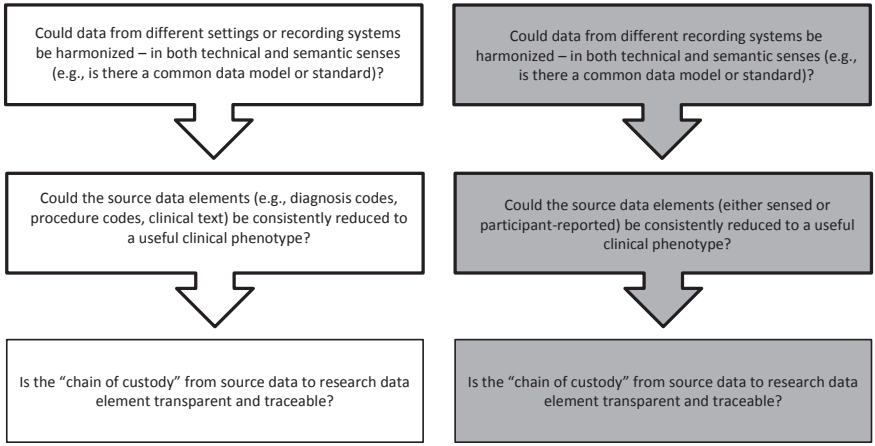


FIGURE 6-3 Continued

BOX 6-1
Feedback on the Decision Aid as Discussed by Individual Workshop Participants

Several workshop participants offered specific feedback on the decision aid "When Is a Real-World Data Element Fit for Assessment of Eligibility, Treatment Exposure, or Outcomes?":

- This decision aid is written from the perspective of the data generator, rather than the data user. There may be a need for another column to reflect the perspective of the data user. In addition, the decision aid is "too high level to be useful" for a regulator who is trying to assess the quality of a dataset. (Ball)
- This decision aid includes questions about data collected in the course of clinical care, as well as outside clinical practice. However, what is missing is a series of questions about existing, aggregated, observational data assets, such as the data used by OptumLabs. (Altan)
- The decision aid needs to include questions about "present bias" for patient-generated health data, and needs a question about the importance of data permission, whether for patient-generated data or data generated in the course of clinical care. (Foschini)
- With slight wording changes, the questions on the decision aid are the same questions asked for every study. However, depending on a huge number of factors, the answers are going to be different. Answering these questions in a binary fashion is very difficult because the answers will vary depending on the research question and the context of the decision. For example, if the decision to be made is about the monetary value of a drug, the necessary level of evidence is different than if the decision to be made is whether to initially approve a drug. (Weiss)

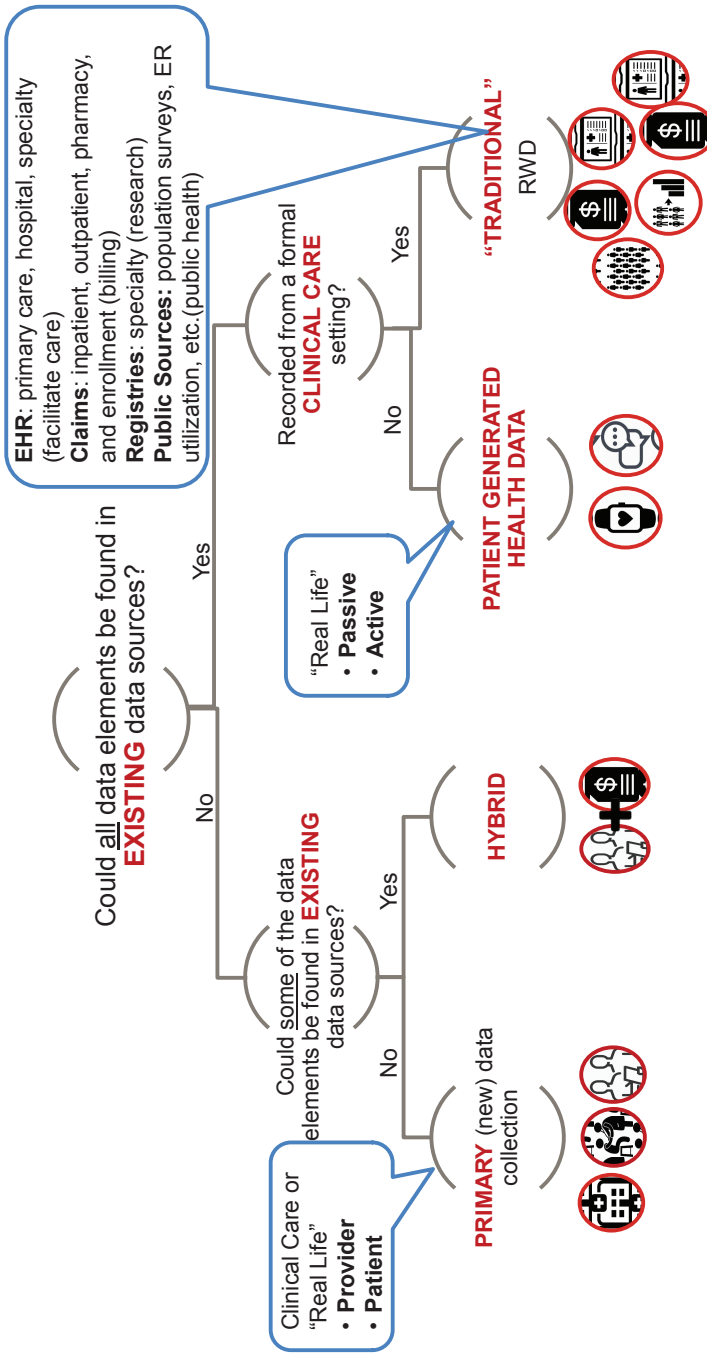


FIGURE 6-4 Decision tree to identify possible data sources.
NOTE: EHR = electronic health record; ER = emergency room; RWD = real-world data.
SOURCE: Yaist presentation, July 17, 2018.

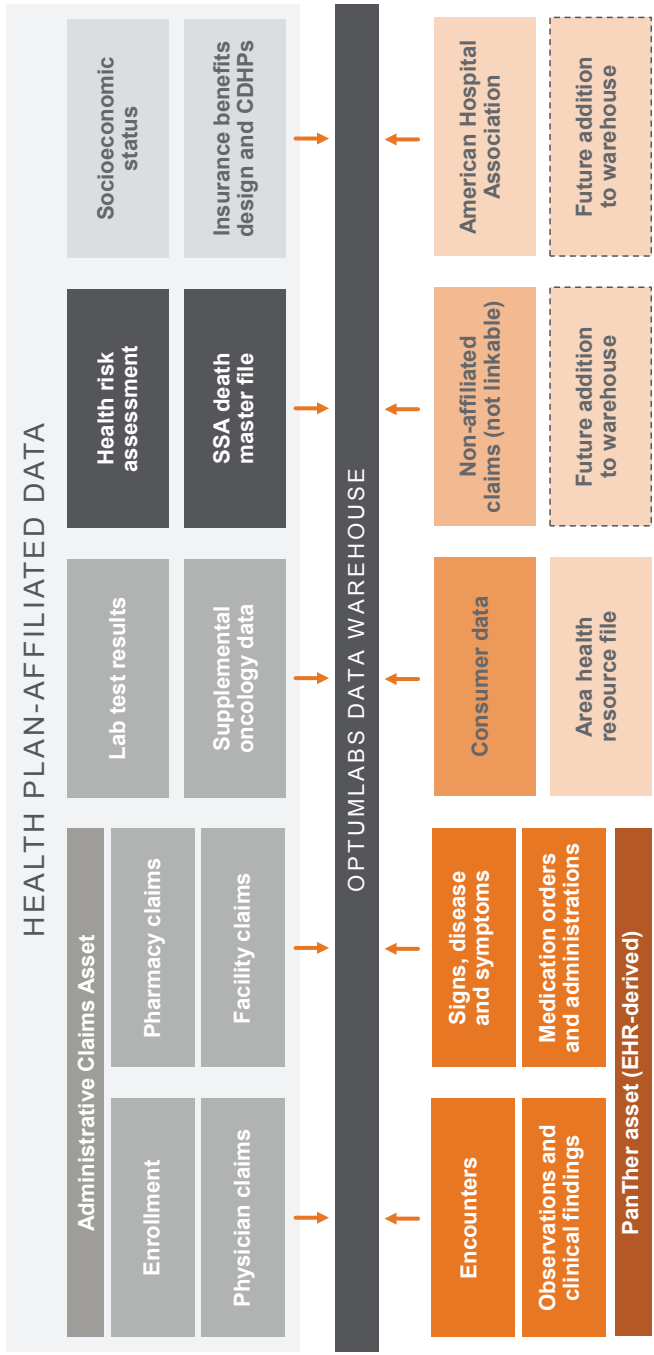


FIGURE 6-5 Linked, de-identified data sources at OptumLabs.
NOTE: CDHP = consumer directed health plan; EHR = electronic health record; SSA = Social Security Administration.
SOURCE: Altan presentation, July 17, 2018.

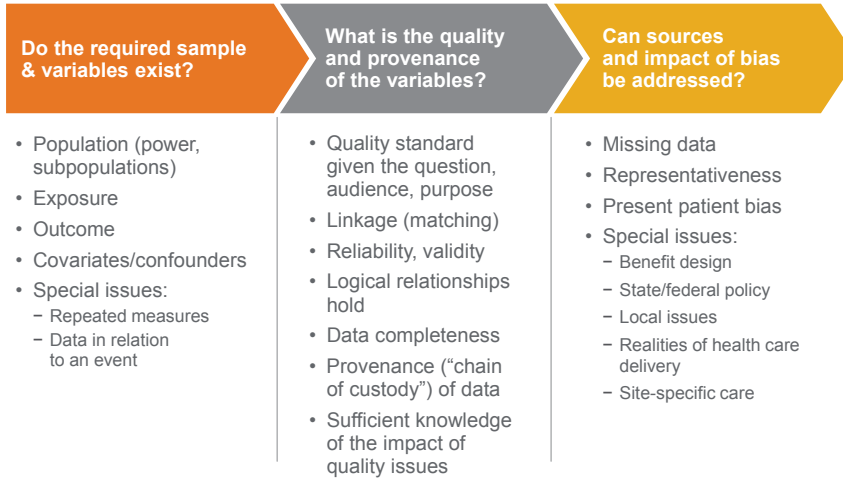


FIGURE 6-6 Decision process with an existing real-world data asset.
SOURCE: Altan presentation, July 17, 2018.

The second step, said Altan, is determining the quality and provenance of the data. Similar to the process outlined by Yaist, this requires considering the research question, the audience, and the proposed purpose for the data, and determining what quality standard would be required. For example, a stakeholder looking to use RWD for regulatory purposes requires the highest quality data, whereas a stakeholder seeking data for marketing research does not. Altan also looks for completeness of the data, the availability of linkages among data and whether they have been well assessed, and whether the provenance of the data can be established. Part of the data quality assessment process, said Altan, is examining whether the impact of any quality issues can be understood.

Finally, any sources of bias and their impact should be assessed, she said. Common sources of bias may include missing data, a lack of representativeness of the sample, and the "present patient bias," where the data only reflect patients who presented for care, but not those who did not seek care. In addition, there may be other biases such as differing policies or care practices. For example, a patient's insurance policy will affect what his or her claims data look like, said Altan. Two patients, one who has insurance with a capitated model and one who has pay-for-service coverage, are going to have very different claims data, even if their care was similar. These types of bias need to be understood and addressed in order for data to be relevant and useful.

Altan gave an example of how this process would work for a variety of data needs and research questions. One common issue, said Altan,

is missing data and leakage. To illustrate, Altan showed a table of data about hospitalizations for myocardial infarction and what medications were administered during the hospitalization. Some sources reported a very low percentage of patients receiving aspirin, which indicates to Altan that "something is wrong." These differences might reflect differences among the sources, for example, different EHR platforms or differences in the types of data that each source shares. Regardless of the source of the difference, a discrepancy like this indicates there are missing data in the dataset, and that the calculated average for aspirin use may not be accurate. Robert Ball, deputy director of the Office of Surveillance and Epidemiology at FDA's CDER, noted that even as RWD sources and the tools for using them improve, it is likely that there will continue to be issues with missing data. One way to address this issue, he said, might be through statistical approaches.

FDA Sufficiency Analyses

Ball shared details of his experience with "sufficiency analyses" in Sentinel, as required by the FDA Amendments Act of 2007, which have the same basic focus as the processes described by the other speakers. Sufficiency analyses, said Ball, ask the questions, "Are the data there for exposures, outcomes, and confounders? Are the methods and tools available? Can it be done with sufficient precision to answer the question of interest?" FDA has conducted these analyses for the use of RWD to study around 100 drug-adverse event outcomes, he said, using Sentinel's System of Active Risk Identification and Analysis (ARIA). ARIA contains data in a common data model, which includes the "most granular elements that are needed to make the assessments," clarified Ball. The curation and quality control of the data occur within each data partner, but the use of a centralized approach and centralized software help to ensure that it is an efficient and standardized process. Ball noted that because of the distributed model of Sentinel, it is important to have access to the people who know the data and the data systems. If a problem arises—such as missing or incomplete data—their knowledge of the data is critical.

PATIENT-GENERATED DATA

Luca Foschini, co-founder and chief data scientist of Evidation Health (Evidation), spoke about the current state and the future potential of patient-generated health data (PGHD). Foschini noted that currently, most RWD still comes from health system data, which is episodic and limited. PGHD, on the other hand, can be continuously collected, helping to illuminate patient experience between clinical visits. A key challenge with these

data, especially when passively collected from wearables and sensors, is that they generate huge volumes of data. For example, he said, a patient who has undergone a knee arthroscopy has perhaps dozens of data points in the claims data over the course of 1 year—the procedures undergone, the drugs taken, and the physical therapy conducted. On the other hand, the same knee arthroscopy can be measured by a device that logs steps to track patterns of physical activity, potentially generating hundreds of thousands of data points over the course of 1 year. At scale, these types of devices are being used to collect RWD for millions of patients, accumulating trillions of data points, he said. Processing this volume of data requires infrastructure designed to intake, clean, and normalize continuous data. In addition to its volume, the complex nature of the data requires analytical approaches such as artificial intelligence and machine learning, he said.

PGHD can come from a wide variety of sources, said Foschini, such as devices attached to a person's wrist or shoe, implantables, and smart devices in the home or car (Gambhir et al., 2018). In addition, there are PGHD available from social media sources, as well as data that patients report themselves such as surveys or diaries. Foschini said that because there are so many potential sources of PGHD, "the hope of having a common data model for [all the disparate kinds of] PGHD is doomed"; researchers should not wait for a common data model before beginning to use these sources, he said.

Foschini discussed the value of PGHD, and how it can be used to answer research questions. He gave examples using several types of PGHD and their value for research. Evidation conducted analysis for a study that used wearable activity trackers to study patients with multiple sclerosis. One finding of the study was that it took people with multiple sclerosis a significantly longer time to fall asleep at night (see Table 6-3). This outcome, said Foschini, had never before been measurable.

TABLE 6-3 Patient-Generated Health Data for a Study of Multiple Sclerosis (MS)

Activity Trackers Only	MS Trackers	Matched Control Trackers
N	498	1,400
Percent of Days with Tracked Steps*	73%	77%
Mean Daily Stepcount*	6,379	7,188
Mean Nightly Sleep Duration (Hours)	6.3	6.5
Max Time to Fall Asleep (Minutes)*	18.58	13.91

NOTE: * $p < 0.001$ corrected for false discovery rate.
SOURCES: Foschini presentation, July 17, 2018; Evidation and Novartis, 2018.

The second example that Foschini presented was using both retrospective passively collected data with data prospectively collected from patients. In this study, Evidation had access to several million participants who had consented to share their data for this analysis, said Foschini. Participants also completed a survey asking if they had had any major medical procedure or surgery, and for more details about it if they had. Evidation then analyzed the types of surgery against the participants' resting heart rate, among other variables, and was able to observe the effect of weight loss procedures on resting heart rate, which had not been possible to capture before in real-world settings (see Figure 6-7). This study, said Foschini, took less than 1 month to conduct. There is a real opportunity to use this model, in which participants are already sharing data, can be consented for a particular use of those data, and can be asked to provide additional information through survey questions, in order to quickly and easily collect fit-for-purpose RWD.

PGHD, said Foschini, can blur the line between RWD and data from clinical trials. For example, at the time of the workshop, Evidation was beginning a new study that ultimately enrolled 10,000 patients and collected prospective data on chronic pain through both devices and self-reported surveys. The data collected, said Foschini, will be an RWD source, but the study process is governed by a traditional clinical trial protocol. In fact, he worked with the Clinical Trials Transformation Initiative (CTTI) on its recently released guidance about how to use mobile devices for data

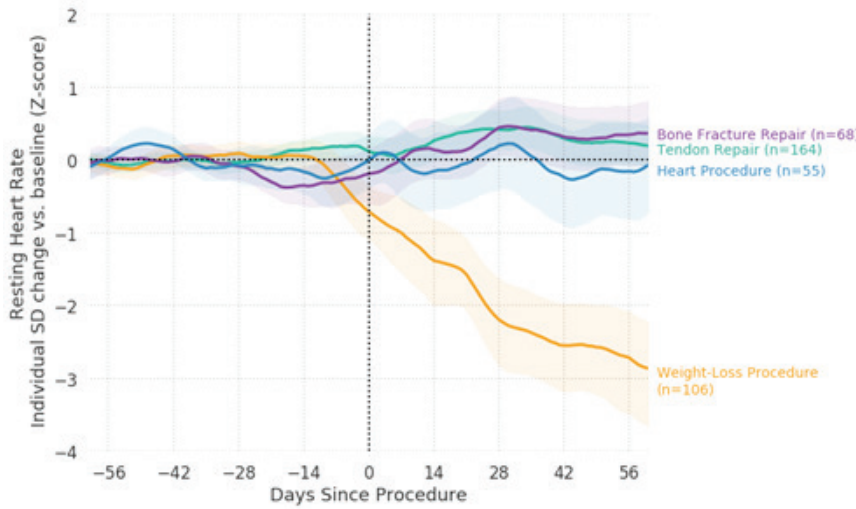


FIGURE 6-7 Effect of medical procedure or surgery on resting heart rate.
NOTE: SD = standard deviation.
SOURCE: Foschini presentation, July 17, 2018.

BOX 6-2
The Importance of Patient-Generated Data as Discussed by Individual Workshop Participants

The growing use of wearables and mobile technologies, David Martin, associate director for Real World Analytics, FDA, said, may enhance our ability to collect patient-generated prospective data, and there are opportunities to link these data with claims or electronic health record (EHR) data. Sebastian Schneeweiss, professor of medicine and epidemiology, Harvard Medical School and Brigham & Women's Hospital, emphasized the potential for these kinds of data links, saying that "there is a huge advantage of marrying those two worlds together." Richard Platt, professor and chair, Department of Population Medicine, and executive director, Harvard Pilgrim Health Care Institute, Harvard Medical School, said that EHR and claims data miss important data, such as whether a person can climb a flight of stairs, or the number of seizures experienced by a patient. These data—which could be provided by patients—are critically important to whether a treatment is worthwhile, but "almost uniformly missing" from EHR or claims data.

capture in clinical trials.[1] Foschini said the findings and recommendations of the CTTI report, as well as from the Duke-Margolis Center for Health Policy work on characterizing RWD quality and relevancy,[2] will be relevant for many sources and uses of PGHD. Other workshop participants discussed the potential of PGHD when linked with additional data sources (see Box 6-2).

"Traditional" RWD are collected in a clinical setting (e.g., EHRs or claims data), said Foschini. In many ways, the considerations about relevancy and quality of data are the same for PGHD and traditional RWD. Regardless of the source, data need to be available for key elements, need to be representative of the population, and need to be accurate and complete, he said. Some of the potential data issues with PGHD also exist in traditional RWD. For example, data collected in a clinical setting can have the "present patient" bias, in which only patients who present for care are represented in the data. A similar bias is possible in PGHD, said Foschini, because patients who are unwell may choose to wear (or not wear) the device, thus "censoring their outcome" in a way similar to patients who present for care.

[1] See https://www.ctti-clinicaltrials.org/programs/mobile-clinical-trials (accessed January 4, 2019).

[2] See https://healthpolicy.duke.edu/sites/default/files/atoms/files/characterizing_rwd.pdf (accessed January 4, 2019).

There are potential types of bias and confounders that are unique to PGHD, said Foschini, and those need to be addressed. For example, in traditional RWD, a provider inputs data into the EHR or claims system, and in general not much thought is given to how this user experiences the system. With PGHD, the end user is the patient or participant, and there may be biases based on the user experience, said Foschini.

Data from PGHD are potentially incredibly valuable and rich, said Simon, but they also may be quite chaotic. A challenge will be in monitoring and interpreting the data stream so that the data are useful rather than overwhelming.

DISCUSSION: REAL-WORLD DATA CONCERNS FOR FUTURE RESEARCH

During the second and third workshops, participants engaged in discussions that reflected on specific questions around the use of RWD (some of which were listed in the draft decision aid), as well as exploring related or new topics that emerged.

Real-World Data Compared with Data from Randomized Controlled Trials

While it is clearly important to ensure that RWD are accurate and reliable, said Robert Califf, it is also important to acknowledge that data from RCTs are not always accurate or reliable. RCTs are a relatively closed system, and the data they collect can be limited and skewed, he said. For example, treatments that are meant for long-term use in the real world cannot be fully assessed in short-term, controlled trials. For this reason, relying solely on RCTs and ignoring the data available in RWD could be "dangerous to the intended population," he said, and called for validation of both RCT data and RWD.

A workshop participant said that clinical trial endpoints do not always align with real-world endpoints, nor are they always relevant to patients. Hernandez responded that while some clinical trial endpoints are extremely relevant—for example, survival or staying out of the hospital—he agreed that there are different preferences over the journey of a patient, and that they should be considered going forward. In mental health practice in particular, said Simon, validated endpoints are quite often not relevant to patients' real-world experiences. For example, a patient does not care about a score on the Yale-Brown Obsessive Compulsive scale, favoring instead daily function and quality of life. In this area, said Simon, it is imperative to not treat these existing measures—which are often used for RCTs—as the gold standard against which to validate RWD. Daniel added that it is

important to make sure that endpoints are relevant to the ultimate decision maker, whether that is the patient, payer, or provider. Daniel said the endpoints that are studied in clinical trials may not necessarily need to be altered, but suggested instead focusing on developing an "evidence package" that includes trials, RWE, and other sources of information to get a well-rounded picture.

In considering the relevance and quality of a data source, said Yaist, one of the main considerations is whether the data are representative of the population of interest. Data that are to be used for a regulatory decision need to reflect the U.S. population, she said. Unfortunately, there are few data sources, including EHRs and claims, that truly reflect the population. When using these large data sources, she said, researchers need to "really think about systematic bias" that may be present. For example, researchers should consider the clinical context of the data source and how certain patients may be represented while others are not.

One potential way to minimize bias from RWD, said Marc Berger, former vice president of Real World Data and Analytics at Pfizer, is to use multiple datasets. "You should never trust any one study," said Berger. Using multiple datasets increases the likelihood that bias will be minimized, because each dataset may have different biases that cancel each other out (see Box 6-3).

BOX 6-3
Potential Sources of Bias in Real-World Data as Discussed by Individual Workshop Participants

Individual workshop participants identified a number of potential biases in real-world data during the discussion, including

- Present patient bias. (Altan)
- Reimbursement policies that encourage patients or providers to choose a specific diagnosis or treatment. (Graff)
- Policies or practices that differ among locations or systems. (Altan)
- User experience for patient-generated data. (Foschini)
- Patient decisions about when to wear a monitoring device. (Foschini)
- Different electronic health record systems that facilitate or inhibit certain types of data collection. (Hernandez)
- Missing or incomplete data. (Altan, Simon)
- Incomplete ascertainment of stigmatized outcomes because patients present for care outside their primary provider or network. (Simon)

Dynamic Data

Rima Izem, senior mathematical statistician at CDER, said one of the unique characteristics of RWD is that they are dynamic—the data change over time, which changes the answers that can be derived from the data. Many studies are predicated on the assumption that the data are static, and algorithms are validated based on the data at one point in time, she said. But with RWD, prospective data are always changing, and even retrospective data can change because data from other sources may be added to the dataset. Altan shared an example of the dynamism of RWD. Researchers were using data from claims and EHRs in a study, and during the course of the study, the provider of the EHR data had a change in their client base that caused data to be removed from the environment and changed the sample size.

Another way in which RWD can change over time, said Altan, is through changes to practice patterns for particular diseases. In a longitudinal study that may last multiple decades, the data about clinical care and clinical outcomes can change substantially as there are regulatory changes, formulary changes, or changes to standard of care. Simon added that coding can also change over time—for example, during a study on suicide prevention, the *International Statistical Classification of Diseases and Related Health Problems* (ICD) codes for self-harm changed dramatically between ICD-9 and ICD-10. As a result, the researchers had to adjust the outcome specifications during the trial, he said.

Berlin stressed that because of this dynamism of data, it may be necessary to reassess the tools used to analyze data—for example, the algorithms that are validated for a certain condition may need to be revalidated after changes have occurred. Most importantly, he said, the data tools and methods need to be transparent, so that all stakeholders can understand how the data have been collected, curated, and analyzed.

The Importance of Sharing

Deven McGraw, deputy director for health information privacy at the U.S. Department of Health and Human Services' Office for Civil Rights at the time of the workshop and currently chief regulatory officer at Ciitizen, observed that nearly any use of RWE and RWD requires some amount of data sharing. Currently, the rules and regulations about data make it "perfectly permissible . . . for you to do nothing but sit on the data that you are collecting," she said. There are perverse incentives that discourage sharing, she said, because when data are shared, there are rules and conditions that must be followed. The hurdles to sharing are even higher for data that are related to sensitive conditions such as mental health or substance abuse, she said. Clearly, patient privacy and security should be

protected. However, the current system tilts the scale so far in the direction of privacy that data are "buried in the backyard" and are not shared and used to their full potential. The regulatory environment, said McGraw, has unintentionally created disincentives for sharing, and it is time to "rethink our regulatory framework and take the thumb off of one side of the scale and try to put it on the other side of the scale." However, allowing any and all kinds of sharing would not necessarily be beneficial, she said, and a framework of accepted principles could be essential for reducing risks and increasing benefits.

Transparency of Data Source and Curation

Berger said there is a lack of transparency around RWD and how they are curated. Transparency would allow researchers to check the original source document against the curated data in order to evaluate the quality of the curation, he said. Grazyna Lieberman, director of regulatory policy at Genentech, added that transparency enables assessment of the completeness of data collection, the accuracy of captured diagnoses, and the reliability of the processes to extract the data.

Another area in which there is a need for transparency, said Simon, is in how EHRs assign ICD codes. Simon said that in mental health, providers often do not type in a specific ICD code; rather, they type a text string that results in suggestions of ICD codes. The algorithms that map the text strings to the suggestions are proprietary, said Simon. Unfortunately, for RWD to be reliable, a researcher may need to know not just the ICD code, but also the text string that prompted the code. Transparency about the transformation from the source data to the analytic dataset is essential, he said, particularly in cases where the diagnosis is less straightforward (e.g., different types of depression versus myocardial infarction). One particular benefit of transparency is that it allows researchers to understand and modify code if the researcher wants a tighter or looser definition of a population, exposure, or outcome, said Simon.

There is a need for publicly available, validated, generally accepted algorithms for identifying core clinical phenomena, said Simon. The process of developing and validating these algorithms should be completely transparent, he said. Simon said this is a "higher level of transparency than we are accustomed to," but it is critical because if a validated algorithm does not work, a researcher needs to be able to look at the building blocks of the algorithm to see where things went wrong. Without transparency, "the assertion that this is a valid phenotype or this is a valid specification will not go far; you need to show your work," he said.

Context of the Decision

Yaist and Altan both emphasized the importance of considering the type of question that RWD are being used to answer. The research question clearly matters when determining the relevance of the data, but just as important is the context of the decision to be made. For example, said Yaist, if there is already an established safety and efficacy profile for a drug, that is a far different scenario than if RWD are being used to answer initial questions about safety and efficacy. The quality of the data that are needed for these two scenarios is very different, she said. Richard Platt added that there may be different standards for data quality when the data are being used to assess the superiority of a drug versus non-inferiority.

Participants discussed the fact that stakeholders all have a shared interest in the quality of the data, and they all have the same basic goal of improving health care. However, Simon noted, the specific data needs of stakeholders may vary considerably. For example, stakeholders trying to define and measure "myocardial infarction" may draw different boundaries around this outcome depending on whether they are clinicians, payers, or regulators. Some may want higher specificity, but some may need higher sensitivity, he said. There is a shared interest in the quality of the data, but the context of the decision to be made is critical.

Need for Systematic Processes

Ball said that FDA assesses a large volume of RWD studies, so "there has to be a very systematic and efficient process for quality assessment." He said that while guiding questions such as the ones in the decision aid (see Figure 6-3) are useful and cover many of the key topics, FDA would need the questions to be systematized into an industrialized process.

Ball noted that the current system for assessing and using RWD is quite resource intensive. For example, he said, to validate algorithms used to identify cases of anaphylaxis, subject-matter experts currently manually review medical records to classify cases as to whether they are anaphylaxis. Then, another set of experts identifies the codes to combine in an algorithm and the algorithm is assessed as to how well it can identify the expert-classified cases of anaphylaxis. This process is "slow and inefficient" and costly. Alternatives that are in development include tools such as natural language processing and machine learning technologies, he said. Simon followed up on this point with the observation that creating an industrialized process for quality assessment would not only be more efficient and cost-effective, but would also help to create a culture of quality. This culture, said Altan, could involve data aggregators rethinking or better documenting the process of data collection, cleaning, and curating.

7

How Tightly Should Investigators Attempt to Control or Restrict Treatment Quality in a Pragmatic or Real-World Trial?

Key Messages Identified by Individual Workshop Participants

- Inclusion and exclusion criteria can be broadened in real-world trials so patients with comorbidities or concomitant treatments are included. These patients may require additional monitoring, and the trials will likely yield more generalizable results. (Califf, Horberg, Katz, London)
- When considering treatment restriction in real-world trials, researchers may consider a specific set of questions or categories for consideration in order to plan a trial that answers the research question and honors research participant safety and autonomy. (Alphs, Stein)
- Research driven by patients is iterative and considers patient needs and priorities, as well as patient experience. The focus of this type of research is to enable patients to make informed decisions about their health care. (Nowell)
- Researchers have two duties: maintaining the trial protocol and caring for the well-being of their patients. If these two duties conflict, caring for patients is the higher priority. (Alphs, Katz, Simon)
- Designing a study requires a decision about how to define the standard of care, which can be highly variable in different regions, for the control arm. This decision can affect the results of the study, and there is an ethical obligation to ensure that

> study participants are not receiving substandard care. (Califf, Hernandez, Stein)

In a real-world trial, there is a delicate balance between the needs of the trial and the needs of the patient, said Larry Alphs, deputy chief medical officer of Newron Pharmaceuticals. To make inferences from the study, researchers need to ensure a certain level of consistent treatment adherence. However, researchers must also ensure that patients receive adequate treatment for their condition, and that there is enough flexibility to handle unique patient needs and adverse events that arise. In this session, workshop participants explored how variability in treatment might affect results, and how to balance the needs of the study with patient autonomy and safety. Gregory Simon noted, "We're talking about real-world treatment, and by that we mean studying treatments with typical providers and typical patients, accepting that there will be highly variable quality of, and adherence to, treatment."

ILLUSTRATIVE EXAMPLES

To explore the issues surrounding treatment quality in real-world trials, speakers at the second and third workshops presented case studies as illustrative examples of the considerations that go into treatment quality control and restriction.

Lithium for Suicide Prevention

For more than two decades, lithium has been proposed as a treatment to prevent suicide in patients with bipolar disorder and major depression, said Ira Katz, senior consultant for program evaluation at the U.S. Department of Veterans Affairs' (VA's) Office of Mental Health and Suicide Prevention. This idea has been supported by evidence from meta-analyses of observational studies and randomized controlled trials (RCTs) that were conducted for other purposes, although a propensity score-matched study of bipolar patients in the VA was equivocal. There had never been an adequately powered clinical trial on lithium, he said, and two previous trials had been terminated due to difficulties in enrollment. Seeing this need, Katz and his colleagues proposed the idea of carrying out a large RCT within the VA to test lithium as a treatment for preventing suicide. The VA, said Katz, was an ideal site for this research because there is a large patient population, approximately 140 medical centers, infrastructure for suicide prevention, and infrastructure for clinical trials.

The study, which is double blinded and placebo controlled, had enrolled around 360 people at the time of the workshop. These people had depres-

sion or bipolar disorder and had survived a suicide attempt. The projected sample size is around 1,600, and it is powered to detect about a 40 percent reduction rate of repeated suicide attempts, said Katz. There have been some problems with recruitment, said Katz, but the biggest issues have been questions about ecological validity (i.e., the extent to which the findings from the clinical trial can be generalized to treatments for patients in real-world settings). Simon noted that this type of study will become necessary with the development of a "new generation" of treatments for suicide prevention. "Our current way of evaluating treatments in mental health will be completely incapable of assessing that question," he said.

Katz walked the workshop participants through a series of questions that the researchers needed to answer, and the considerations that went into answering the questions.

Does the Selection of Participants for RCTs Lower the Outcome Rate?

To conduct a study evaluating interventions to prevent suicidal behavior with reasonable sample sizes, Katz emphasized the need to enroll people who are at an increased risk for suicide. However, the process of enrolling in and consenting to an RCT makes it likely that people who are truly at risk will be "filtered out," said Katz. Data from electronic medical records suggested that among those with depression or bipolar disorder who survive a suicide attempt, 15 percent reattempt during the year, he said. Researchers expected to see a far higher number during the course of the clinical trial because they assumed that many suicide attempts that were not documented in the medical record would be captured with the assessments included in the RCT protocol. However, experience in the study has been consistent with the 15 percent rate, said Katz, which was "really surprising and a puzzle." After examination of other data, the researchers determined that what was likely happening was a counterbalancing effect: The suicide attempt rate was probably decreased due to filtering out from the RCT process, and at the same time, it was probably increased due to the RCT protocol that uncovered suicide attempts.

How Rigorous Should the Study Be About Diagnoses?

A second issue that the researchers faced, said Katz, was determining the inclusion and exclusion criteria for participants. Most RCTs want "clean patients," that is, those with the diagnosis but without comorbidities. However, requiring "clean patients" could exclude those at highest risk for suicidal behavior, said Katz. The researchers decided to permit comorbidities, such as posttraumatic stress disorder (PTSD) or substance abuse, and there was no attempt to filter out primary versus secondary

diagnoses of depression. Researchers also struggled with concomitant medications. Some medications, such as diuretics and angiotensin converting enzyme (ACE) inhibitors, make it more difficult to manage patients on lithium. The researchers proposed performing extra monitoring for these patients, but the Institutional Review Board (IRB) decided they could not be enrolled. After the research began, they discovered that 30 percent of otherwise eligible patients were excluded specifically because they were on ACE inhibitors. Katz and his colleagues successfully argued to the IRB that these patients had to be included if the trial participants were going to be a representative sample of the true patient population.

How Can Researchers Balance the Flexibility of Caring for Patients with the Need for Adherence to RCT Protocol?

Third, said Katz, there was the issue of finding a balance between providing appropriate care for patients and adhering closely to the RCT protocol. "Flexibility is essential but difficult," said Katz. He relayed the story of a patient who had depression, PTSD, and a personality disorder, and was difficult to manage. The patient was frequently late for appointments or missed them entirely. In the sixth month of the study, the patient missed the required blood test and assessments. The investigators, however, continued to send the patient supplies of the study medication while encouraging the patient to come in for study assessments and care. The investigators believed this to be good clinical management for the patient, though they recognized that it may not have been consistent with good clinical practice for research. The study monitor criticized this decision and judged it to be a serious protocol violation. The question in designing clinical trials relevant to patients and complex issues, said Katz, is whether the protocol should fit the patient, or whether the patient should fit the protocol. This is a difficult dilemma to navigate, particularly in clinically difficult situations like suicide prevention, said Katz.

INTERSEPT and PRIDE: Real-World Mental Health Trials

Developing real-world evidence (RWE), Alphs said, is an iterative process that evolves over time. At the third workshop, Alphs explained that RWE is aimed at answering the basic questions, "Is this drug safe?" and "Is this drug effective?" As the body of information grows, researchers are able to focus their questions on specific populations that may be complex. The ultimate question is whether a drug is safe or effective for specific individuals within the broader population for which the drug has been shown to be safe and effective. However, it is unlikely there will ever be sufficient RWE to answer these questions at the individual patient level.

In real-world pragmatic clinical trials, said Alphs, patient treatment may need to be restricted to patients without many complicating medical problems for a period of time in order to generate answers about a drug's safety and efficacy. When considering treatment restriction, a number of general questions need to be asked:

- Which patients' treatments will be restricted from inclusion in the trial? What are their vulnerabilities that lead to these restrictions?
- What specific restrictions will be placed on treatment? What is the impact of these treatment restrictions?
- In which treatment settings will restrictions be applied? Are treatment practice and ethical considerations similar in all areas?
- How long are the restrictions to be in place? Will the treatment restrictions have enduring impact on morbidity and mortality?
- What is the value of the restrictions? What are the risk–benefit considerations of imposing these restrictions?

To answer these questions, Alphs and his colleagues developed a template that identifies the specific considerations when designing clinical trials, which are divided into six domains (see Table 7-1).

Alphs presented two case studies to demonstrate how the issue of treatment restriction has been dealt with in real-world trials.

InterSePT: Clozapine Versus Olanzapine for Suicide Prevention

The first case study was an international trial that compared the use of clozapine and olanzapine for suicide prevention in patients with schizophrenia or schizoaffective disorder. Patients with these disorders, said Alphs, are at high risk of suicide behavior, with a lifetime risk of suicide attempts at 25 to 50 percent and a lifetime risk of death by suicide at 5 percent. This is an undertreated life-threatening mental health condition, and represents a major public health problem, said Alphs. There is stigma surrounding both schizophrenia and suicide, making them difficult problems to address.

The trial, called InterSePT (International Suicide Prevention Trial), was a 2-year, multicenter, randomized, open-label, rater-blinded study, said Alphs. The trial enrolled 980 patients at high risk for suicide. They were randomized to receive either clozapine or olanzapine and were followed for 2 years. At the time of the study, olanzapine was the leading state-of-the-art treatment for schizophrenia, although the standard of care has changed somewhat since the study was completed, noted Alphs. Several endpoints were assessed: significant worsening of suicidality, hospitalization to prevent suicide attempt, suicide attempt, and death by suicide. Blinded raters made these assessments and a blinded endpoint monitoring board judged whether or not an event had

TABLE 7-1 Considerations for Study Design Restrictions

Domain	Definition of Domain Terminology
Participant eligibility criteria	Considerations include the intended treatment population of interest as identified by the study's authors.
Intervention flexibility	Considerations include posology, dose, dosing interval, windows allowed for dosing; permitted concomitant treatments. The domain should be considered separately for experimental and comparisons treatment interventions.
Medical practice setting/ practitioner expertise	Considerations include experience, skills, and resources of the practitioner and the treatment team; the health care delivery system; standards of care at the site; and local cultural practices that may influence medical delivery or outcomes. The domain should be considered separately for experimental and comparisons treatment interventions.
Follow-up intensity and duration	Considerations include frequency and length of visits and the number and the scope of the assessments.
Outcome(s)	Considerations include evaluation of measure(s) by which the interventions' effects are assessed and how well they reflect outcomes that are used and considered important to real-world practice.
Participant adherence	Considerations include the degree to which the subjects are encouraged and tracked for adherence to study-related procedures.

SOURCES: Alphs presentation, July 17, 2018; Alphs and Bossie, 2016.

occurred. Alphs said that at the time of the study, scales to measure suicidality as a clinical trial endpoint did not exist, so the research team developed new scales to use for regularly monitoring suicidality.

The trial was not blinded, said Alphs, for several reasons. First, the side effect profiles of the two drugs are dramatically different, so it would have been very difficult to keep patients from knowing which treatment they were taking. Second, clozapine causes agranulocytosis in about 1 percent of the population, and as a consequence, requires regular blood draws that were not required for olanzapine treatment and, thus, unnecessary for patients randomized to that treatment arm. Finally, the study was not blinded because it would have been unethical to do so, said Alphs. The enrolled patients were at a high risk of suicide, having either been hospitalized for suicide ideation or having attempted suicide in the past year. Because the potential outcomes were so severe, and ethical considerations required that the research design minimize suicide attempts and deaths, the patients' clinicians needed the knowledge of treatment to flexibly manage their patients should they become suicidal.

Participants in the clozapine arm of the trial required regular visits for blood draws. Alphs noted these visits could have impacted the patients' suicidality, so all of the participants were required to come in for visits with the clinical staff on the same schedule, regardless of treatment arm. He said this requirement did not reflect the reality of olanzapine treatment in the real world, but it was "an absolute requirement for the safety of the study." At each of these visits, the participants' suicidality was assessed. Alphs noted that this assessment is "good clinical practice that should probably be done in every case," including real-world clinical practice. If the assessment found that the patient was highly suicidal, the patient was hospitalized to prevent a suicide attempt. If the patient's suicidality had worsened significantly, this was also considered an endpoint for the study.

The results of the study, said Alphs, indicated that patients treated with clozapine were less likely to exhibit suicidal behavior or be judged at imminent risk for suicide than patients treated with olanzapine (see Figure 7-1). Significantly fewer patients in the clozapine treatment arm attempted suicide, required hospitalization, required treatment with concomitant drugs to prevent suicide, or died by suicide. Alphs noted that in both treatment groups, the extensive study surveillance and regular clinical assessment likely prevented suicide in many of these high-risk patients. This study contributed to a U.S. Food and Drug Administration (FDA) decision to

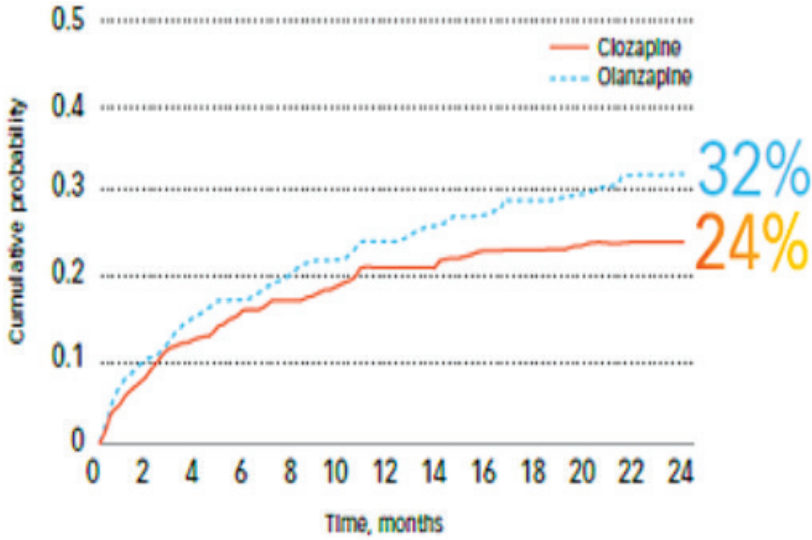


FIGURE 7-1 Cumulative probability of experiencing a significant suicide attempt or hospitalization to prevent suicide.
SOURCE: Alphs presentation, July 17, 2018.

approve clozapine for reducing the risk of suicide in high-risk schizophrenic or schizoaffective patients, said Alphs.

PRIDE: Oral Versus Injectable Antipsychotic for Treatment of Schizophrenia

The second case study that Alphs presented was a trial called PRIDE (Alphs et al., 2015). This study sought to determine if treatment with a long-acting injectable antipsychotic had advantages over oral antipsychotic treatments when provided to recently incarcerated persons with schizophrenia. Alphs said that "deinstitutionalization of the mentally ill over the past 50 years and changes in health policy have shifted the burden of care for mental illness [from mental hospitals] to jails and prisons." Many mentally ill people in the United States are incarcerated, and the risk of reincarceration is high when people are not given access to treatment after leaving jail or prison. This study examined not only the potential clinical benefits of an injectable antipsychotic (see also Box 7-1), but also the potential economic benefits that could be gained if patients are adequately treated for mental illness.

The trial was a 15-month multicenter, randomized, open-label, rater-blinded study. Participants were randomized to either an injectable antipsychotic (paliperidone palmitate), or to oral antipsychotic treatment (one of seven frequently used oral antipsychotics). Participants could de-select the medications in this group if the medications were considered by the participant or his/her treating physician to be ineffective for them, said Alphs. After de-selection of unacceptable candidate oral treatments, patients randomized to oral treatment were further randomly assigned to one of the remaining acceptable oral antipsychotics. The trial was not blinded because it was known that all of the drugs were relatively safe and effective, and the difference between an injectable and an oral drug would have been impractical to blind. The endpoints, all of which were considered "treatment failure," were time to hospitalization, time to suicide, time to arrest or reincarceration, and time to an intervention to prevent hospitalization or arrest.

The study found that treatment failure was 1.4 times more likely to occur during oral antipsychotic treatment than with injectable antipsychotic treatment, said Alphs. The mean days to treatment failure were nearly 6 months more for patients who received the injectable antipsychotic treatment. Using results from this study, the researchers applied economic modeling to stable schizophrenic Medicaid patients with similar clinical characteristics to predict health economic outcomes. This model estimated that, using the injectable treatment among all Medicaid patients with similar characteristics to those in the PRIDE study, more than $3 billion could be saved over an 18-month period, Alphs said.

BOX 7-1
Patient Compliance as Discussed by Individual Workshop Participants

Gregory Simon, senior investigator, Kaiser Permanente Washington Health Research Institute, asked Larry Alphs, deputy chief medical officer of Newron Pharmaceuticals, about how decisions about patient care were made for the PRIDE study on oral versus injectable antipsychotics. Alphs started by noting that in order to benefit from oral antipsychotics, the patient must take the pill daily. However, patients often have difficulty taking the medication every day, and this subpopulation (recently incarcerated persons) is particularly susceptible to noncompliance. The researchers "consciously did not intervene in any way in terms of the compliance" because this factor was part of the study question, he said. While the researchers were monitoring safety in general and encouraged adherence to the prescribed treatment in both treatment arms, no special "compliance checks" were made to ensure patients were taking their medication. Simon noted that in an optimal environment, oral medication would be delivered with "adequate psychosocial support, with close monitoring of adherence, and likely with a good, high-quality intensive case management program for people who have been hospitalized for exacerbation of psychotic disorder." However, he said, this is not "the world we live in." Poor compliance is responsible for much of the bad health in this country, said Robert Temple, deputy director for clinical science at the U.S. Food and Drug Administration's (FDA's) Center for Drug Evaluation and Research (CDER). While studying the safety and efficacy of new drugs is important, "a large part of our misery is due to *not* taking the drugs that work." Temple concluded that promoting or facilitating compliance was "probably the most important thing the health care system could do."

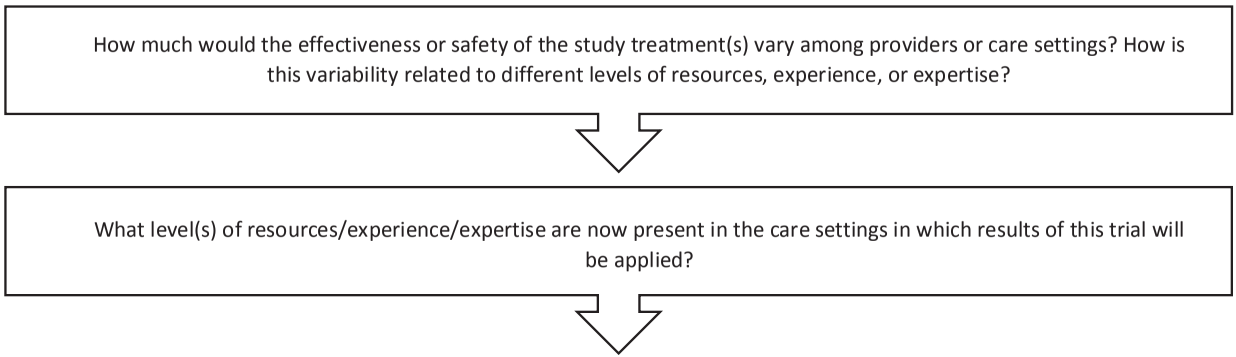
DECISION AID

The general issues discussed by individual workshop participants in the first and second workshops were used to develop a decision aid for the third workshop (see Figure 7-2). As with the other decision aids, the intention was to outline some questions to consider in order to make thoughtful choices in RWE study design. Session moderator Jennifer Graff, vice president of comparative effectiveness research at the National Pharmaceutical Council, encouraged workshop participants to consider how much variation could be desired or accepted in terms of treatment, setting, or provider, and what elements of trial participant safety and autonomy could be most important. She asked, "What is our obligation to deliver safe care if we are watching what happens in the real world? What do we do and what are our obligations to deliver state-of-the-art care to patients enrolled in pragmatic trials?" Participants at the third workshop reflected on these questions and offered feedback on the decision aid throughout the course of their discussions.

How Tightly Should Investigators Attempt to Control or Restrict Treatment Quality in a Pragmatic or "Real-World" Trial?

Starting assumptions for applying this decision aid

- The study question is clearly defined, including the decision to be made and decision maker the study should inform.
- Data are of adequate quality to assess eligibility, key prognostic factors, treatment exposure, and outcomes.
- Treatments are assigned randomly or by some other method that supports valid inference.



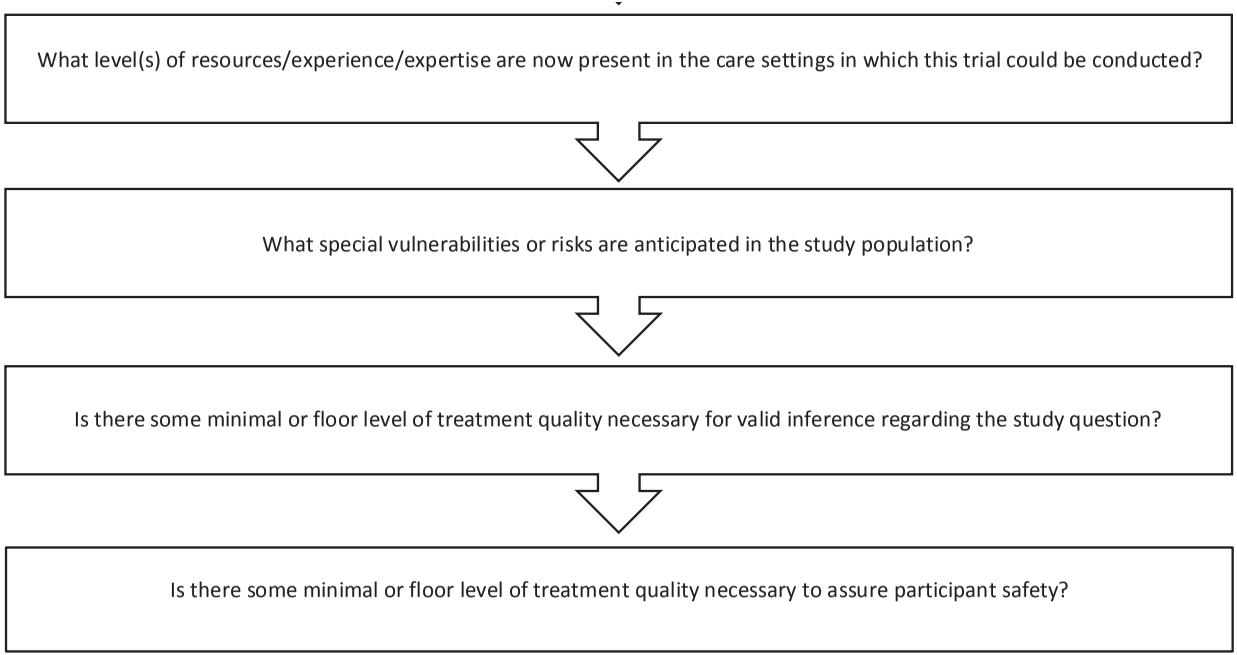


FIGURE 7-2 Decision aid on questions to consider regarding participant safety and investigator control of treatments in a trial taking place in a community care setting.
NOTE: This decision aid was drafted by some individual workshop participants based on the discussions of individual workshop participants at the first and second workshops in the real-world evidence series. The questions raised are those of the individual participants and do not necessarily represent the views of all workshop participants; the planning committee; or the National Academies of Sciences, Engineering, and Medicine, and the figure should not be construed as reflecting any group consensus.
SOURCE: Graff presentation, July 17, 2018.

PATIENT-CENTERED RESEARCH

W. Benjamin Nowell, director of patient-centered research at the Global Healthy Living Foundation, spoke about his experiences developing a patient-powered research network (PPRN) for arthritis patients. The PPRN—called ArthritisPower—is a research registry with more than 16,000 patient participants who have rheumatic and musculoskeletal diseases. It was created in 2015 with support from the Patient-Centered Outcomes Research Institute, and is 1 of 33 networks within PCORnet, the National Patient-Centered Clinical Research Network. ArthritisPower has a smartphone app that is used to collect patient-reported outcomes (PROs), said Nowell. The app has a number of features: Patients can input and track their own symptoms and treatments, run analytics on their own data, send reports to providers, connect to other patients, and learn about research opportunities (see Box 7-2). Nowell noted that the app has evolved as patients have provided feedback. For example, the app initially only allowed patients to enter information about arthritis-specific drugs, but patients wanted to be able to enter information about all of their medications.

The fundamental assumption that drives ArthritisPower as a platform for patient-centered research, said Nowell, is that "it enables patients to make a decision about their health care." In order to enable patients to make good decisions, facilitating access to relevant evidence and choosing

BOX 7-2
Features of ArthritisPower Research Registry, with Smartphone Application and User-Friendly Interface

- **Tracking**—symptoms; active/past medications; complementary treatments;[a] add other measures (flares,[a] depression, disease impact on social satisfaction); personal symptom note entries for context;[a] import VectraDA® lab results;[a] future biosensor innovation
- **Analytics**—longitudinal results graphing; overlay medication usage to see how symptoms change with new medications[a]
- **Share**—electronic reports can be sent to doctors, caregivers, and others
- **Research Opportunities**—browse available studies for participation
- **Connect**—invite other patients or caregivers to connect through in-app messaging with option to share notes, analytics[a]
- **Education**—CreakyJoints social media feed, including patient blogs/Twitter; disease-specific content on treatments, coping, and support[a]

[a] New features developed for version 2.
SOURCE: Nowell presentation, July 17, 2018.

study designs that are best suited to generating that evidence are needed, he said. Most important, he said, is determining which study designs and data sources will "permit us to answer the research question *and* engage our partners." The patient-driven research process is iterative, said Nowell, and requires ongoing consideration of patient needs and priorities, the patient experience in the study, transparency, and consent.

Nowell gave an example of a study in which ArthritisPower is currently involved. The CHOICE (Comparative Health Outcomes in Immune-mediated diseases Collaborative) study, said Nowell, is a PCORnet demonstration project that involves multiple networks, including ArthritisPower and other PPRNs. The study aims to evaluate the comparative risks for infection, heart attack, and stroke, and to evaluate the comparative clinical effectiveness of various medications using PROs. Evaluating the effectiveness of medication, said Nowell, is of utmost importance to patients and providers—"Patients and doctors want to know what treatment works best for whom and under what circumstances." Providers and patients need to make challenging decisions about treatment options, he said, which can be difficult to do when there are limited data. While most approved drugs work reasonably well for most patients most of the time, it is generally unknown exactly how well or how quickly the treatments work depending on the characteristics and preferences of a particular patient; this kind of evidence can be generated through PROs. Ultimately, he said, a patient is the only one who can determine how well a treatment is working to improve his or her quality of life. Data from multiple individual patients can be turned into information that physicians and other patients can use to make decisions about treatments, said Nowell.

Nowell emphasized that engaging patients is more than a "one and done" conversation, and that patients need to be engaged in different ways at different times throughout the process. In addition, different types of patients with different perspectives need to be engaged, he said. Although some patients are very familiar with the terms and concepts of clinical research, other patients need help understanding how research works and how it can impact them. In response to Nowell's presentation, Graff suggested that the patient's perspective be somehow captured in the decision aid (see Figure 7-2).

CONTEXT OF THE DECISION

As with any decision about how to design a study, said Peter Stein, deputy director of the Office of New Drugs at FDA's Center for Drug Evaluation and Research, the first consideration is the context of the decision to be made: What is the research question? What is the intended use of the evidence that is generated? What is the level of evidence that is needed to

make the decision? Stein said the answers to these questions can help guide decisions about patient treatment within a real-world study. For example, if the study is aimed at establishing efficacy for a new intervention, the researchers may want to tightly control patient treatment to find the most precise estimate of efficacy. However, if the study is about an intervention that is already known to be safe and efficacious, and the research question is broader, a study design that allows for flexibility in patient treatment may be more appropriate. For example, if the research question is how an intervention works in an expanded population in the real world, a trial in which patient treatment was tightly controlled might not generate generalizable evidence about real-world usage. Stein pointed to the study on oral versus injectable antipsychotics presented by Alphs, and said that the decisions about patient treatment in this case were based on what the study sought to discover. That is, the study was designed to look at patient outcomes on these two drugs in a real-world setting. Closely monitoring the group that was taking the oral antipsychotic—or ensuring compliance through directly observed therapy—would have defeated the purposes of the research. If research is being conducted for the purpose of a regulatory decision, said Stein, there are specific parameters of how the treatment and the comparative treatment were administered in order to make decisions regarding relative efficacy (see Box 7-3 about regulatory decision making).

Another consideration when making decisions about patient treatment, said Stein, is the clinical context and the patients to be studied: What is the nature of the disease? Is it progressive or non-progressive? Are vulnerable populations involved and what is their susceptibility to harm? What are the current available treatments?

BOX 7-3
Regulatory Decision Making Regarding Clinical Strategy as Discussed by Schneeweiss

As discussed, the utility of a treatment depends not just on the molecular makeup of the drug and how it impacts a disease, but on a host of usage factors, including access, compliance, dosing schedules, need for monitoring, and patient preferences. Sebastian Schneeweiss, professor of medicine and epidemiology, Harvard Medical School and Brigham & Women's Hospital, asked: "To what extent does the regulator have the discretion to go down the road of regulating clinical strategies versus the molecules very narrowly?" For example, if an oral antipsychotic works better than an injectable antipsychotic on a molecular level (i.e., when patients are perfectly compliant), but the injectable antipsychotic is more effective in practice, could a regulator approve this indication based on this information?

There is a trade-off, said Stein, between internal validity and generalizability. Depending on the specific research question, the intended use of the evidence, and safety concerns, a researcher might choose to emphasize one of these factors over the other.

OBLIGATION TO PATIENTS

Investigators, said Simon, have "dual interests." Investigators have a duty to uphold the protocol of a trial to ensure that the research question is answered; meanwhile, they also have a duty to the safety and well-being of participants. These dual interests can sometimes conflict, and as Simon noted, "We want to make sure the duty to the participant always trumps the duty to the protocol."

Safety Monitoring

Researchers have an obligation to ensure that participants in research are receiving treatment that is appropriate for their condition, and that the treatments are safe, said Alphs. When a relatively new treatment is under investigation, said Alphs, the obligation to patients is greater; there are going to be more restrictions on treatment and there is a need for substantial safety monitoring. By contrast, when a treatment has been used safely and effectively for many years, some of the requirements and safety monitoring can be relaxed. Simon explained two reasons to monitor for safety:

1. For a new treatment, safety monitoring is essential to learn about the unknown adverse effects and to make inferences about how the treatment may or may not work in the real world; and
2. For a treatment for which the adverse effects are already known, safety monitoring is about "doing the right thing" for the participants.

These two uses of safety monitoring, said Simon, require different procedures and study designs. Alphs noted that in the trials on treatment for suicidality, the side effects of agranulocytosis and weight gain were known. In this case, the research was not aimed primarily at discovering more about clozapine's (or olanzapine's) adverse effects, but there was still an obligation to the patients to monitor for potential safety events.

Inclusion of Patients and Real-World Experiences

Researchers' responsibility to patients, said Simon, includes an obligation to include in trials patients who "do not behave as we would always

hope." Patients who behave in understandably human ways need to be involved in trials so that results are generalizable to the real-world population, he said. Robert Califf added that patients with comorbidities or concomitant drugs are more likely to experience adverse events, but in traditional RCTs, these patients are excluded. In addition, the issue of how a treatment works for a real-life patient—with comorbidities and concomitant drugs—is what providers and patients "really want to know," said Michael Horberg. If treatments are only tested on patients without these complexities, providers are left grasping for answers on how to treat their real-world patients, such as an older person with hypertension, diabetes, and advanced HIV. "There are real consequences" of limiting trials in this way, concluded Califf.

Acknowledging and welcoming human behavior is sometimes the only way to answer certain types of questions, said Horberg. For example, a study on HIV preexposure prophylaxis consistently advised sero-discordant couples to practice safer sex and to use condoms. However, during the course of the trial, there were more than 100 pregnancies, and the fact that these pregnancies happened without the transmission of HIV was one way that researchers could show that the therapy was effective, he said. Another example of necessary inclusion, said Simon, is seen in the case of the lithium trial for prevention of suicide. Katz had reported that patients who were taking diuretics or ACE inhibitors were initially excluded from the study. However, said Califf, the people most at risk for suicide may be exactly this population of older men with chronic health problems who are taking such medications. Excluding them from the trial not only reduces generalizability, but leaves this at-risk population without knowledge of what treatments might work.

At the third workshop, Alex John London, Clara L. West Professor of Ethics and Philosophy at Carnegie Mellon University, added a comment about how the view on vulnerable populations in research has changed over the years, through the lens of the Council for International Organizations of Medical Sciences (CIOMS) guidelines. The initial view was "protectionist," he said; there were concerns about involving vulnerable populations in research. Now, CIOMS has revised its guidelines to a "justice-based approach," which encourages the participation of vulnerable patients so that research can generate the information necessary to treat these populations. Simon concluded, "Unless we welcome in the way the real world works, we are never going to answer these questions."

Standard of Care

Califf said that when researchers are designing trials, they have to decide on one of three options for the treatment for the control group: the

best available care, usual care, or "whatever care is available where you are doing it." If the research question is specifically about how a treatment works in a real-world setting, using a standard of care that is above the usual care for the control group may not provide an accurate answer, said Stein. However, he noted, there is always an obligation to ensure that the care delivered to the control arm is not substandard care. Simon said that although the effect of a treatment might be magnified if the control arm received substandard care, doing so would be unethical. Adrian Hernandez noted that research often seeks to understand whether one treatment is better than the standard of care, but that the standard of care can vary significantly by region, due to clinical practice and access issues. This variation can make it challenging to detect whether and to what extent a new treatment is better than the standard.

Simon asked Katz about balancing the needs of both the research question and the extremely vulnerable patients in the study on lithium for suicide prevention. Simon noted that the researchers had an obligation to stay engaged with the patients in the placebo arm, and to closely monitor for suicide and to hospitalize the patients if they suspected the possibility of a suicide attempt. However, at some point, the researchers might control and augment the care of participants so much that it would prevent the outcome from ever occurring and guarantee a null result, he said. Katz responded that there was extensive debate about this issue in defining outcomes for the study. The final approach to addressing this problem, at least in part, was to include hospital admissions with documentation that the reason for admission was specifically for prevention of suicidal behavior. This addresses the issue, but it "softens" the outcome. To minimize bias in deciding what events should be considered outcomes, the study uses a process of outcome adjudication, based on reviews of study documents and medical records by independent clinician-investigators blinded to treatment assignment. In general, the VA may be a unique site for this type of research because the baseline of care for suicide prevention at the VA is probably above the community standard. There is significant infrastructure in place, and there are requirements for flagging patients at risk of suicide and facilitating their access to care.

Marc Berger added that another option is to use observational data; these could allow researchers to study the outcomes of usual care without making fraught decisions about obligations to research subjects. Observational research can provide great insight into real-world effectiveness, he said. He gave the example of an oral drug and an inhaled corticosteroid used for childhood asthma. The oral drug was not as efficacious, but was equally effective in the real world because children did not want to use their inhalers at school, he said. Gaining this insight from an RCT would not be possible, but it is important information that can be used to make patient

treatment decisions, he said. Evidence from observational data can be used, said Berger, along with evidence from RCTs or pragmatic clinical trials to build a "corpus of evidence" about a treatment. (See Chapter 9 for further discussion of observational trials.)

8

Obscuring Intervention Allocation in Trials to Generate Real-World Evidence: Why, Who, and How?

Key Messages Identified by Individual Workshop Participants

- Obscuring intervention allocation, or blinding, allows researchers to study the effects of an intervention without influence from patient and provider perceptions. However, it may not always be appropriate or feasible. (London, Watanabe)
- The appropriateness of blinding is highly dependent on the context of the study, the research question at hand, and the specific uncertainties for that study. (Critchlow, Dreyer, London, Smith, Vardeny)
- Uncertainties in a study can be classified along two axes: ensemble efficacy and utilization factors. The interaction of these two for a particular study can indicate whether blinding is appropriate. (London)
- Blinding can impart risks to the study, such as patient exposure to either sham procedures or to incomplete care from their provider. These and other risks are balanced against the benefits of blinding in order to make a decision. (Dreyer, Hernandez, London, Simon)
- Decisions on blinding can also be influenced by practical considerations, such as effects on study cost, feasibility of masking or delivering the treatments, patient preferences, and the generalizability of data from blinded study sites. (Dreyer, Reynolds, Smith)

- Patient and provider bias, and how it may influence a study, can be difficult to predict in advance. Bias may not affect some outcomes, such as quantitative laboratory readings or all-cause mortality, but can affect other more subjective outcomes or have other effects such as ascertainment or treatment bias. (Critchlow, Dreyer, Graham, Smith)

Obscuring intervention allocation—commonly known as blinding—is used to reduce the likelihood that those involved in the trial will be influenced by the treatment assignment, said James P. Smith, deputy director of the Division of Metabolism and Endocrinology Products at the U.S. Food and Drug Administration's Center for Drug Evaluation and Research. If those involved are influenced by treatment assignment, it may affect the outcome of the trial. Blinding allows researchers to study the effects of the intervention without the influence of patient and provider perceptions. However, blinding is not always appropriate or feasible, said Alex John London. Trial design features such as blinding have to be justified by the contribution they make to the evidence quality and the relative risks and costs compared with alternative designs, he said. This session in the third workshop examined the topic of blinding, and session moderator Jonathan Watanabe, associate professor of clinical pharmacy and National Academy of Medicine Anniversary Fellow in Pharmacy, University of California, San Diego, asked participants to consider the following:

- How might variability in knowledge of treatment group assignment affect provider and patient adherence and outcomes?
- How might variability in knowledge of treatment assignment affect study cost and reliability?
- What key factors could affect decisions to obscure intervention allocation?

This topic was highlighted by some participants at the second workshop as an area that needed further exploration, so a session was held on the topic at the third workshop.

ILLUSTRATIVE EXAMPLES

To explore the issues surrounding obscuring treatment intervention and blinding, speakers at the second and third workshops presented case studies as illustrative examples of the considerations that go into designing and conducting a real-world study.

INVESTED Trial

Orly Vardeny, associate professor of medicine at the University of Minnesota and of the Minneapolis VA (U.S. Department of Veterans Affairs) Center for Chronic Disease Outcomes Research, told workshop participants about the INVESTED (INfluenza Vaccine to Effectively Stop Cardio Thoracic Events and Decompensated heart failure) trial. Influenza leads to significant morbidity and mortality, said Vardeny, and there have been several analyses that documented a temporal association between influenza infection and cardiovascular events (Madjid et al., 2007; Thompson et al., 2003, 2004). Researchers have sought to further understand how the influenza vaccination could affect cardiovascular events. The first step, said Vardeny, was a meta-analysis of six studies that demonstrated that influenza vaccination could reduce cardiovascular events (Udell et al., 2013). The next research question, said Vardeny, was to determine if patients with heart failure exhibited the same immune response to the flu vaccine. This analysis showed that heart failure patients had a reduced antibody response compared with patients without heart failure (Vardeny et al., 2009). Researchers were able to increase the antibody response by giving a higher dose of the flu vaccine, she said (Van Ermen et al., 2013).

Based on this initial research, the INVESTED[1] researchers set out to examine whether the increased immune response from the high-dose vaccine would translate into better cardiovascular outcomes (see Figure 8-1). The researchers designed a randomized trial comparing high-dose trivalent influenza vaccine to standard-dose quadrivalent influenza vaccine. Vardeny noted the standard-dose quadrivalent vaccine for the active control group was chosen because it would be unethical to have an unvaccinated control group, and the quadrivalent vaccine is the current standard of care. Researchers have enrolled approximately 3,000 participants out of a target of 9,300 participants for this pragmatic study so far, and have contacted them up to four times per year by phone to ascertain endpoints. The primary endpoint is a composite of death or cardiopulmonary hospitalization.

Both providers and participants are blinded in the study, which is accomplished by a third-party vendor affixing identical labels over the individual-dose syringes. Vardeny said that because there were some ways to tell the two vaccines apart (e.g., the high dose causes more pain during injection), the study was designed so that the person who administers the vaccine is not the same person who conducts the ascertainment phone calls later in the season.

Vardeny said the choice to conduct a double-blind study was based on a few considerations (see Box 8-1). There were reasons not to blind,

[1] See https://clinicaltrials.gov/ct2/show/NCT02787044 (accessed January 4, 2019).

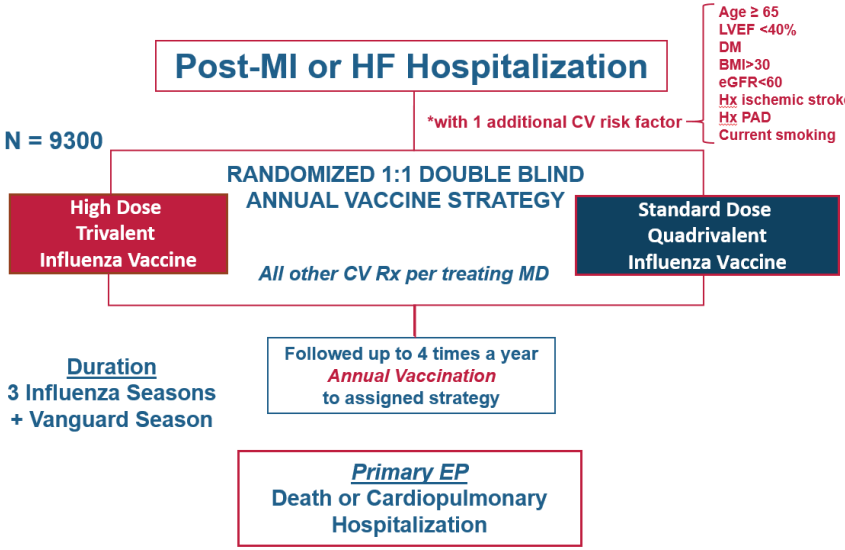


FIGURE 8-1 Study design for the INVESTED trial.
NOTE: BMI = body mass index; CV = cardiovascular; DM = diabetes mellitus; eGFR = estimated glomerular filtration rate; EP = end point; HF = heart failure; Hx = history; Hx PAD = history of peripheral artery disease; LVEF = left ventricular ejection fraction; MD = Doctor of Medicine; MI = myocardial infarction; Rx = prescription.
SOURCE: Vardeny presentation, July 17, 2018.

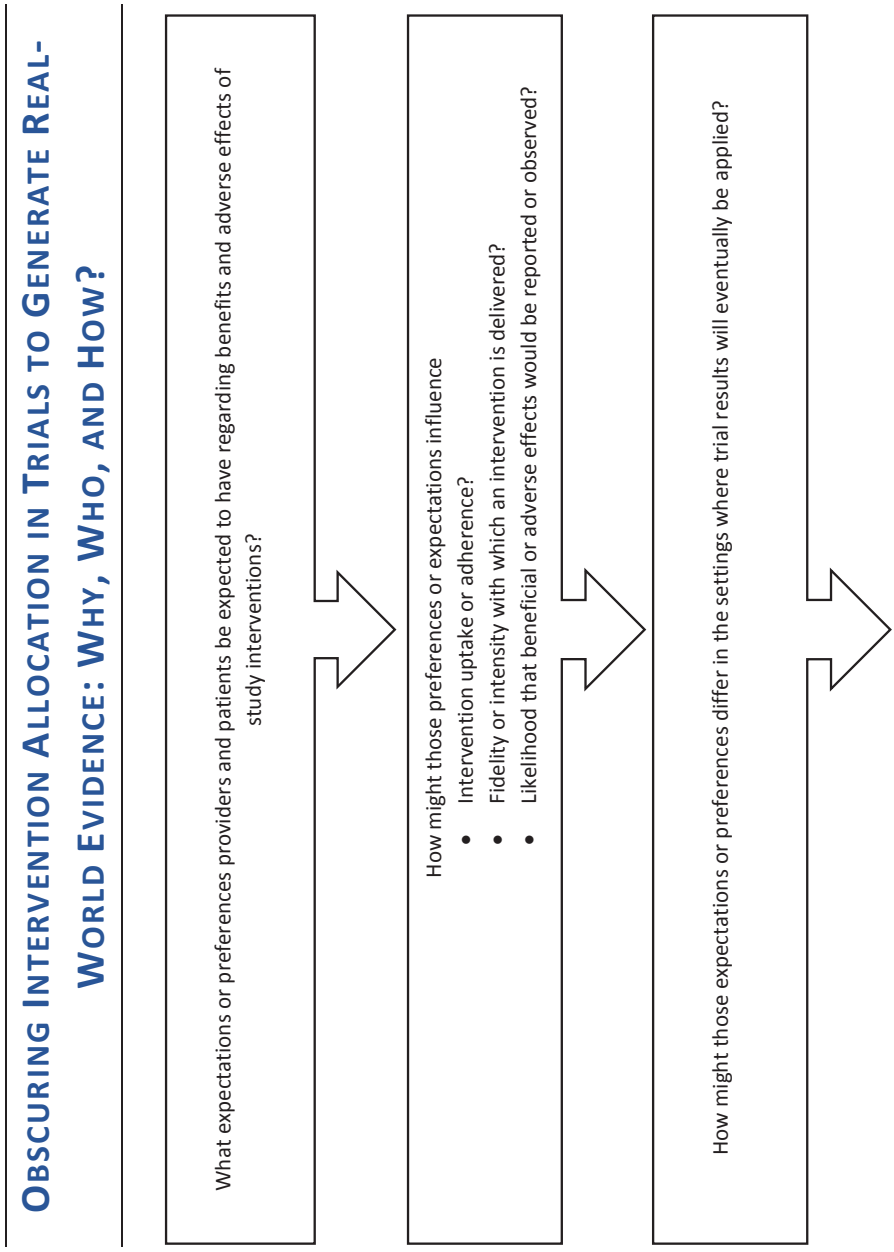
she said. Blinding delays the study by several weeks while the third-party vendor blinds and distributes the vaccines. During this period, the flu vaccine becomes available to the general public, which reduces the number of people eligible to participate in the trial because they already received the vaccine. However, said Vardeny, the reasons for blinding outweighed these considerations. There are perceived differences in efficacy of these vaccines, she said. For example, the high-dose vaccine is only approved for individuals aged 65 and over. Although there is no recommendation that the standard of care for people over 65 should be the high-dose vaccine, some providers or patients may perceive that all people over 65 should get the high-dose vaccine. Without double blinding, these perceptions may have resulted in systematic bias in terms of patients presenting at the hospital, or providers urging hospitalization.

BOX 8-1
Other Examples of Blinded and Non-Blinded Studies Discussed Throughout the Workshop Series as Presented by Individual Workshop Participants

- **The Salford Lung Studies to understand the effectiveness of asthma and chronic obstructive pulmonary disease drugs in the usual care setting** (discussed in detail by Gibson and in Chapter 3). Graham discussed in this session that patients were randomized to either group, and the study was conducted open label (i.e., not blinded). There were several reasons why the researchers chose not to blind patients in the study, said Graham. First, in the United Kingdom (where the study was conducted), the standard practice is for patients to receive their medicine from a pharmacist who educates them on the use of the drug and the importance of adherence. Had the study been blinded, this real-world aspect of clinical care would have been altered and potentially caused confusion or changes in adherence for patients. In addition, said Graham, the delivery mechanism of inhalers can impact outcomes, aside from the specific drug that is inhaled. Blinding may have complicated the ability to see the benefit of the inhaler itself, he said.
- **Novel oral anticoagulants (NOACs) versus warfarin for controlling blood pressure.** Hernandez presented trials comparing NOACs with warfarin; three were double-blind trials (see Chapter 6 for full details). He noted that randomizing more than 14,000 patients to two treatment arms in a double-blind study was "quite a challenge." He also noted that any study that involves warfarin is challenging to maintain the blind because warfarin requires frequent monitoring. Researchers must also ensure that patients in the warfarin arm are appropriately anticoagulated, without revealing the treatment assignment.
- **Lithium for suicide prevention.** Ira Katz presented a research project that evaluated lithium for suicide prevention (see this chapter for full details). The trial was double blinded and placebo controlled.
- **The INTERSEPT and PRIDE trials.** Larry Alphs presented two pragmatic trials (see this chapter for full details). The INTERSEPT trial tested two oral antipsychotic drugs, and was not blinded because of differences in known side effect profiles, additional blood draw monitoring required by only one of the drugs, and because the investigators judged that it was unethical for this study to be blinded because that may have inhibited individual providers' ability to give the best care to their patients. The PRIDE trial compared an oral to a long-acting injectable antipsychotic and was open label because the different delivery mechanisms would have been impractical to blind.

DECISION AID

The general issues discussed by individual workshop participants in the first and second workshops were used to develop a decision aid for the third workshop (see Figure 8-2). As with the other decision aids, the intention was to outline some questions to consider in order to make thoughtful



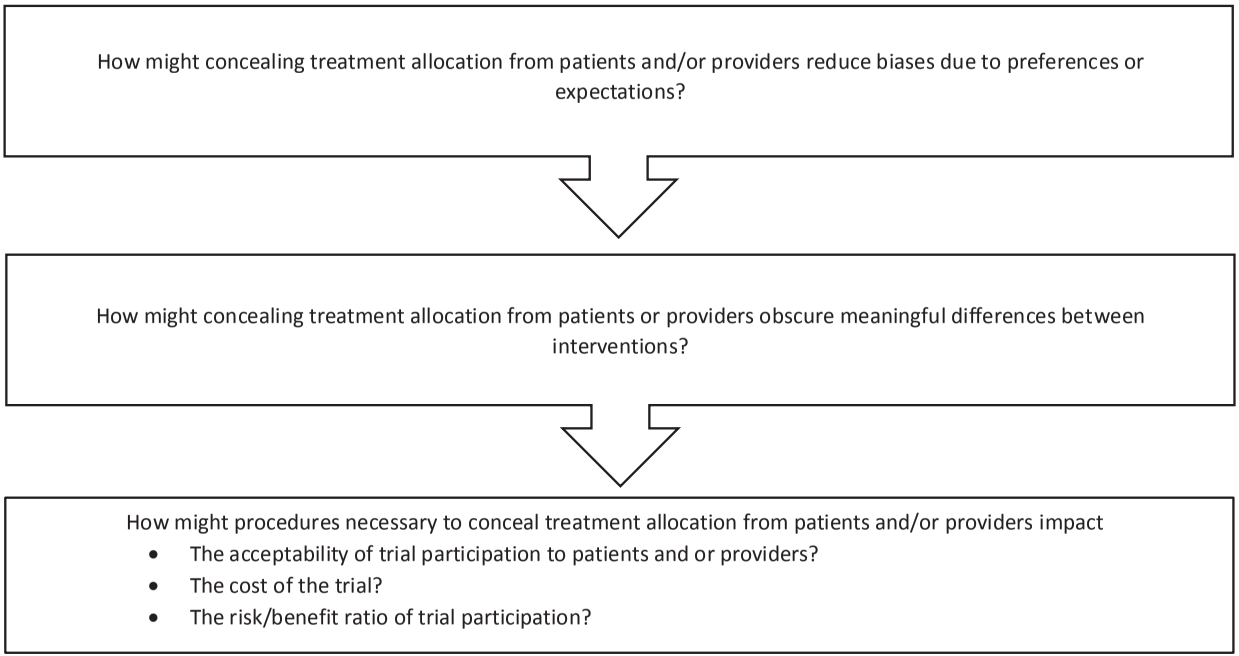


FIGURE 8-2 Decision aid on questions to consider regarding when and whether to obscure intervention allocation (commonly known as blinding) in trials intended to generate real-world evidence.
NOTE: This decision aid was drafted by some individual workshop participants based on the discussions of individual workshop participants at the first and second workshops in the real-world evidence series. The questions raised are those of the individual participants and do not necessarily represent the views of all workshop participants; the planning committee; or the National Academies of Sciences, Engineering, and Medicine, and the figure should not be construed as reflecting any group consensus.
SOURCE: Watanabe presentation, July 17, 2018.

choices in real-world evidence (RWE) study design. Participants at the third workshop reflected on these questions and offered feedback on the decision aid throughout the course of their discussions.

In addition to offering specific feedback on the decision aid (see Box 8-2), workshop participants also discussed the difficulty of answering these types of questions prior to conducting a study. Simon summarized that although the questions in the decision aid may be good questions to consider, making decisions about the risks and benefits of blinding in a study a priori is enormously difficult. Watanabe asked the participants if they thought that the answers to the questions on the decision aid could be quantified a priori—for example, if researchers could, before a study, quantify the risks and benefits of blinding and make a decision based on these numbers. Cathy Critchlow, vice president at the Center for Observational Research at Amgen Inc.; John Graham; and Vardeny all said that quantifying would be enormously difficult. Critchlow added that this sort of quantification could be risky because it could make the researchers "feel better about making a decision that may not be based on something real." Richard Platt wondered if some of these a priori decisions could be based on empirical data on bias from previous research—for example, data on how aggressively vaccinated patients are treated versus non-vaccinated patients. Vardeny said that vaccinated patients are so different

BOX 8-2
Feedback on the Decision Aid as Discussed by Individual Workshop Participants

Several workshop participants offered specific feedback on the decision aid "Obscuring Intervention Allocation in Trials to Generate Real-World Evidence: Why, Who, and How?"

- The evaluation of a potential real-world trial needs to consider the existing knowledge base and what other studies have been conducted. For example, the Salford Lung Studies were planned after two standard, randomized, placebo-controlled trials had been conducted. This is a very different evaluation from using the Salford Studies for regulatory approval. The decision aid should reflect this contextual difference. (London)
- One of the questions on the decision aid asks, "How might concealing treatment allocations reduce bias?" This question should instead ask, "How *much* might concealing treatment allocations reduce bias?" (Dreyer)
- These questions will not generally have yes/no answers; rather the goal should be to estimate *how much* a research design decision (e.g., blinding) will affect bias or other issues. (Graham)

from unvaccinated patients that it would be difficult to isolate the effect of any provider bias. Smith said an alternative approach could be statistically adjusting for the bias after the research was complete, which would still require a judgment, but may be conceptually easier than quantifying the bias a priori.

Simon suggested that there may be value, after a study is completed, in analyzing decisions made in the study to learn from any mistakes that may have been made. Smith concurred and said that "an autopsy of trials that do not go the way one expected can only be an incredibly valuable exercise," and could be a best practice. Graham added that trial "autopsies" could be done not just for trials that gave a surprising result, but for all trials. He said that researchers could spend more time looking back at their decisions and the impact the decisions had on the results, and could include this information in their published results. Each study can give researchers insight into how to improve the next study, he said.

DISCUSSION

Context of the Decision

Critchlow echoed an earlier discussion (see Chapter 6) in which participants stressed the importance of considering the context of the decision when designing a study that uses real-world data (RWD). The decision about whether or not to blind participants in a study, she said, should consider what the study is meant to accomplish: Is it to answer a question about effectiveness, is it to inform treatment guidelines, is it to get approval for a new treatment, or is it to expand the label indications for an already approved treatment? Depending on the use of the study, blinding may be more or less appropriate. Researchers, she said, need to ask themselves: Is blinding helping us to answer the right question and make the right decision? Nancy Dreyer, chief scientific officer at IQVIA, agreed with this framing, and added a more specific question: What is the expected effect estimate, and how likely is it to be missed? For example, if a study that will use RWD is about a chronic disease for which the treatments might make a small but important difference, blinding might be necessary to get the needed effect size for the decision to be made, she said. Smith agreed that making a decision about blinding depends on the research question. For example, he said, if the research is seeking answers about whether patients are more likely to adhere to an injectable therapy or to an oral daily tablet, a blinded study cannot answer that question.

London concurred with the other speakers, saying that trial design decisions should be based on the specific uncertainties that a study is aimed at mitigating (see Box 8-3).

BOX 8-3
Considerations Around Uncertainty in Study Design

Uncertainty matters:

1. Nature of the uncertainty should influence trial design.
2. Uncertainty about relative risks/potential benefits should be clearly communicated to participants.
   a. Uncertainty versus disagreement
3. Role of design in addressing uncertainty should be explained.

SOURCES: London presentation, July 17, 2018; concept from London, 2018.

For example, if there are uncertainties about efficacy, blinding may be necessary, whereas if there are uncertainties about effectiveness in the real world, blinding would not be warranted. London explained that real-world effectiveness is a function of two variables: ensemble efficacy and utilization factors.

- The term "ensemble efficacy" acknowledges the fact that a drug is not efficacious on its own—the therapeutic effect depends on the proper dose and schedule, the right population, and co-interventions or diagnostic requirements.
- Utilization factors include patient and provider awareness and preferences, cost, adherence, clinical capacity, and tolerability. Even if an ensemble is efficacious, said London, the real-world effectiveness of the drug depends on the utilization factors. For example, a drug could be efficacious for reducing blood pressure, but if it is unaffordable, difficult to adhere to, or poorly tolerated, it will not be effective in the real world because patients will not use it.

Blinding may be more or less appropriate, depending on the state of knowledge about the ensemble efficacy and utilization factors, said London. If two interventions are being compared for real-world effectiveness, and it is known that Intervention 1 is more efficacious than Intervention 2, but the utilization factors are unknown, blinding would not be warranted (see Figure 8-3). The research question in this scenario, said London, is how these real-world factors such as patient preference will impact the effectiveness of an intervention. If patients and providers were blinded, the utilization factors would remain unknown and thus the real-world effectiveness would not be understood.

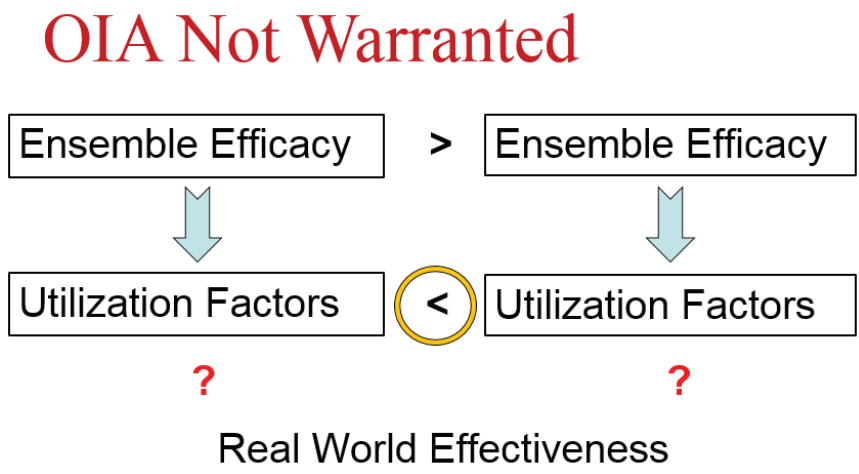


FIGURE 8-3 Blinding not warranted when utilization factors are unknown.
NOTE: OIA = obscuring intervention allocation.
SOURCES: London presentation, July 17, 2018; concepts from Kimmelman and London, 2015, and Moseley et al., 2002.

However, he said, a measure of real-world effectiveness may not always be sufficient for making policy or care decisions, compared to a measure of efficacy. London gave the example of arthroscopic surgery for osteoarthritis of the knee. This procedure, he said, was performed about 650,000 times annually at a cost of $3.25 billion. However, when a randomized controlled trial was performed to compare the procedure to a sham procedure, it was found to have no benefit (Moseley et al., 2002). The sham procedure, said London, was "largely theater" (manipulations, surface cuts, and bandages), but was necessary to obscure the intervention allocation. In this case, blinding was essential for a clear picture of intervention efficacy, and these efficacy data were necessary to determine the clinical merit of the procedure.

Risks of Blinding

When blinding is necessary to fully understand the effectiveness of an intervention, participants in the control group receive some sort of placebo drug or intervention, London said. This could range from a sugar pill, to a sham surgery that is "largely theater," to invasive surgical procedures such as implanting cannulas in the brain to deliver saline in Parkinson's trials, said London. There are risks to these procedures, and the benefits of a placebo are limited. The risks of these control procedures need to be disclosed to patients, he said. However, it is important to note that there

are also risks involved with the active procedure, and there may also be no benefit. For example, while there are minor risks involved with a sham knee surgery, the real knee surgery is far more invasive and risky. If the real knee surgery does not have any benefit, London said, the risks far outweigh the benefits. Both the risks from the placebo or sham procedure and the risks from receiving a potentially ineffective active intervention should be communicated to participants, London said. (See Chapter 7 for more discussion of patient safety in real-world studies.)

Another potential risk of blinding is that the participant population may be more homogeneous and more tightly controlled, which may make the results less generalizable, said Adrian Hernandez. The process of enrolling in a trial, and consenting to all of the processes necessary for blinding (e.g., complying with regular monitoring), may filter out some populations of participants.

Blinding can "obscure the truth," said Simon. "We fundamentally distort the nature of the treatment" by making two active treatments seem identical and by ensuring that the patient and provider experiences with the treatments are identical. For example, he said, if two treatments have very different dosing schedules, this difference may have an important bearing on the acceptability and adherence to those treatments, he said. While there are obvious advantages of blinding (e.g., mitigating bias), it can also constrain opportunities to learn about how a treatment will work in the real world.

In discussions about whether or not to blind, said Dreyer, much attention is given to the risk of bias if a provider knows to which treatment group a patient belongs. However, she said, there are risks to the patient if a provider does not know which treatment he or she is receiving. For example, a provider may not know what adverse effects to monitor, or when a change in health status may indicate a significant problem. Dreyer said that while blinding can have benefits for the study, these benefits need to be weighed against the risks to the patient.

Cost Considerations

Researchers, said Dreyer, aim to get precise answers to research questions and then compare their answers to those that others get in order to confirm or refute the original hypothesis. However, Dreyer said that as a consumer and generator of data, researchers should be looking to see if two estimates are close, and certainly in the same direction, rather than expecting them to match exactly. Rather than designing a study to generate the most exact answer—which can increase costs significantly—she said, researchers should consider the big picture of the RWD's intended use, and design studies accordingly. Dreyer said the economic value of different

study designs could be considered in the designs; if the answer can be found with an open-label, real-world study rather than from a more expensive randomized blinded trial—and it can be shown that bias is unlikely to explain the observed effect—it may be a better use of resources to choose the less expensive design, she said. Smith had a slightly different perspective. He said that while blinding can certainly increase the cost of a trial, an unblinded trial may "give you results that nobody believes," which wastes time and resources.

Practical Considerations

For some interventions, such as vaccines, blinding is quite feasible and does not overly complicate the study, said Dreyer. However, for other interventions that involve treatments that are complex, sequential, or require monitoring and dose adjustment, it may be impractical or impossible to blind while keeping patient safety in mind. There are situations, said Smith, where blinding is simply not feasible. For example, some drugs have severe side effects, and to try to duplicate these side effects in the control group would be impossible or unethical. Blinding is a "tool that can be valuable in a narrow set of circumstances," Dreyer said, but it should not necessarily be employed in all circumstances.

Hernandez noted that blinding could have the advantage of keeping participants in a trial. A participant who does not know if he or she is getting the active treatment may be more likely to continue care, he said.

Conducting a blinded study in a real-world setting may be difficult, said Rob Reynolds, vice president and Global Health, Epidemiology, Pfizer. Blinded RCTs are difficult enough, he said, but in a real-world environment with diverse health care sites and patient populations, the feasibility challenges are even greater. Dreyer joked that "a lot of things are feasible if you spend enough money." She added that when doing blinded studies in the real world, researchers generally perform the studies at academic centers that are more able to carry out this type of research. However, this may result in a lack of generalizability because these centers are often especially high-quality facilities with a high quality of care and follow practices that may not be routinely conducted in community settings.

Patient Preferences

One question on the decision aid asks, "How might procedures necessary to conceal treatment allocation from patients and/or providers impact the acceptability of trial participation to patients and/or providers?" Smith said the best way to answer this question is to ask the patients directly.

Smith relayed an experience he had working with patients with a rare disease. He consistently heard that it would be impossible to conduct a placebo-controlled blinded trial in this population, but when meeting with a group of patients directly, the patients said if blinding "will give you the best data if we would be willing to go on a placebo for 6 months, we are happy to do so." Smith said that assumptions about patients' willingness to participate in blinded trials may not always be accurate. Critchlow added that in a context where there are fewer therapeutic options, patients might be more willing to take the risk of participating in a trial, whereas in a context with plenty of existing therapeutic options, the risk–benefit analysis of patients and providers is quite different.

On another patient-related topic, John Burch, an investor with the Mid-America Angels Capital Network, asked whether participants in a blinded trial are ever able to get information about which treatment they received. "If patients are entitled to their own information," he said, "does that include patient-specific data from research?" Graham said that from an ethical perspective, research is likely progressing to a point where patients will be able to get information about their treatment in a trial.

Biases

Blinding, said Smith, is one mechanism to reduce the possibility of bias that would affect the trial results. There are multiple types of bias that can be associated with treatment assignment, some of which are discussed below. Smith said bias is often subconscious, and it can be difficult to predict the direction and magnitude of the bias. Different patients or providers within the same trial might have biases that affect the results in different directions, he said. For example, if aware of treatment assignment, one patient may believe that a novel drug will help, and therefore may overstate its benefit. Another patient may believe that the standard of care is fine and could be skeptical of a company's development of a novel drug. The direction and magnitude of various biases may differ depending on the specific disease area, type of intervention, or patient population. Overall, Smith said, "It is very difficult to predict how bias is going to creep in and what direction and magnitude it is going to be."

Patient Perception

Without blinding, there is potential for bias due to patient perception, said Critchlow. For example, patient perception may impact the likelihood of complying with a treatment regimen as intended. Dreyer said that bias due to patient perception can differ, depending on the subjectivity of the measurements. For example, if the outcome of interest is hemoglobin A1C

levels, patient perception is unlikely to have an effect. On the other hand, if the outcome of interest is pain level as reported by the patient, this could be influenced dramatically by perception, she said. Dreyer noted that for some outcomes, patient perception likely has no effect—"dead is dead." Graham added that depending on the time line of the study, patient perception bias may wane over time. For example, if a patient gets a vaccine and feels ill 3 weeks later, does the patient's initial idea of what treatment he received still affect his likelihood of seeking medical care? Graham said this is a very difficult area to have a black or white answer.

Ascertainment Bias

If the people who are assessing outcomes are not blinded, said Critchlow, it can result in ascertainment bias. The potential for ascertainment bias may depend on whether the endpoints are subjective or objective. For example, if the endpoint is a quantitative laboratory result, it is unlikely to be biased by the assessor knowing to which group the patient is assigned. Outcomes assessors, said Smith, can be blinded to mitigate ascertainment bias, but it must be recognized that "they only assess what has been given to them." For example, a blinded pathologist can read a biopsy specimen, with no knowledge of the clinical course or treatment. However, blinding the assessor does not automatically remove the potential introduction of bias. For example, although the pathologist may be blinded, the decision by the patient's provider to conduct a biopsy could have been influenced by their knowledge of treatment assignment. In this scenario, said Smith, "the bias has already been introduced before the outcome got to the outcomes assessor."

Treatment Bias

When providers are not blinded during a trial (i.e., double blinded), said Smith, there is the worry that patients in one group are being treated differently than patients in the other group. For example, if a provider knows that a patient is on a placebo for suicide prevention versus an active treatment, the provider may be more likely to closely monitor the patient for signs of suicidality. The bias can also run in the other direction—if, for example, a provider knows that a patient is receiving an active but not well-understood treatment, the provider might monitor more closely for adverse events.

Channeling Bias

Channeling bias occurs later in the life cycle of a product when the adverse events associated with a product are well known, Jesse Berlin said during the second workshop. Patients who are at higher risk for these events are prescribed a different product; when these high-risk patients experience the event, the other product "starts to look like it's increasing risk." For example, non-steroidal anti-inflammatory drugs (NSAIDs) are associated with an increased risk of myocardial infarction (MI). Because of this, patients who are already at high risk of MI may be prescribed acetaminophen instead of NSAIDs, which results in a larger-than-expected group of acetaminophen patients experiencing MI.

9

Gaining Confidence in Observational Comparisons

Key Messages Identified by Individual Workshop Participants

- Confidence in database studies, or observational comparisons, is related to the type of effect being detected. (Schneeweiss)
- Database studies may be more appropriate when the outcomes and exposures are measurable in the data; when two active treatments are compared; and when the key confounding variables are measurable. (Schneeweiss)
- For disease areas and interventions where additional evidence could be beneficial but randomized trials are impractical, evidence from observational studies could be used to fill the gap. (Madigan)
- Conveying uncertainty in any study can aid in assessing what purposes the study could be used for and how to interpret the results. (Madigan, Martin)
- Transparent reporting of methods used in studies can promote replicability and also aid in assessing study validity. (Madigan, Schneeweiss)
- Replicating randomized clinical trial results with observational databases can help establish a process and criteria for conducting observational studies more widely. (Franklin)
- New methods such as predictive analytics and machine learning can potentially be used to predict outcomes for individual patients or to identify associations. (Jimenez, van der Laan)

- Regulators consider defined, sequenced criteria to evaluate evidence, including whether the data are relevant for the proposed indication, the outcomes are well assessed, the methods minimize bias, and the statistical analyses are rigorous. (Gormley)
- Regulators at the Center for Devices and Radiological Health have recently developed a methodology to support premarket observational studies that minimizes bias by restricting statisticians' access to the outcome data before the analysis is run. (Li)

Randomization, by its nature, minimizes the effect of unmeasured confounders and other biases and allows straightforward inferences to be made from the gathered data. However, it is not always feasible and can be costly. When using observational data, what are ways to account for these biases and unmeasured confounders? What steps can be taken to ensure that the results from observational research are sound? In these sessions, participants explored these questions and heard from speakers who have developed methods to gain confidence in unrandomized observational comparisons. Throughout the presentations and discussions, workshop participants also heard an alternative perspective from some speakers who emphasized the continuing importance of randomization and discussed methods to make randomization easier in real-world settings.

ILLUSTRATIVE EXAMPLES

To explore the issues surrounding bias in observational comparisons, speakers at the second and third workshops presented case studies as illustrative examples of the considerations that go into designing and conducting observational data analyses.

Health Care Databases for Regulatory Decision Making

Data for effectiveness research in health care, said Sebastian Schneeweiss, professor of medicine and epidemiology at Harvard Medical School and Brigham & Women's Hospital, can come from a number of different sources (see Figure 9-1). The first categorization, Schneeweiss said at the second workshop in this series, is between traditional randomized controlled trials (RCTs) and non-interventional studies. Within non-interventional studies, there are two major different sources of data, he said. One source is primary data that are generated for the purpose of conducting research on individuals and the investigator controls what, how, and when to measure; the other source is transactional data that are generated for other purposes, but used secondarily for research. Within transactional data, there are numerous cate-

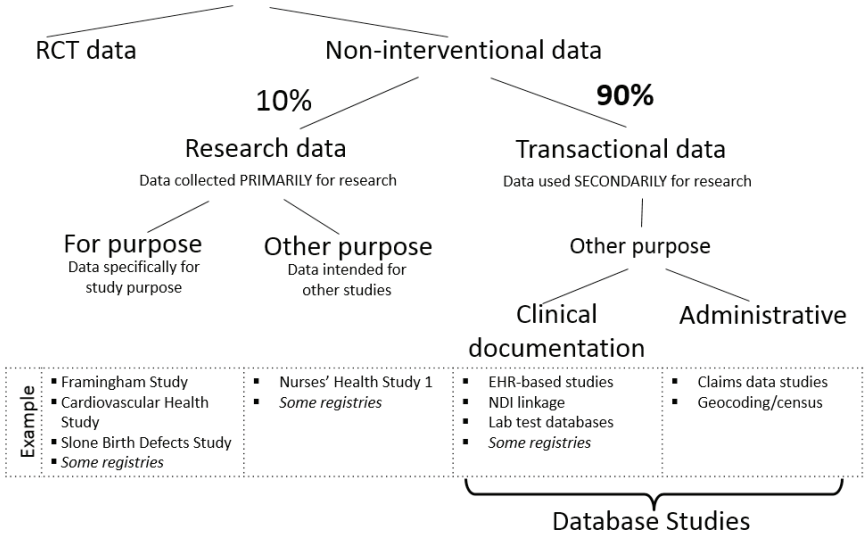


FIGURE 9-1 Sources of data for research.
NOTE: EHR = electronic health record; NDI = National Death Index; RCT = randomized controlled trial.
SOURCES: Schneeweiss presentation, March 6, 2018; Franklin and Schneeweiss, 2017.

gories: electronic health records, laboratory results, and administrative data such as insurance claims. Schneeweiss said that in pharmacoepidemiology, approximately 90 percent of research is done with these types of transactional data.

Schneeweiss walked workshop participants through the process of using transactional data for research. First, he said, there is a dynamic database that records an ongoing stream of new health care records for all enrolled patients. These data are stabilized into a "snapshot" for research purposes. This ensures that the research will be replicable if the analysis is run again, said Schneeweiss. The study rules are then applied to all of the health encounters of individual patients (e.g., hospitalizations, diagnoses, procedures).

Real-world data (RWD) have four different potential uses in regulatory decision making, said Schneeweiss. RWD can be used as synthetic control arms for single-arm trials for primary approval. RWD can be collected and analyzed for secondary indications—such as a different outcome or different population—for a product that is already on the market. RWD can also be used as part of the initial approval process, when the indication has been

based on surrogate endpoints with the understanding that evidence of clinical endpoints will be generated before the product receives full approval. Finally, RWD can be used for safety assessment and monitoring, either in the immediate postmarket time frame, or if a safety concern has arisen later in the product's lifetime.

Schneeweiss provided several examples of results from database studies that came to the same causal conclusion as RCTs (Connolly et al., 2009; Fergusson et al., 2008; Fralick et al., 2017; Giles et al., 2016; Kim et al., 2017; Neal et al., 2017; Patorno et al., 2018; Schneeweiss et al., 2008; Seeger et al., 2015; Zinman et al., 2015). Schneeweiss pointed out that in these examples, a new-user, active-comparator design was chosen; determining why these database studies and RCTs came to the same conclusion would be key to indicating when database studies could be appropriate and reliable in future research. More importantly, said Schneeweiss, "How confident can we be that a new database study will generate results that are comparable to those from an RCT?" Confidence in non-experimental database studies, said Schneeweiss, depends on two factors. The first is whether the study examines beneficial effects or harmful effects, and second, whether it looks for the intended treatment effect or discovers unintended effects (see Figure 9-2). Depending on the answers to these questions, said Schneeweiss, the a priori confidence level in RWD will be higher or lower. For example, confidence in RWD is higher when the study finds an unintended (and unknown) effect, in part because the provider is "blinded" to the effect so the confounding is less strong, he said.

There are three reasons, said Schneeweiss, why researchers prefer RCTs to observational studies. First, and most obvious, is that RCTs use random treatment assignment. Second, there is controlled outcome measurement. Third, the implementation of RCTs is clear and easy to understand. Schneeweiss emphasized that "sweat and tears" go into the design of an RCT, but the actual analysis is quite straightforward compared to the complex analytics used in database studies. Despite these advantages of RCTs, Schneeweiss offered his perspective on when a researcher might feel comfortable conducting a database study instead:

- First, an active comparator is preferred. Using a database to study the difference between two active treatments is far easier than comparing a treatment to patients who did not receive treatment because "there is usually a reason why they did not get treated."
- The second requirement for a valid database study is that outcomes and exposures need to be measurable and observable in the data.
- Finally, the key confounding variables need to also be measurable and observable in the system.

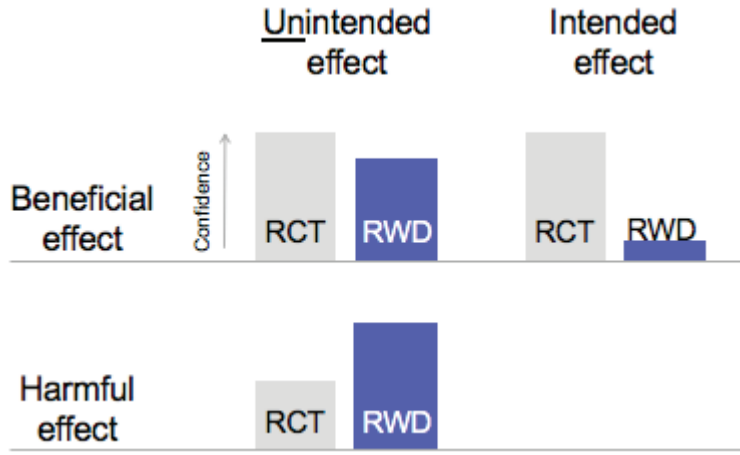


FIGURE 9-2 A priori confidence in validity of study findings.
NOTE: RCT = randomized controlled trial; RWD = real-world data.
SOURCE: Schneeweiss presentation, March 6, 2018.

There are research questions that need to be answered by RCTs, and research questions that can be answered through RWD analyses, said Schneeweiss. There is an unknown area of overlap between the two. "If you can identify that group of questions that are relevant for decision makers and that we feel confident we can answer without randomization" and without putting patients at risk, RWD analysis could be used with high confidence, he said.

Schneeweiss presented a possible pathway for deciding to conduct RWD analysis. At each step along the pathway, if the answer to a question is "no," an RCT would be more appropriate. Schneeweiss described the checkpoints along the pathway:

- Is the setting adequate for an RWD analysis?
- Is data quality fit for purpose?
- Is the data analysis plan based on epidemiologic study design principles?
- Was balance in confounding factors among treatment groups achieved?

Zostavax Vaccine Effectiveness and Duration of Effectiveness Project

Hector Izurieta, epidemiologist at the Office of Biostatistics and Epidemiology at the U.S. Food and Drug Administration's (FDA's) Center for

Biologics Evaluation and Research, described a real-world study on the effectiveness and duration of effectiveness of the herpes zoster—commonly known as shingles—vaccine, Zostavax (Izurieta et al., 2017). The study used data on Medicare Part D beneficiaries and compared outcomes of those who had been vaccinated with those who had not received the vaccine. Outcomes included herpes zoster and ophthalmic zoster (a subtype of shingles in which the characteristic rash presents at the forehead and around the eyes) medical office visits, postherpetic neuralgia, and herpes zoster hospitalization. The researchers, said Izurieta, used Cox regression models to estimate the risks of herpes zoster and postherpetic neuralgia in the vaccinated and unvaccinated populations, adjusted for the main known characteristics, and measured the risk at different time intervals because vaccine protection varies over time.

After this basic overview of the study, Izurieta went through the draft decision aid (see Figure 9-5 later in this chapter) and discussed each question included in the decision aid.

What Are the Clinical and Epidemiologic Justifications for the Comparator Selected (and the Margin, If Applicable)?

Despite previous clinical trials, questions regarding the effectiveness of the vaccine and the duration of effectiveness lingered, said Izurieta. This study demonstrated effectiveness, and was also among the first studies to examine postherpetic neuralgia and hospitalization for herpes zoster, he said.

Where Has the Study Been Registered Prior to Initiation (e.g., ClinicalTrials.gov)? If It Is a Regulatory Study, or a Study Initiated by a Regulatory Agency, Which Regulatory Agencies Have Examined the Protocol?

The study was not registered prior to implementation. As regulators, said Izurieta, the researchers worked within the team to prepare the protocol and analyze the data according to the prespecified protocol; they considered this to be appropriate for the time.

How Can Reporting Be Structured to Enable Replication by a Regulator or Another Research Team?

To facilitate replication, said Izurieta, the researchers published appendixes and supplementary material that included all of the covariates used, all of the codes used, and all of the analyses and subanalyses for the groups.

Does There Appear to Be Appropriate Balance Between the Treatment Cohorts After Matching/Weighting? After Matching or Weighting for Balance, Do the Analytic Cohorts Appear to Represent Clinically Meaningful Groups for Study (e.g., Has Utility or Generalizability Been Sacrificed)?

To achieve balance, Izurieta and his colleagues adopted an approach developed by Rubin and Thomas (2000) that combines propensity scores and Mahalanobis metric matching. This approach, said Izurieta, allowed adjustment for heterogeneity between variables and controls, using a broad list of covariates that are plausibly related to herpes zoster, while generating cohorts that are closely matched on a subset of key covariates. The researchers used propensity scoring to adjust for covariates, including demographic factors, socioeconomic conditions, health care usage characteristics, and frailty characteristics. Mahalanobis distance was used to match essential covariates, including age, gender, race, and low-income subsidy. The database that was used, said Izurieta, included about 1 million people who had been vaccinated, and about 7 million who had not been vaccinated. This disproportion allowed the researchers to use nearly all of the vaccinated individuals and to be very selective about choosing non-vaccinated individuals for one-to-one matching, said Izurieta. For each vaccine recipient, a control population was found whose propensity scores fell within an acceptable range. Among these beneficiaries, one vaccine recipient was matched to one control with the minimum Mahalanobis distance from the vaccine recipient. Researchers used standardized mean difference statistics and falsification outcomes in order to assess cohort balance.

Before matching, the two populations were different with respect to key covariates. For example, the unvaccinated population was slightly older, was more racially diverse, and was more likely to receive low-income subsidies. These differences resulted in variations in the propensity score distribution between groups, said Izurieta (see Figure 9-3). After using the two-step matching process with propensity scores and Mahalanobis distance, the propensity score distributions were "wonderfully consistent" for all covariates (see Figure 9-4).

Are There Specific Unmeasured Confounders Thought to Be Sufficiently Influential That They Might Alter the Statistical Inference from the Study? Is There a Supplemental Way to Measure These Confounders? If Not, Can Sensitivity Analyses Be Designed to Examine Their Potential Influence?

While the cohort was well matched on many important covariates, said Izurieta, the problem of unmeasured confounders still existed. For example, vaccine recipients might seek health care more readily than non-vaccinated individuals. To address this issue, the researchers conducted a secondary

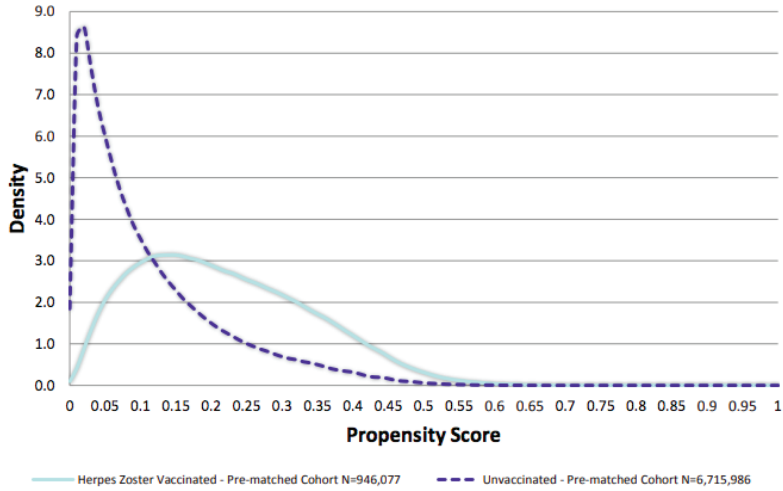


FIGURE 9-3 Prematched herpes zoster vaccinated and unvaccinated cohorts.
SOURCE: Izurieta presentation, July 17, 2018.

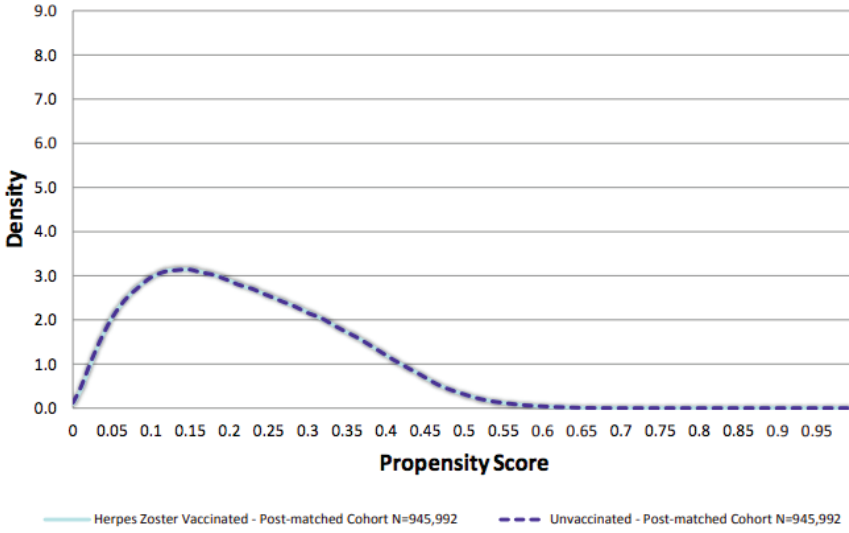


FIGURE 9-4 Postmatched herpes zoster vaccinated and unvaccinated cohorts.
SOURCE: Izurieta presentation, July 17, 2018.

analysis that compared the herpes zoster vaccine recipients with a cohort of people who received the pneumococcal vaccine (but not the herpes zoster vaccine). The researchers used 13 negative endpoints (e.g., thrombosis, hip fracture, hemorrhoids) in order to check the balance of these cohorts, and found no substantial difference between the two groups (Tseng et al., 2011).

After the herpes zoster vaccine study was published, said Izurieta, the researchers initiated a second project to use a Medicare survey to check the match of the herpes zoster study cohorts and to augment the cohort data using multiple imputation, something that had not been done previously. The Medicare Current Beneficiary Survey is distributed to a random sample of beneficiaries every year, and asks questions regarding health usage, vaccine history, education level, and frailty. The researchers compared approximately 900 herpes zoster vaccine recipients with about 900 non-vaccinated individuals from the survey respondents, and used multiple imputation to reanalyze the vaccine effectiveness after linking the data to the survey.

DISCUSSION: OBSERVATIONAL STUDIES AND RANDOMIZATION

Perspective from Some Workshop Participants: Continued Importance of Randomization

Supplementing knowledge through observational comparisons is useful, said Robert Califf. However, while he agreed with the basic principles for RWD analyses laid out by Schneeweiss during the second workshop, Califf asserted that "where possible, you should randomize." David Madigan, professor of statistics, and dean, Faculty of Arts and Sciences, Columbia University, agreed in part with Califf. However, he said, there are many disease areas and interventions for which evidence is needed, but there is never going to be a randomized trial. In these situations, evidence from observational studies is better than no evidence. Madigan said, "Randomization is an incredibly important tool that can provide very high-quality answers to important questions, but there are situations where we cannot randomize." Observational database studies are one way to answer these questions (e.g., for rare diseases), and so there is a need to "keep improving the way we do these studies," he added.

One way to incorporate randomization into RWD, said Califf, would be a system in which patients can consent to participating in randomized studies in the course of clinical practice. Califf said this type of randomization could be used in a situation in which two reasonable clinicians would make a different choice for no specific reason. For example, if a patient has hypertension and there is substantial uncertainty about which drug is more effective, the patient could be randomized to receive one or the other. The data resulting from this randomization, he said, could be used to fill

a chasm in evidence to answer questions considered most important by patients and clinicians, including which of several approved treatments is best for which patient.

Capturing the Uncertainty of Observational Trials

Every scientific study has some amount of uncertainty, said Madigan. While language such as "rely on" or "trust" is often used to describe a study, he said, this is the wrong mindset: The focus instead should be on characterizing and conveying the uncertainty in the study results. When a study is aimed at estimating the effect size of an intervention in some population, there are varying levels of accuracy depending on the study design. A well-conducted RCT can be quite accurate; an observational study that is not randomized, said Madigan, tends to be less accurate in general. However, the observational study may be "accurate enough" for the decisions that need to be made. Researchers have a duty to better capture and characterize all of the types of uncertainty in a study, he said.

Madigan described a statistical approach that he said can help illuminate the level of uncertainty in an observational study. Propensity scoring (which Izurieta used), he said, is a commonly used approach that attempts to estimate the probability that an individual would have an exposure (e.g., be treated with a drug, be a smoker). In an RCT, researchers know the exact likelihood that someone will be exposed (i.e., 50 percent if it is a 50–50 randomized trial), but in the real world, the likelihood of being exposed is due to a large number of factors such as socioeconomic status, access to health care, age, etc. Propensity scoring attempts to account for these factors by adjusting the likelihood of an exposure accordingly. Paul Rosenbaum (Rosenbaum, 2010) has extended this idea further, said Madigan, by attempting to account not just for known factors that affect probability of exposure, but also for the unknown factors that are not included in the propensity score. Rosenbaum accounts for these unknown factors by calculating how large a difference there would need to be in the probability that two people have an exposure in order for the results to not be statistically significant. An example of this approach, said Madigan, can be found using a 1954 matched-pair study on smoking (Hammond and Horn, 1954). This study matched each person who smoked heavily with one person who did not smoke and compared the outcomes. After adjusting for known covariates, there was a statistically significant large effect of smoking on lung cancer. Cornfield et al. (1959) calculated that in order for this effect to not be statistically significant, there would need to be an unknown factor that made it nine times more likely for a person to smoke. Madigan said "it is pretty hard to conceive of" an unknown factor that would increase a person's risk of smoking nine-fold. This "gamma

analysis," as Rosenbaum calls it, can give researchers confidence in findings from observational studies, even if the effect estimate is not as accurate as from an RCT. David Martin, associate director for Real-World Evidence Analytics at FDA's Center for Drug Evaluation and Research (CDER), gave a regulator's perspective on this approach, and said that this type of analysis could help move the conversation forward in a way that addresses concerns on both sides. On one side, it helps to mitigate the concerns of people who dismiss observational studies as being fraught with unmeasured confounding, and on the other, it forces observational researchers to truly examine, consider, and account for the effect of all potential variables.

Multiple Unreported Analyses

One issue with observational studies, said Califf, is the possibility of researchers conducting multiple, unreported analyses and then choosing only the most favorable ones to report. When there are an unlimited number of chances to obtain a certain result, the reported outcome is likely not a valid representation of the truth, he said. Although results in the published literature may point to a certain finding, Califf said, "We have no idea how many things were looked at that were never published." Observational data can now be analyzed with the push of a button because of automated programs; this ease of analysis increases the chances that the published results are a highly selective sample of all of the results compared with clinical trials. He added that there is a need for principles and systems to address this issue for observational studies. Schneeweiss agreed that many database studies are done "in the dark" and said that practices such as preregistration of database studies could help address the issue. Schneeweiss said that preregistration is particularly appropriate for confirmatory studies, and Martin agreed, noting that research that is performed to evaluate or confirm hypotheses needs transparency. Martin suggested that other countermeasures to multiple unreported analyses could be in-house replication by FDA, or support from FDA for complex data analysis.

Richard Platt told workshop participants about how Sentinel deals with multiple unreported analyses. Before a study is implemented, he said there is extensive discussion between FDA scientists and the researchers at Sentinel, and all of the specifications are decided in advance. Multiple analyses may still be done, he said, but each one is specified in advance and is shown in the report, so it is a transparent process. For regulatory decisions, said Martin, this type of prespecification is a necessary best practice. Another workshop participant added that there are technological approaches used to help ensure that there is a prespecified protocol and that the first analysis is reported as such (rather than running multiple analyses and picking the "best" one to report).

Gregory Simon noted that there are two potential solutions to the problems with multiple unreported analyses. The first would be to allow only a certain number of prespecified analyses. The second would be to allow analyses to be unlimited, but require that all analyses be reported and compared. Schneeweiss disagreed with this second approach, saying that "you just get a mess of data and you have no idea how to interpret it."

Several other individual speakers emphasized the importance of prespecification of the analytic plan; Mark van der Laan, professor of biostatistics and statistics at the University of California, Berkeley, called the problem of multiple unplanned analyses "the biggest problem in observational studies."

Reproducibility

Several individual participants noted problems related to nonreproducibility of database studies. One issue, said Madigan, is that researchers write custom code for a particular study rather than using validated tools. If the code is not available to other researchers, the study cannot be reproduced. Another issue, he said, is that authors are not always transparent about their design and analytic choices. For example, he cited a paper that said the study was "adjusted for age." However, no further details were given about how this was done, and Madigan had to contact the authors to understand that they had grouped the participants into 5-year ranges. If the participants were grouped by different age ranges, the data generated a different answer, Madigan said, adding, "The level of irreproducibility we are living with right now is unacceptable."

RWD analyses, said Schneeweiss, require multiple difficult choices about study design, dealing with non-standardized observations, and complex analytic methods. Because of this complexity, he emphasized that it is essential that researchers present studies with transparency and promote replicability. Decision makers—particularly regulators—need to be able to determine exactly how and when decisions were made in order to assess the quality of the research and replicate the analysis. "Transparent, structured reporting of complex methodology clarifies study validity for decision makers," Schneeweiss observed.

DECISION AID

The general issues discussed by individual workshop participants in the first and second workshops were used to develop a decision aid for the third workshop (see Figure 9-5). As with the other decision aids, the intention was to outline some questions to consider to make thoughtful choices in real-world evidence (RWE) study design. Participants at the third workshop

reflected on these questions and offered feedback on the decision aid specifically (see Box 9-1) and throughout the course of their discussions.

PRESENTATIONS: OBSERVATIONAL STUDIES AND BIAS

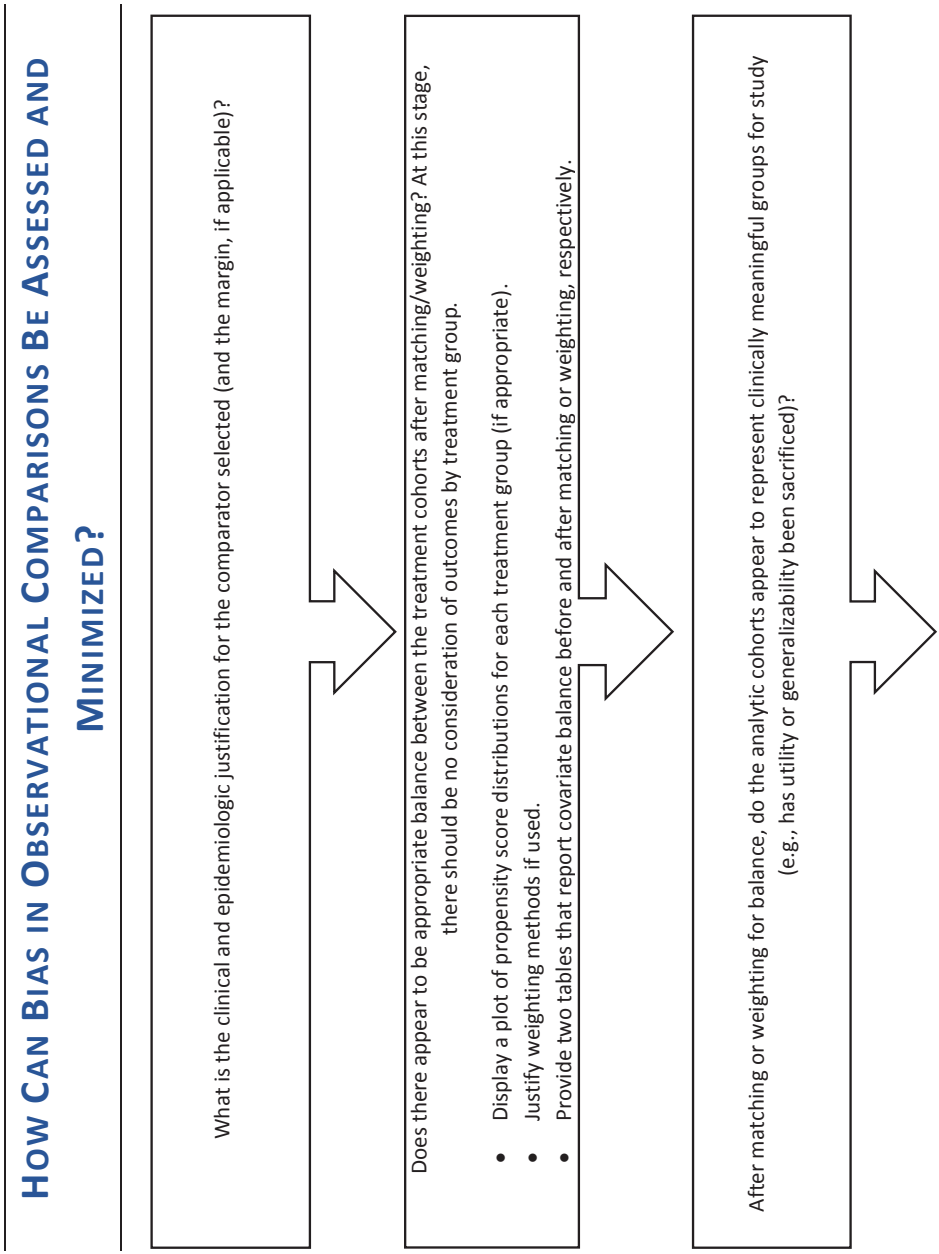
Replicating RCTs to Gain Confidence

One way to increase confidence in observational studies, said Jessica Franklin, assistant professor of medicine at Harvard Medical School, is to replicate results from RCTs using RWE. If RCT results can be consistently replicated across a range of clinical questions, this "gives us confidence going forward into new clinical questions," she said. Franklin and Schneeweiss are working on a project that will replicate a number of RCTs using RWD sources. Thirty RCTs have been selected, and the researchers are setting up the protocol for conducting the replications, she said. The process for determining which RCTs are appropriate for replicating with RWD, she said, is similar to the process that Schneeweiss outlined for determining whether a question can be appropriately answered with RWD. First, the setting and data quality must be assessed for the specific research question. Next, a statistical analysis plan is drafted, and initial analyses are conducted to test feasibility and validity. If the study passes these initial steps, the study is registered, and the analyses are specified before being conducted. Once results are reported, additional analyses can be conducted if necessary or appropriate. This process could serve as a model for how to use RWE to inform policy, said Franklin.

Marc Berger added that replicating RCTs can be used to identify fit-for-purpose datasets. If a dataset and an RCT generate substantially similar results, other analyses that use the same dataset "should be more credible," he said.

Predictive Analytics and Machine Learning

Javier Jimenez, vice president and global head for Real-World Evidence and Clinical Outcomes at Sanofi, suggested that new tools such as machine learning and predictive modeling may be useful for analyzing RWD. In addition to other uses, machine learning and predictive modeling present an opportunity to evaluate unmeasured confounders through proxies from other information that has been collected, he said. Jimenez presented a study on insulin in which predictive analytics were used to build models that evaluated the probability of an outcome for different populations, based on information in the Optum Database. The predictive models were based on individual variables as well as the interactions of all the variables, said Jimenez. The models were applied to the overall population in order to understand the



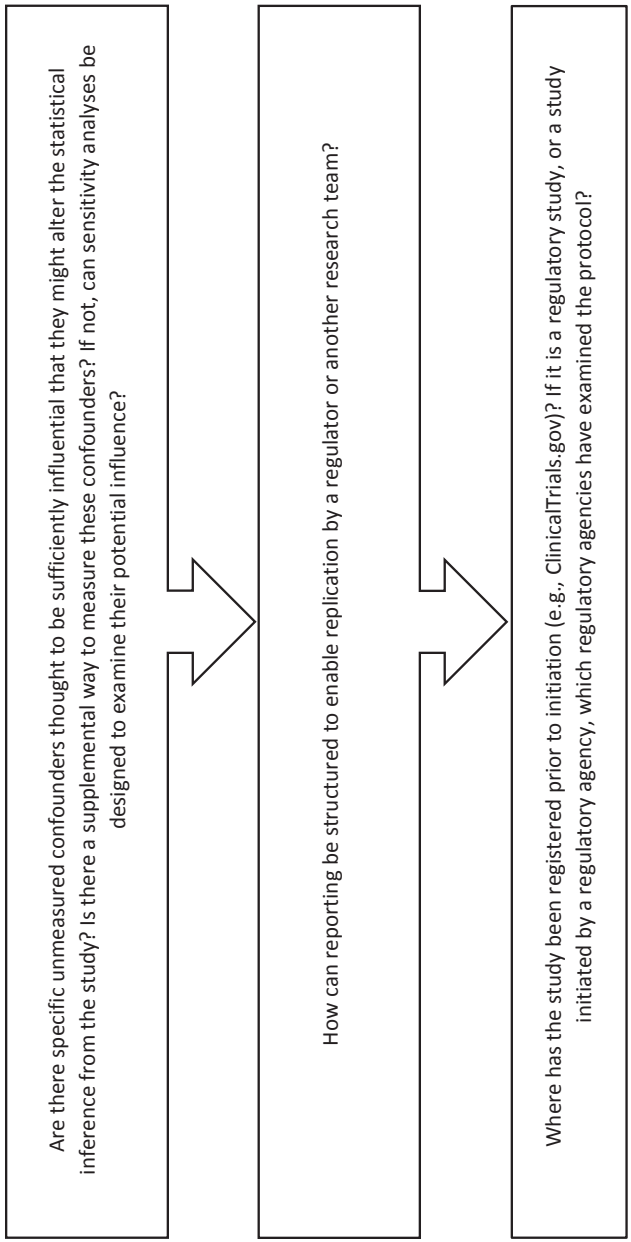


FIGURE 9-5 Decision aid on questions to consider to assess and minimize bias in non-randomized observational comparisons.
NOTE: This decision aid was drafted by some individual workshop participants based on the discussions of individual workshop participants at the first and second workshops in the real-world evidence series. The questions raised are those of the individual participants and do not necessarily represent the views of all workshop participants; the planning committee; or the National Academies of Sciences, Engineering, and Medicine, and the figure should not be construed as reflecting any group consensus.
SOURCE: Simon presentation, July 18, 2018.

BOX 9-1
Feedback on the Decision Aid as Discussed by Individual Workshop Participants

Several workshop participants offered specific feedback on the decision aid "How Can Bias in Observational Comparisons Be Assessed and Minimized?":

- The clinical context and the decision to be made should be included in the decision aid because they are critical to the choices that a researcher makes in the design of an observational study. (Daniel)
- Incorporating common methodological issues, most of which are entirely avoidable, could improve the decision aid. Examples of common mistakes include immortal person-time, overadjustment, and using an inappropriate comparator (Patorno et al., 2018). (Franklin)
- The decision aid itself is overly simplistic; the aid is tailored to very specific methods (e.g., propensity scores) when in reality, a number of different methods can be used. (van der Laan)

expected outcome if all patients were to use the particular insulin product. The results of this model, he said, were consistent with the results from a real-life study. In addition, the analysis revealed a particular subgroup that benefited more from the product. This model, said Jimenez, could be used to predict the probability of a specific outcome for a patient, using all of the available information.

van der Laan outlined another modern approach for conducting observational studies using computer systems to learn from data, known as "targeted machine learning." This approach, he said, is always based on a roadmap of causal inference, with defined steps that are followed for every analysis:

- The first step is understanding the question of interest from a causal perspective—that is, what are the outcomes, interventions, and other variables of interest—and developing a causal model describing the causal relations among the variables. The causal model allows a researcher to define the counterfactual data that would have been seen under a particular intervention on the intervention variables and defines the causal question of interest.
- Second, the researcher must determine what observed data are available and link them to the underlying counterfactual data that define the causal question of interest.
- Third, researchers determine whether causation can be established from the available data. To answer this, researchers use mathematical techniques to establish identification of the answer to the

causal question from the observed data distribution under a specified set of causal assumptions. These underlying assumptions generally cannot be tested. Any such identification result defines an estimand.
- Next, the researcher must commit to an estimand that "best" approximates the desired causal quantity and develop an a priori specified estimator and method of inference. It is essential that this be a priori specified in order for the research to be reproducible and transparent.
- The final step is sensitivity analysis to establish how confidence intervals and p-values change under different levels of assumed discrepancy of the estimand and the desired causal quantity, due to unmeasured confounding or other violations of the non-testable assumptions.

van der Laan described an approach called "targeted learning," which combines causal modeling, state-of-the-art machine learning, and deep statistical learning to get more precise answers for causal questions of interest, while providing formal statistical inference in terms of confidence intervals and p-values. Targeted learning is a technique to minimize estimation bias and to maximize precision in observational studies.

Regulatory Perspective on Observational Studies for Drugs

Nicole Gormley, clinical team leader within the Division of Hematology Products at CDER, said although the 21st Century Cures Act has put new focus on RWD and RWE to expand and expedite drug development, the "evidentiary criteria and standard really doesn't change." To approve a drug label, FDA needs a "demonstration of substantial evidence of efficacy with adequate demonstration of safety to enable the safe and effective use of the product," Gormley said. RWE can serve as primary or supportive evidence when FDA evaluates a product for approval, and there are specific aspects that regulators would consider to evaluate evidence, she said:

- First, data should be relevant for the proposed indication for the product; that is, the data should represent the population of interest and the setting in which the product would be used. For example, in an RCT, there are strict inclusion and exclusion criteria that define the patient population. In RWD, by contrast, the patient's inclusion in a treatment group is due to multiple factors, which can introduce bias.
- Second, a critical factor is that the outcomes being measured are clear and well assessed. In some disease settings, outcomes are relatively easy to assess (e.g., survival rates). Other endpoints may be

more difficult to assess, particularly in the real world. For example, an outcome of "attaining transfusion independence" is easy to assess in theory through the absence or presence of transfusions. However, in practice, some patients may be receiving transfusions at other centers that are not captured in the data, or data may be missing for other reasons.

- A third aspect of RWE that regulators would consider is the methods used to collect the data. Some methods give more confidence in the data collected, including well-designed protocols that minimize bias, account for confounders, and mitigate the impact of missing data.
- Finally, the statistical analyses applied to the data should be "robust and of significant rigor." While RWE has the potential to expedite drug development and complement the evidence from RCTs, it is important to ensure that RWE is collected and analyzed in a way that minimizes bias and increases reliability.

Gormley emphasized that when stakeholders are considering using RWE for regulatory purposes, they should engage in dialogue with FDA early in the process.

Regulatory Perspective on Observational Studies for Devices

Premarket observational studies are more common in devices than in drugs, said Heng Li, mathematical statistician at FDA's Center for Devices and Radiological Health (CDRH). This is due in part to a difference in the evidentiary standards for devices; for drugs, Li summarized that FDA requires "substantial evidence from well-controlled investigations." For devices, however, FDA requires "reasonable assurance based on valid scientific evidence," Li said. This evidence can come from well-controlled investigations as well as partially controlled studies, studies without matched controls, or well-documented case histories conducted by qualified experts. Premarket observational studies for devices generally use prospective enrollment for the treatment arm, and use an RWD source as a concurrent or historical control, said Li.

Statisticians at CDRH recently developed a streamlined procedure for designing premarket observational studies, said Li. This procedure uses propensity score methodology to balance baseline covariates, and uses an "outcome-free" design principle. This principle requires the propensity score development and assessment of covariate balance to be performed without knowledge of any outcome data. This procedure has two stages, he said:

- The first stage involves specifying a comprehensive list of baseline covariates, choosing an appropriate data source for the control group, identifying an independent statistician, and estimating the sample size.
- The second stage begins after the patients have been enrolled and all of the baseline covariate data have been collected, cleaned, and locked. In this stage, the independent statistician estimates the propensity score, performs matching or weighting, assesses baseline covariate balance, and finalizes the sample size and statistical analysis plan.

All of this work is done, said Li, without access to any outcome data. By blinding the statistician to the outcomes, a source of potential bias in the analysis is eliminated. This procedure has been implemented successfully in the premarket space for devices, he said.

DISCUSSION: THE FUTURE OF OBSERVATIONAL STUDIES

Context of the Decision

The choices in design and analysis of an observational study, said Gregory Daniel, depend on the context in which a decision is being made. Is it a chronic disease or is it a rare disease? What is already known about the safety and efficacy of a product? What is the expected treatment effect? What is the regulatory question at hand? For example, is it a brand new indication or an extension of the label? Gormley agreed, and said that evidence from observational studies needs to be examined on a "case-by-case basis." The context is enormously important, she said; for example, the standard of evidence may be higher for a situation with the potential for serious, life-threatening illness.

Best Practices for Observational Studies

While the questions on the decision aid are useful, it is a "very hard task" to dictate exactly how to do design or analysis, said Franklin. For any given clinical question, she said, the methods are going to vary. It can be useful to give people examples of practices that have worked in the past or to point out some of the common mistakes in observational studies, she said. For example, she pointed to Schneeweiss's presentation about which characteristics lead to a more valid observational study: an active comparator, a new user design, well-specified outcomes, and data sources with good longitudinal exposure measurement. In addition to these, propensity scores and sensitivity analyses can be useful, she said.

Berger agreed that there is a need for adoption of best practices in observational studies. In clinical trials, he said, there are rules and structures that enforce best practices for study design and conduct. For example, in RCTs, hypotheses must be registered, and if a hypothesis changes, this needs to be reported. Berger advocated for similar requirements and practices for observational studies: "We need to bring observational data up to the same level of scrutiny as we have for RCTs" before discussing when and whether RWD are fit for purpose.

Schneeweiss emphasized that the reliability of an observational study is driven by the underlying data. If the data are not fit for answering the question, the results will not be reliable. Unfortunately, he said, this is an ongoing process because "there is no one single dataset that will be fit" for all research questions.

10

Looking Ahead

Key Messages Identified by Individual Workshop Participants

- Health technology assessment relies on understanding the comparative effectiveness of new treatments; real-world evidence (RWE) can play a role in supplementing evidence from randomized clinical trials. (Jonsson)
- Using RWE can potentially engage patients more deeply in their care and in research, particularly with increases in usage of mobile technology and patients' ability to aggregate and share data about their own health. (Stem)
- Supporting a patient-centric shift in health research and care may require rethinking legislation and incentives, and forming new types of partnerships. (Stem)
- Newer clinical trial designs, such as adaptive designs, platform trials, or incorporating RWE, have the potential to significantly reduce cost and time investments required for medical product development. (Levy)
- The U.S. Food and Drug Administration's (FDA's) Center for Drug Evaluation and Research (CDER) uses RWE to routinely support postmarketing safety evaluation and to a limited extent for effectiveness in certain rare diseases, including oncology diseases; CDER's experience with Sentinel and support of several demonstration projects can inform policies going forward. (Corrigan-Curay)

- FDA's Center for Biologics Evaluation and Research uses several population-based data systems to conduct RWE safety and effectiveness studies, including a new program, Biologics Effectiveness and Safety (BEST), which is designed to build data infrastructure, tools, and expertise. (Anderson)
- FDA's Center for Devices and Radiological Health (CDRH) routinely uses RWE in its product evaluations for both pre/postmarket decisions. It has also started two programs to combine registry data with other forms of real-world data to better address regulatory needs. (Shuren)
- CDRH released guidance on use of RWE in 2017, which pointed to relevance and reliability as two critical considerations in evaluating RWE. (Shuren)
- All three medical product centers at FDA are interested in using RWE, but would like to have more experience with regulatory applications of the evidence and acknowledge that evidence for regulatory purposes is necessarily different from that suitable for other purposes. (Anderson, Corrigan-Curay, Shuren)

The workshops' discussions on real-world evidence (RWE) concluded with sessions about how RWE can be used to improve different facets of the health care system, including any potential to improve health technology assessment, to transform product research and development, or to bring patients in as partners for research. Workshop participants also discussed how—based on U.S. Food and Drug Administration (FDA) input—RWE can inform regulatory decisions for biologics, drugs, and devices in the United States and abroad.

REAL-WORLD EVIDENCE TO IMPROVE HEALTH TECHNOLOGY ASSESSMENT

Health technology assessment has a problem, said Pall Jonsson. The function of the UK National Institute for Health and Care Excellence (NICE) is to understand the comparative effectiveness and comparative cost-effectiveness of new treatments compared with standard practice. However, obtaining data suited for health technology assessment is becoming increasingly difficult. For example, new treatments in orphan diseases are becoming available, and the treatments are not supported by large randomized controlled trials (RCTs) that are traditionally used in assessments. Drugs are receiving market authorization based on single-arm trials, particularly in orphan diseases or areas of unmet needs. The lack of head-to-head trial data on these products makes them difficult to assess,

he said. Some drugs are receiving accelerated approval through new regulatory mechanisms. This acceleration means the products are getting to patients quicker, but it also means the data are less mature and there is more reliance on observational data. All of these factors combined, said Jonsson, mean there is "increased uncertainty around the decisions that we have to make."

RWE has a role to play in providing data for comparative assessments, said Jonsson. While NICE's assessment framework traditionally relies on RCTs, real-world data (RWD) sources may provide more useful data in some cases. For example, RCTs have narrow inclusion criteria, which reduces evidence about how a product works in patients with comorbidities or with more severe diseases. RCTs are time limited, which restricts the ability to gather evidence on long-term effectiveness, particularly for patients with chronic disease. Comparators in RCTs sometimes do not reflect the local practice or clinical practice, which makes it difficult to assess real-life comparative effectiveness of a product, he said. Of course, he added, using RWE also has challenges, including limited availability of data at time of assessment, potential for bias, poor quality or missing data, and data sources that are not established for research purposes. The acceptability of RWE varies across Europe; different countries have different views on the importance and credibility of RWE and the potential value of RWE in the future (see Figure 10-1).

To address the potential use of RWE, as well as the challenges involved, the GetReal[1] project was established approximately 4 years ago by the Innovative Medicines Initiative (IMI) in Europe, said Jonsson. IMI brought together stakeholders—including regulators, payers, patients, assessors, clinicians, and drug developers—to identify issues with current evidence-generation practices, and to explore how RWE might be useful for key decision makers.

The stakeholders involved in GetReal identified several pressing needs in the RWE space, said Jonsson, including

- Integrity, quality, access, and privacy protection of RWD sources;
- Guidance on RWE study design, evidence synthesis, and interpretation in decision making;
- Training and education in RWE; and
- Broader involvement of stakeholders in RWE generation and use of RWD.

A key finding of GetReal, said Jonsson, was that more attention should be paid to the "whole journey" of RWE, from designing studies to imple-

[1] See http://www.imi-getreal.eu (accessed November 4, 2018).

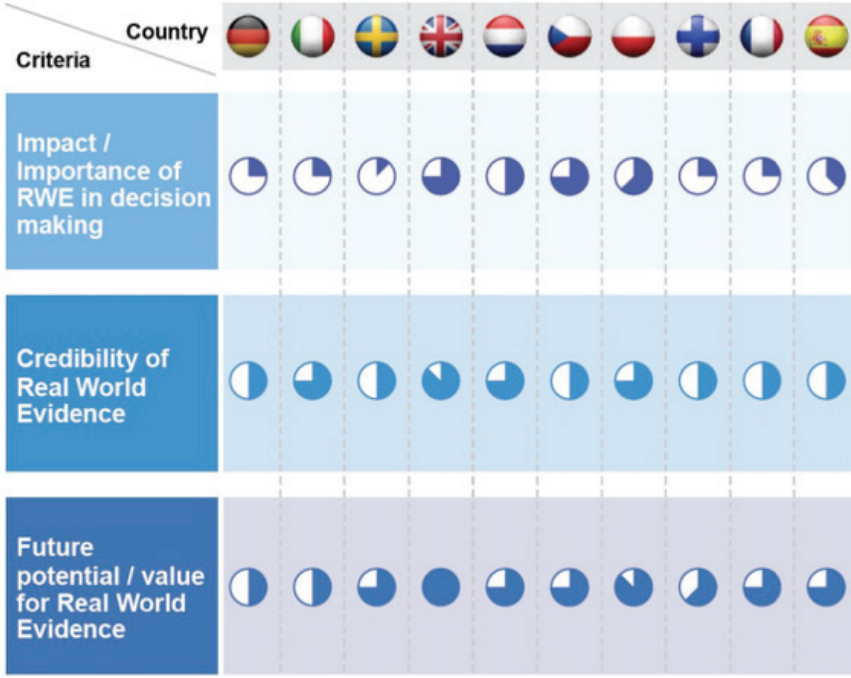


FIGURE 10-1 Acceptability of real-world evidence (RWE) across Europe.
NOTE: The shaded portions of the circles indicate the value and importance placed on RWE in each country (indicated by flag across the top).
SOURCES: Jonsson presentation, July 17, 2018; Gill et al., 2016.

menting and analyzing studies. To focus on the whole journey, GetReal has received additional funding to create a sustainable, self-funded entity to continue this work in its next iteration. The next generation of GetReal will continue to drive international consensus and use of RWE in decision making, provide tools to deliver high-quality RWE, and provide the education and training required to generate and use RWE. International thought leaders within the entity will "act as ambassadors" for the use of RWE by broadly engaging with stakeholders to drive debate and facilitate uptake of best practices.

Jonsson provided three examples of how the GetReal initiative has affected the work of NICE in the past several years. Jonsson noted that NICE is a unique stakeholder in that it is not a regulator or a payer, but sits "somewhere in between" and provides guidance to a broad range of stakeholders, including health care, public health, and social care organizations. The approach that NICE takes in regard to RWE, said Jonsson,

needs to be compatible across all of these areas and all of the different stakeholder needs.

The first example of NICE progression in incorporating RWE, said Jonsson, is the development of an internal health care and data analytics team that can help identify research questions, work with data owners, analyze data, and provide quality assurance of data. This team will be supported by an external advisory group that consists of members, data owners, experts, and industry, he said. Future partnerships may include working with the UK National Health Service to facilitate access to data and to understand how NICE guidance impacts the health care system, or working with health policy agencies to implement pilot projects.

The second example Jonsson discussed is a NICE statement that incorporates lessons learned from the GetReal project. The manual describes a range of possible situations in which RWD can be used in the context of NICE guidelines, and is meant to encourage guideline developers to consider whether and when analysis of RWD could be used to support decision making. The manual "flags key areas where we think there is a role potentially for real-world data," Jonsson said. For example, applications might include addressing the efficacy–effectiveness gap, extrapolating treatment beyond the duration of clinical trials, and understanding the impact of treatments on the health care system.

Finally, NICE's Science Policy and Research team is prioritizing areas for methods development, said Jonsson. The team is engaging in research projects with partners in order to develop best practices for applying adjustment methods for confounding, explore the use of big data in health care decision making, and consider the use of advanced analytics and artificial intelligence for RWE analysis.

REAL-WORLD EVIDENCE TO TURN PATIENTS INTO PARTNERS

In RCTs, said Komathi Stem, founder and chief executive officer of monARC Bionetworks, patients are passive participants. Data generated by RCTs are clean, structured, and easy to analyze, but at the same time, are limited, expensive, time consuming, and not always generalizable. RWE presents an opportunity to change these dynamics, and to turn patients into partners, she said.

Stem discussed three major trends that are facilitating a shift toward a more patient-centered, real-world approach:

- First, the point of care is shifting from the clinic to the smartphone (see Figure 10-2). Smartphones can be used to track health information, conduct telemedicine visits, and facilitate communication between providers and patients. Ninety-five percent of Americans

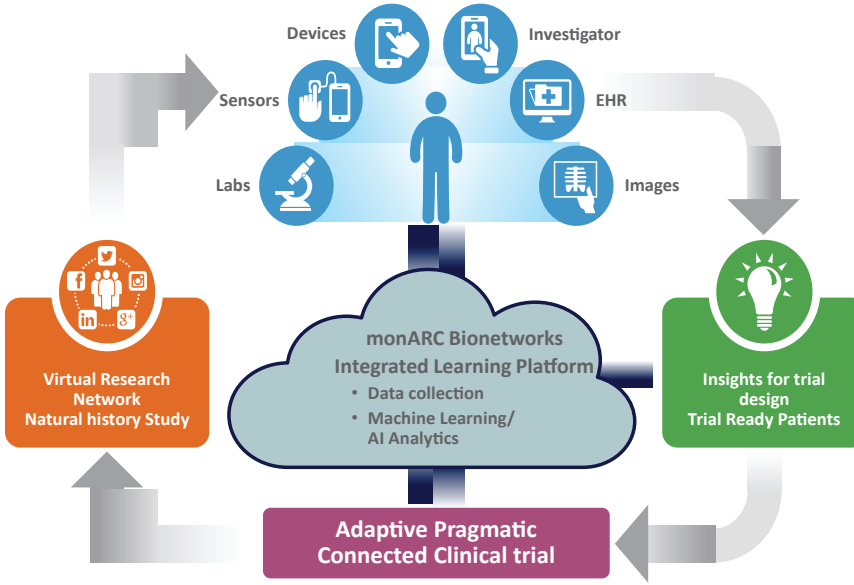


FIGURE 10-2 monARC Bionetworks Integrated Learning Platform.
NOTE: AI = artificial intelligence; EHR = electronic health record.
SOURCE: Stem presentation, July 17, 2018.

have a cell phone, 77 percent have a smartphone, and 20 percent have online access only through a smartphone (Pew Charitable Trusts, 2018). Seventy-five percent of physicians are using mobile tools in their clinic, and 75 percent of patients are willing to use mobile tools or telemedicine (Physicians Foundation, 2018). "This is the new clinical care setting, and if we continue to develop drugs in our traditional brick-and-mortar medium, we're going to be developing tools and drugs for a market that is completely different."

- Second, data are abundant, but highly compartmentalized and noisy. There are data from wearables, social media, smartphones, and electronic health records (EHRs), and these data are complex and variable. However, there are issues with interoperability, and with mitigating privacy concerns and proprietary obstacles to data sharing. These issues limit the potential of what can be done with the abundant data that exist.
- Third, there is a developing shift from data ownership to data access. Ultimately, being able to access, integrate, and use a variety of data streams is what delivers value. Players in other industries—

such as Amazon, Facebook, and Apple—have begun this shift, and health care should follow suit.

In this new world, patients will be the most important partners for research, Stem said. Patients are ideal aggregators of RWD; people already own a considerable amount of data from smartphones and other technologies, and they have and can grant access to their own health records. Patients are willing to share their data—96 percent have expressed a willingness to share, provided that security is protected and that the data will be used for a trustworthy purpose, she said. Patients can provide not just RWD, but also real-world insights. Stem said that patients can help develop meaningful endpoints for research, influence the design of research, and improve the decision-making process, similar to W. Benjamin Nowell's presentation in Chapter 7. Patients expect to be partners in this new world, said Stem, and expect to access continuous and updated information, such as personalized and dynamic drug labels on their smartphones.

To capitalize on these trends and to involve patients as partners, said Stem, monARC Bionetworks has developed an integrated RWD learning platform. monARC worked directly with patients to generate virtual research networks where patients can share data, and be engaged and recruited for research (see Figure 10-2). This system, said Stem, means that researchers can design better trials from the beginning and accelerate trial development by using data from patients and by recruiting and qualifying patients for whom data already exist in the system. Stem gave an example of an observational study on home spirometry that went from Institutional Review Board approval to published poster in 4 months.

These types of big changes to health research and practice, said Stem, will require big changes to incentives and legislation, as well as new partnerships. New incentives for data sharing across researchers and sponsors are needed, as well as incentives to develop novel endpoints. Clarity and improvement are needed in legislation about telemedicine, drug shipments, and research, she said. For example, the use of smartphones as a point of care or in research is highly limited by these laws. Finally, she said, there should be collaboration among industries such as social media, mobile devices, and artificial intelligence, and "bold partnerships" with patients to leverage the RWD and real-world insights they have to offer. Stem advocated for making processes simple so patients can participate in research, and for ensuring patients have access to or the ability to learn from the research to which they contribute. Echoing Nowell, Stem said patients have unique and invaluable insights; they "know a lot about their own condition, maybe more than we do, because they're living it."

REAL-WORLD EVIDENCE TO TRANSFORM RESEARCH AND DEVELOPMENT

"We are living in extraordinary times," said Elliott Levy, head of Development at Amgen Inc. There is a confluence of new technologies and the availability of new data sources that have the potential to transform medical product development. Over the past several decades, he said, medical technologies have been developed that have incredible potential. For example, T cells can be reprogrammed to express a chimeric antigen receptor, and these CAR-T cells have been used with response rates exceeding 80 percent in patients with highly refractory malignancies. Researchers have access to unprecedented amounts of data—such as genetic code—that can be used to personalize and target therapies. These technologies and new data sources are exciting, said Levy, who cautioned that "it will be expensive" to do the research and build the infrastructure necessary to fully exploit the opportunities presented. Unfortunately, the cost of developing a new medicine has been steadily increasing over the past few decades, while revenues of pharmaceutical companies have remained flat or risen slowly. Because of this, the challenge, said Levy, will be in finding a way to exploit new technologies and new data sources, without the luxury of having new funds to invest.

Several factors increase the cost of developing drugs, Levy said. The failure rate of new drug candidates is high, with only about 1 in 10 reaching the market (DiMasi et al., 2016). Levy said this rate suggests that despite a growing understanding of human biology, the research community's "ability to identify targets outstrips our ability to validate and confirm their importance." The data requirements for new products are increasing, with a growing demand for active comparator data, patient-reported outcomes, and long-term safety follow-up data. Finally, the most burdensome part of drug development is the cost of collecting the data in the physician's office, he said, a cost that continues to rise quickly.

In addition to the cost of developing new drugs, there is a significant cost to maintaining drugs that are on the market, Levy observed. There are demands for postmarket safety data, data on how products affect the health system, and further investigation into extending the product into pediatric or other special populations. The costs of developing and maintaining products present a "formidable challenge to our ability to realize the promise of new technologies and new data sources, and bring to market truly transformative medicines," said Levy.

RWE can play a critical role in addressing this issue in a sustainable way, he said. Levy noted that traditional cost-cutting measures typically yield significant savings initially, but the savings rate diminishes over time. For example, if a company moves a data management unit to India, there

are significant initial savings, but no future decreases to this cost. The challenge is to identify ways to stop or even reverse the increase in drug development costs, and this may require "radical changes" in the design and methodology of clinical trials. Levy shared three potential approaches to reduce the cost burden of generating evidence.

One potential change is a shift to adaptive trial designs, which have the potential to increase the probability of making correct decisions with reduced sample size and cost, and on a shortened time line. In essence, adaptive designs allow researchers to "fail efficiently and succeed efficiently," Levy said. He shared a hypothetical example of a traditional design compared to a Bayesian adaptive design (see Figure 10-3). It showed that the adaptive design needed to enroll about 20 percent fewer subjects, and saved 4 to 6 months in the time needed to reach a decision on product efficacy or futility. This type of design, said Levy, means that such decisions can be reached more quickly, and patient exposure can be limited earlier if a product is found to be ineffective. Levy noted that adaptive designs are currently used extensively in Phase I, to some extent in Phase II, but only sparingly in Phase III, where the greatest costs are incurred.

A second promising approach, said Levy, is the platform trial, a concept mentioned during the first workshop by Janet Woodcock of FDA. This design allows more than one agent to be studied in a single trial, or a single agent can be studied in multiple disease types. For example, two anticancer treatments that would normally be studied in two parallel trials (see Figure 10-4) could potentially be combined into a single platform trial that has a single master protocol, consolidated start-up procedures, and a combined comparator arm (see Figure 10-5). Consolidating two trials into one, said Levy, saves about 25 percent in terms of cost, number of patients randomized, and trial duration. A more complex platform trial (e.g., combining a larger number of experimental agents) could lead to even greater savings, he said.

The third approach for reducing costs, said Levy, is to use RWE to augment RCTs or perhaps for label expansions on existing products. RWE can be incorporated into trial design in a number of ways, he said. For example, a simplified pragmatic trial could randomize patients and collect data on baseline status, adverse events, and key endpoints, but rely on information collected in the usual course of care for all other data. By reducing the frequency of study visits and the number of laboratory assessments, this can reduce the cost per patient by about two-thirds, said Levy. In some cases, RCT evidence could be replaced entirely by RWE. If a product is already known to be safe and effective, but evidence is needed for marketing the product to a new geography or a new indication, a retrospective RWE study can use data that already exist, and reduce the time and cost significantly. Levy estimated that replacing RCTs with RWE would reduce the cost of a

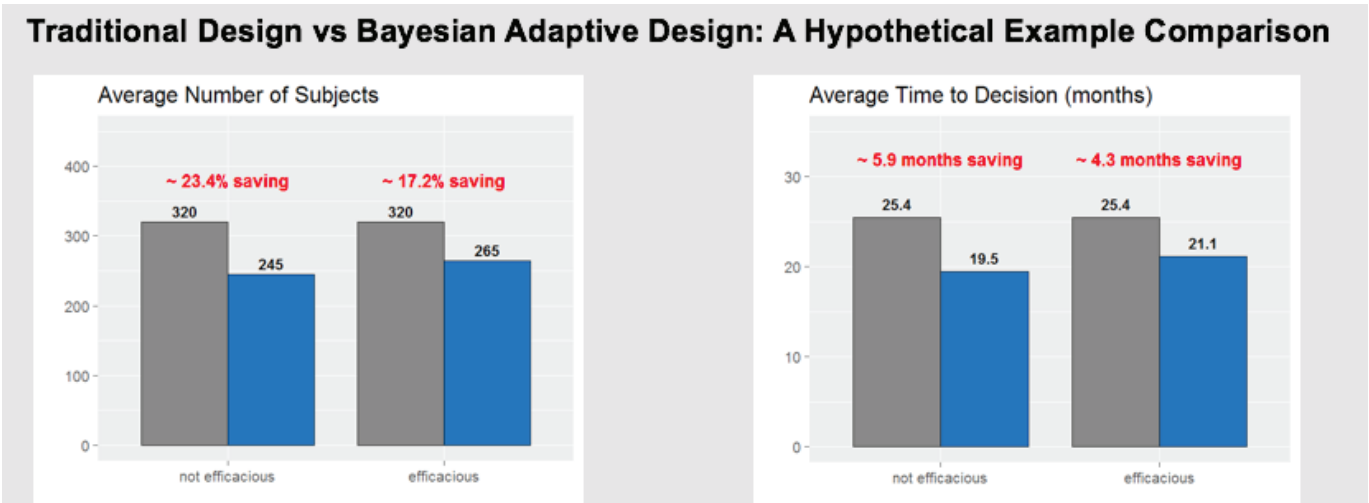


FIGURE 10-3 Hypothetical example of traditional design versus Bayesian adaptive design.
SOURCE: Levy presentation, July 17, 2018.

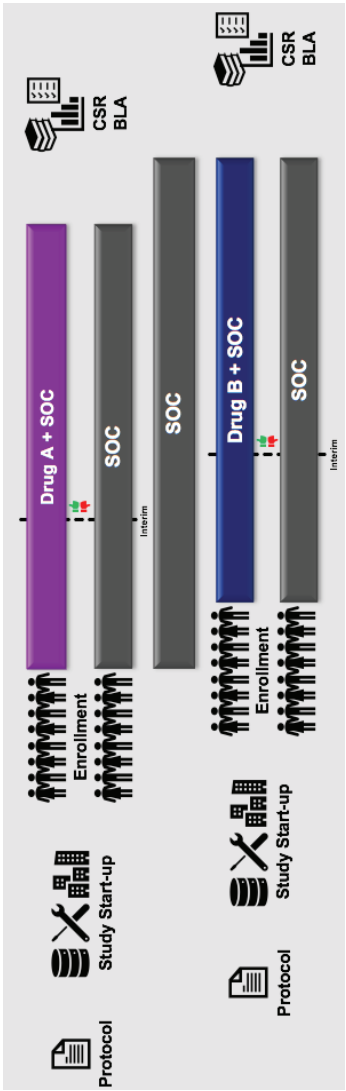


FIGURE 10-4 Standard separate trials for drugs A and B.
NOTE: BLA = biologics license application; CSR = clinical study report; SOC = standard of care.
SOURCE: Levy presentation, July 17, 2018.

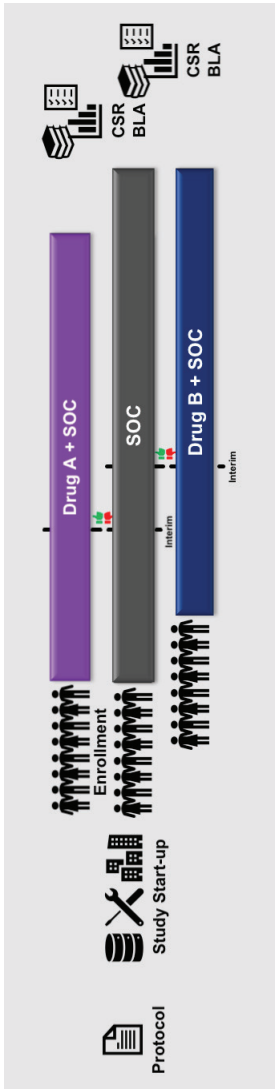


FIGURE 10-5 Combined platform trial for drugs A and B.
NOTE: BLA = biologics license application; CSR = clinical study report; SOC = standard of care.
SOURCE: Levy presentation, July 17, 2018.

study by up to 90 percent and the duration of the trial by up to 80 percent. Berger added that another way to incorporate RWE is to nest trials within registries or networks where data are already being collected on a systematic basis.

While all of these approaches—adaptive design, platform trials, and using RWE to augment trials or for label expansions—can generate significant cost and time savings, said Levy, RWE has the potential for the most radical cost savings (see Table 10-1). The conventional approaches that drug development companies have pursued over the past decade for improving cost efficiency have already been fully exploited and will not yield meaningful additional gains, said Levy, so other changes will be necessary. Greater improvements in trial costs and time lines will require radical reductions in the amount of data collected specifically for the trial, instead replacing it with the RWD collected in the process of providing care, he said. Levy noted that RWE is not appropriate for every setting, and that RCTs will continue to play an important role. However, the overall cost of developing drugs could be reduced by using new trial designs and incorporating RWE whenever feasible and appropriate. Hernandez emphasized that the timing and scope of a study can greatly impact the cost of the study. A trial that aims to assess a product early in its life cycle likely requires significant data collection and must contend with many adverse events, concomitant therapies, and analysis of millions of data points. By contrast, a trial that assesses a product later in its life cycle can be more streamlined, collect fewer data, and cost significantly less.

TABLE 10-1 Hypothetical Examples of Cost Savings

Example	Savings
Risk-based monitoring	~5 percent of total clinical trial spend
Adaptive trials designs	Average saving of ~20 percent study subjects and up to 25 percent reduction in trial timeline
Platform trials	~20–40 percent saving of trial costs and reduction of timeline by ~25 percent
Highly simplified trials	~50 percent reduction in per-site and per-patient costs
Real-world evidence (in lieu of randomized controlled trial)	Up to ~90 percent reduction in total trial cost and ~75 percent reduction in trial timeline

SOURCE: Levy presentation, July 17, 2018.

REAL-WORLD EVIDENCE TO INFORM REGULATORY DECISIONS

In this session of the third workshop, participants first heard about the European perspective on RWE, and then heard from three directors at FDA about the use of RWE for making regulatory decisions on biologics, drugs, and devices.

European Perspective

The field of RWE is active in Europe in several areas, said Alasdair Breckenridge, emeritus professor of clinical pharmacology at the University of Liverpool. Activity is still mainly in the traditional area of safety, but increasingly in dosing, drug–drug interactions, sequence of therapies, expansion to subpopulations, and new indications. A particularly interesting new use, he said, is in applying RWE-generation techniques to the data collected to meet postmarketing requirements in order to suggest new indications. Breckenridge told workshop participants about RWE work being done by three European bodies.

The UK Academy of Medical Sciences has held two workshops on the topics of RWD and RWE, he said. The first workshop was held in September 2015, and aimed to explore the acceptability of RWE in regulatory and health technology assessment decision making, to address the challenges, and to suggest practical steps to address them. The workshop brought together a number of stakeholders, including the European Medicines Agency (EMA), FDA, NICE, and the regulated industry. This workshop reached the conclusion, said Breckenridge, that the role of RWE in regulatory decision making remains to be defined. While RWE is being used, there is a need for better definitions and standards, he said. The second workshop was held in early 2018, and participants at the workshop concluded that compared with the United States, progress in Europe on RWE was limited and many of the challenges identified in 2015 remained unresolved. The workshop also explored definitions, and defined RWD as a "subset of big data relating to patient health status, delivery of routine health care, collected from a variety of sources [including] electronic health records, claims, product and disease registries, and social media." RWE, said Breckenridge, was considered to be evidence drawn from RWD through the application of research methods. The workshop participants suggested there should be a way to define "regulatory-grade RWE." Regulatory-grade RWE, said Breckenridge, would meet five criteria:

1. Define the scientific question to be answered.
2. Identify study design.
3. Be specific in terms of the RWD used.

4. Be rigorous in data standards and analytic methods.
5. Comply with regulatory standards.

EMA, said Breckenridge, is very active in the pharmacovigilance field, routinely using RWD for safety monitoring. The EudraVigilance system, operated by EMA, received 1 million safety reports in 2016, of which 2,000 signals were detected and 48 were validated. Breckenridge discussed EMA's involvement in two efficacy studies: the Salford Lung Studies (discussed in detail in Chapter 3) and the Phase II single-arm study on Zalmoxis, an immunotherapy for high-risk hematological malignancies, which used historical controls from a transplantation registry. Breckenridge noted that EMA gave conditional marketing authorization for this product, but asked for postauthorization efficacy and safety studies. EMA's biggest contribution, said Breckenridge, has been its work on the Medicines Adaptive Pathways to Patients (MAPPs, previously known as Adaptive Licensing). The MAPP program is a prospectively planned, adaptive approach to give early access to important new medicines for patients with unmet needs, with lower premarket evidence requirements. The program shifts the burden of evidence from the pre- to postmarket space, and emphasizes postauthorization efficacy and safety studies. MAPP uses the existing European Union legal framework, he said, and requires the ongoing involvement of the company, regulators, health technology assessment experts, payers, and patients. As of July 2016, there were 18 accepted pilot projects, all of which included plans for the use of RWD that went beyond the traditional use of registries for pharmacovigilance. Breckenridge noted, however, that most of these projects were "vague in terms of the purpose of collecting RWD to supplement RCTs."

IMI (discussed earlier in this chapter by Jonsson) is a public–private consortium launched in 2008, which has a budget of approximately 3 billion euros and currently supports 50 projects, said Breckenridge. He briefly outlined three IMI-sponsored projects of interest, in addition to IMI's GetReal project, presented by Jonsson earlier in this chapter. IMI PROTECT (Pharmacoepidemiological Research Outcomes of Therapeutics by European Consortium) is a project designed to strengthen the "monitoring of benefit/risk" of treatments in Europe, with the involvement of patients and the public. The project addressed the limitations of the current methods for monitoring, and an evaluation of the project found that it had met its objectives in terms of signal detection and evaluation, and the use of routine pharmacovigilance. IMI ADAPT SMART is a project to facilitate and accelerate the availability of MAPP's related activities. It was established in 2017 for a 2-year period to develop next-generation vaccines and medicines, to tackle Europe's growing health care challenges, and to ensure future competitiveness of Europe's pharmaceutical industry, he said. Finally, IMI WEB RADR (Recognising

Adverse Drug Reactions) is a project that developed a mobile application for patients and health care professionals to report suspected adverse events. Reports received from the application are compared to reports received via established reporting schemes in order to evaluate whether the application is an effective way to collect this information.

U.S. Food and Drug Administration: Drugs

FDA is "taking our obligations very seriously under 21st Century Cures," said Jacqueline Corrigan-Curay, director of the Office of Medical Policy at FDA's Center for Drug Evaluation and Research (CDER). This is not just because it is a requirement, she said, but "because we do want to make sure we are doing everything we can be doing to bring things efficiently to the market" while keeping strong evidence standards. RWD has been used a number of times as part of the premarket evaluation of drugs for rare diseases, said Corrigan-Curay, including examples of non-randomized, unblinded trials against historical controls (see Figure 10-6).

Drug	Indication	Status	Data
Lutathera *(lutetium 177 dotate)*	GEP-NET Gastropanc. Neuroendo tumors	Approved 2017	▪ Open label clinical trial ▪ **Analysis of 360 patients in an investigator sponsored, expanded access protocol** of 1214 patients*
Voraxaze *(glucarpidase)*	Treatment of MTX toxicity	Approved 2012	▪ Approval based on open-label, NIH **compassionate Use Protocol**
Uridine Triacetate	Treatment of 5 FU overdose	Approved 2015	▪ Two single-arm, open label expanded access trial of **135 patients compared to case history control**
Blincynto (Blinatumomab)	Treatment of Acute Lymphoblastic Leukemia	Approved 2014	▪ Single arm trial ▪ Reference for effect weighted analysis of patient level **data on chart review of 694 patients at EU and US study sites***
Carbaglu® (carglumic acid) Tablets	Treatment of NAGS deficiency	Approved 2010	▪ Retrospective, non-random, un-blinded **case series of 23 patients** compared to **historical control group**
Myozyme® (alglucosidase alfa)	Treatment of Pompe disease	Approved 2004	▪ Open-label, non-randomized study of 18 patients compared to **historical control group of 62 untreated patients**
Refludan®	Anti-coagulation in heparin-induced thrombocytopenia	Approved 1998	▪ Two non-randomized, open-label multicenter trials using **historical control comparator group from HIT Registry**

NOT EXHAUSTIVE

FIGURE 10-6 Use of real-world data in the evaluation of drugs for rare diseases.
NOTES: Bolded text = examples using real-world evidence. 5 FU = 5-fluorouracil; EU = European Union; GEP-NET = gastroenteropancreatic neuroendocrine tumor; HIT = heparin-induced thrombocytopenia; MTX = methotrexate; NAGS = N-acetylglutamate synthetase; NIH = National Institutes of Health; US = United States.
SOURCES: Corrigan-Curay presentation, July 18, 2018; * data from Gökbuget et al., 2016.

RWD has also been used in the postmarket space for some limited indication expansions; for example, in 2017, FDA expanded the indications for Kalydeco to include 23 new mutations in cystic fibrosis based on clinical and in vitro data. After approval, FDA asked for postmarketing observational study using the cystic fibrosis registry, she said.

In 2013, FDA released guidance on the best practices for conducting and reporting pharmacoepidemiologic safety studies using EHR data.[2] These practices can likely inform best practices in other areas of RWD analysis and RWE use, said Corrigan-Curay. For example, one goal of this guidance is to ensure that patients whose electronic health care data have been used in an outcomes analysis have actually experienced the event; it suggests practical steps such as ensuring that the code or algorithm has either been validated previously or that its predictive value was calculated, and describing the sensitivity of the outcome.

Claims data are a common source of RWD. Corrigan-Curay noted that even though use of claims data in RWE applications is well understood, systemic changes such as the switch from the *International Statistical Classification of Diseases and Related Health Problems, Ninth Revision* (ICD-9) to ICD-10 will create challenges for RWD-based research. ICD-10 contains 70,000 diagnoses, compared with 14,000 in ICD-9, and this transition has created some issues, she said. For example, a surveillance of hospitalizations with a diagnosis of opioid use disorder saw an uptick of about 14 percent when the transition happened; other systems observed a decrease in the likelihood of correctly reporting confounding comorbidities with the new system. In addition, published studies using different claims data sources are sometimes in conflict. Corrigan-Curay emphasized that the point was not to decide which study was correct, but rather to understand the reasons for differential results. These challenges will need to be dealt with so that data sources are valid and reliable.

To begin answering some of the questions on how to work with RWD for regulatory purposes, Corrigan-Curay discussed several demonstration projects that are being supported by FDA. One example was the OneSource[3] checklist project, which supports data collection in a way that will meet the needs of both researchers and practicing clinicians. Corrigan-Curay also mentioned the promise of data networks like Sentinel or the National Patient-Centered Clinical Research Network, and the potential of also using observational data networks in the future.

[2] See https://www.fda.gov/downloads/drugs/guidances/ucm243537.pdf (accessed January 4, 2019).

[3] See https://www.fda.gov/ScienceResearch/SpecialTopics/RegulatoryScience/ucm574079.htm (accessed January 4, 2019).

There is a wide spectrum of potential uses of RWE in clinical studies, said Corrigan-Curay. However, she emphasized a need to be thoughtful about adopting new methodologies and data sources because "as we adapt the tools and methods of traditional trials to real-world settings, we must consider the components of such trials that are critical to obtaining valid results and minimizing bias" (Sherman et al., 2016, p. 2294). Corrigan-Curay emphasized a need to build confidence and experience in using new data streams, new technologies, and new analytic methodologies for RWE, as well as building expertise and commitments from multiple stakeholders to realize the potential of RWE. Corrigan-Curay outlined the potential future of RWE, which could include

- Research fully embedded in care settings (no data are wasted);
- Integrated/connected systems throughout the entire health care continuum with feedback loops;
- Seamless and integrated auditing and quality control mechanisms;
- Flexible and linkable on-demand data aggregation from databases/registries;
- All stakeholders engaged, including potentially increased patient engagement with mobile technologies or data capture;
- Secured and traceable access and management of data (blockchain); and/or
- RWE continuously used to support decision-making processes.

U.S. Food and Drug Administration: Biologics

The 21st Century Cures Act is a main driver of the RWD and RWE initiatives at FDA, said Steven Anderson, director of the Office of Biostatistics and Epidemiology at Center for Biologics Evaluation and Research (CBER). CBER, which regulates vaccines, blood and blood products, tissue and tissue products, and cellular and gene therapy products, has undertaken several RWE initiatives in recent years, he said.

CBER uses a number of population-based data systems to conduct RWE safety and effectiveness studies, including a system being developed specifically for CBER called the Biologics Effectiveness and Safety (BEST) program. BEST is part of the Sentinel initiative, and was launched in September 2017. One goal of BEST, said Anderson, is to build data infrastructure and tools and develop expertise to conduct queries and studies of biologic products. The second goal is to automate adverse event reporting by using methods such as machine learning and natural language processing in order to mine adverse events related to biologics from EHRs and automatically submit them to FDA. The BEST program, said Anderson, will help CBER to better meet its regulatory needs by building "better,

faster, cheaper systems" to generate evidence about safety and effectiveness of biologic products. CBER also uses the Sentinel system (described in Chapter 3) to generate evidence on biologics; CBER has conducted dozens of safety assessments and more than 100 "rapid queries" to address safety questions, as well as a pilot study on vaccine effectiveness. Anderson gave two examples of RWE studies on safety of biologic products being successfully used for label changes or regulatory action. The first was a study on immune globulins and thrombotic events, and used data from the Centers for Medicare & Medicaid Services (CMS) database (Daniel et al., 2012). The second, discussed in Chapter 3, was a Sentinel study on rotavirus vaccines and intussusception.

Anderson highlighted CBER's use of RWE for real-time analysis. CBER uses the CMS claims data system to conduct near-real-time analysis of the annual flu vaccine and related adverse events such as Guillain-Barre syndrome, he said. One specific study that was performed in 2017–2018 was a rapid response effectiveness study of cell versus egg-based influenza vaccines. Using CMS data, FDA and CMS examined about 13 million vaccine doses, and found that cell-cultured vaccines were slightly more effective. These types of studies, said Anderson, can provide near-real-time information (within 4 to 6 weeks) to inform regulatory decisions concerning current and future influenza vaccines.

In addition to these activities, CBER is considering ways to collect patient input and information on patient preferences, said Anderson. In March 2018, the agency began soliciting proposals to collect patient input in five disease areas: sickle cell anemia, brittle diabetes, hemophilia, rheumatoid arthritis, and retinal dystrophy. CBER is also working with the Sentinel staff to develop a mobile app to collect patient input, said Anderson.

There are several challenges in using RWD for regulatory decision making, said Anderson, including bias, quality of data, missing data, and how well a patient's exposure and outcome can be captured. In addition, there are challenges in linking different forms of data, such as data from EHRs and registries.

U.S. Food and Drug Administration: Devices

FDA's Center for Devices and Radiological Health (CDRH) has been on "an 8-plus-year journey in our use of RWE," said CDRH Director Jeff Shuren. He discussed three major CDRH efforts regarding RWE.

First, in 2010, the Medical Device Epidemiology Network (MDEpiNet) was established with other stakeholders from government, industry, and academia. Since its inception, MDEpiNet partners have published more than 190 studies, said Shuren. MDEpiNet has developed registries, including coordinated registry networks. These networks, he said, address the fact

that some individual registries may not contain fit-for-purpose data, but when a registry is combined with other RWD sources (e.g., claims data), it may be fit for purpose. MDEpiNet has also worked to develop active surveillance methodologies, conducted studies exploring the utility of claims and EHR data, and worked on evidence synthesis through in silico modeling and other approaches. Several dozen projects are in the MDEpiNet pipeline, Shuren said, including developing tools to move clinical data from EHRs into the Women's Health Coordinated Registry Network[4]; implementing the Delta System for active surveillance in the transcatheter valve therapy registry and the cardioverter–defibrillator registry; and testing the capabilities of state-based claims.

In 2012, CDRH developed a strategy for a national system to address the limitations in the use of RWD, and to facilitate the systematic generation of RWE by a broad range of public and private entities. This effort, said Shuren, is called the National Evaluation System for health Technologies (NEST).[5] NEST will soon begin a series of demonstration projects with 11 data partners, with the purpose of test driving the systems' capabilities for addressing important device questions. At the time of the workshop, the NEST partners include approximately 150 hospitals and more than 3,000 outpatient clinics, which provide access to more than 469 million patient records. NEST has also established committees on methodology and data quality to develop standards and best practices.

CDRH's third effort in the RWE space, said Shuren, is its involvement in the International Medical Device Regulators Forum (IMDRF). IMDRF released a series of three principle documents that focused on the infrastructure, methods, and tools for assessing the usability of registries for regulatory decision making. There is currently a study under way that aims to test these principles through research on devices for ruptured abdominal-aortic aneurysms.

CDRH has been accepting and leveraging RWE as valid scientific evidence in support of both pre- and postmarket regulatory decisions for many years, said Shuren. Since 2015, there have been at least 50 regulatory decisions for which CDRH relied on RWE, he noted, including decisions about expanded labeling and new device approval. In 2017, CDRH released final guidance on the use of RWE in regulatory decisions to provide greater clarity about when RWE is "regulatory grade." The guidance pointed to two critical considerations in the evaluation of RWE: relevance and reliability.

Shuren gave an example of how RWE can be used in an iterative manner in the postmarket space. In 2011, FDA approved first-generation

[4] For more information about the Women's Health Coordinated Registry Network, see mdepinet.org/womens-health-crn (accessed November 2, 2018).

[5] See https://nestcc.org/demo-announcement (accessed January 4, 2019).

transcatheter aortic valve replacement (TAVR), being the 42nd country to do so, and worked with the American College of Cardiology, the Society of Thoracic Surgeons, and CMS to establish a registry. At the same time, CMS issued a national coverage determination that approved coverage for the FDA-approved indication for TAVR, with the requirement that data be entered into the registry. The coverage determination also stated that if FDA expanded the indication for TAVR, coverage would automatically be expanded. In 2017, FDA approved third-generation TAVR for intermediate-risk patients 19 days after the European Union granted a CE (European conformity) mark for a similar device, said Shuren, and approved it for mitral valve-in-valve indication—the first country to do so. This represents a "really dramatic change," he said, in FDA's ability to obtain data to answer questions that were not previously addressable, and to do so more quickly and at a lower cost. CDRH conducted a return-on-investment analysis for decisions like this that leverage a registry, and found that it cost approximately $25 million for 20 studies to support 22 FDA decisions, compared with $127 to $134 million to conduct these studies if there had been no registry. This cost savings, he said, does not consider that a quicker time to market results in lives saved and improved quality of life for patients, as well as additional economic benefits to the companies.

The use of RWE for making regulatory decisions about devices is not a thing that is "nice to have," it is a "need to have," said Shuren. RCTs have inherent limitations that restrict their utility, particularly for devices. For example, when devices are used in the real world, there is a learning curve for providers who are implanting the device. This information cannot be generated from RCTs, he said. In addition, RWE can inform subgroup analysis on race, gender, and other patient characteristics. RWE is necessary for assessing these types of technologies in the real world, particularly when they are used in a broad range of patients and providers.

CDRH has long been trying to apply a total product life cycle approach for devices, said Shuren, rather than making an artificial divide into pre- and postmarket spaces. However, CDRH is not optimally structured organizationally to do this, so there is a reorganization under way. Part of the reorganization will involve the creation of a new Office of Clinical Evidence and Analysis that brings together experts on clinical trials, epidemiology, RWE, biostatistics, and surveillance. The reason for this, said Shuren, is that ultimately, the source of the clinical data is irrelevant; the true key attributes instead are whether the data are relevant and reliable.

Shuren concluded with three "lessons learned" for workshop participants to carry forward into their work with RWE:

- "Don't let the perfect be the enemy of the good": A single RWD source may not always be perfectly fit for purpose, but it can be

incredibly useful when linked with other RWD sources. Other approaches include validating an RWD source by comparing the results generated from it with a previously conducted RCT, or to conduct separate analyses with separate RWD sources to "triangulate" results; if the results match then confidence in the result is much greater.

- Consider how data are collected: Clinical research needs to be incorporated into the workflow of routine clinical practice in a systematic way that ensures consistent, high-quality data.
- Do not underestimate the value of good data scientists: Evidence generation requires expertise and appropriate methodology to turn good data into good evidence.

Discussion

After the presentations on the regulatory perspective on RWE, Breckenridge moderated a broad-ranging discussion among the panelists and workshop participants. Corrigan-Curay began by noting that there is a "chicken and egg" problem with the use of RWE in the regulatory space: companies are waiting for FDA to release guidance, and FDA is waiting for companies to come forward to inform the guidance. She said that meetings and workshops, such as this series, are enormously helpful for identifying gaps that need more discussion, as well as areas where FDA may be able to move forward. Anderson concurred and noted that one of the biggest gaps is in ensuring data quality and reliability. For example, he said, depending on the source of the data (e.g., EHRs, claims data), there can be different biases introduced and there can be issues with fully and accurately capturing the necessary information. In addition, linking the various sources of data remains an enormous challenge. Anderson noted that there needs to be a higher standard for evidence used for regulatory decision making. He said that companies should engage in conversation with FDA early and often in order to improve the likelihood that their studies will produce this type of "regulatory-grade evidence." Shuren broadened this issue slightly, noting that evidence should be produced that is fit for purpose, whether that purpose is regulatory decision making or some other purpose such as clinical decision making. He added that FDA and other regulators do not look at RWE in a vacuum, but rather as part of the totality of the evidence from a variety of sources.

Breckenridge asked the panelists how the three Centers within FDA—CDER, CBER, and CDRH—are working together on the issue of RWE. Corrigan-Curay responded that while each Center may have different needs for RWE, communication among Centers is essential so that "we are all talking about the same thing across the agency," including harmonizing

terminology. Shuren added that while open communication and working together are essential, it is sometimes appropriate and even necessary for different organizations within FDA to take a different approach to RWE. He said one benefit of FDA's structure is that different parts of FDA can experiment with different approaches and the other organizations can learn from those experiences and adapt them to their own needs. The panelists noted that FDA is also working with EMA and other international regulators on the issue of RWE.

Gregory Simon asked if FDA will be moving away from solely being a "referee" of evidence produced by others to becoming a "starring player" by producing evidence, like in the Zostavax vaccine effectiveness study presented by Hector Izurieta (see Chapter 9). Corrigan-Curay responded that CDER will still primarily be recipients of evidence rather than data generators, but will try to work closely with stakeholders bringing evidence to FDA. Anderson noted that these types of studies are costly and take years to conduct, and that CBER would likely continue to "be the referee rather than [primarily generating] the data." Shuren responded that CDRH both generates data and has built infrastructure and partnerships that enable CDRH to use RWE for regulatory decisions. For example, said Shuren, there was a clinical trial proposed by a sponsor in order to expand the indication for a device. CDRH looked at the existing evidence in the registries about the device and found that it was sufficient for expanding the indication without further study. This reduced the amount of time necessary for the expanded indication, said Shuren, from 1 or 2 years to a couple of weeks.

Mark McClellan closed the discussion by asking each of the panelists where they would choose to direct resources to best accelerate effective use of RWE. Anderson replied that if additional resources were available, he would direct them to hiring new staff to help coordinate communications among different levels and organizations at the agency. Corrigan-Curay said she would invest in demonstration and validation projects so she could better understand the best methods and approaches for turning RWD into useful, fit-for-purpose evidence. She added that in particular, there is a need for further work to ensure the accuracy and relevancy of data from EHRs. Shuren said he would use additional resources to build out the active surveillance capabilities of CDRH as well as to adopt universal device identifiers.

FINAL THOUGHTS

The third workshop concluded with McClellan offering some overarching thoughts on the series of workshops and on the decision aids that had been discussed. He noted that the decision aids were "viewed as useful"

by workshop participants, and that they captured many key considerations that could be included in designing approaches for collecting, evaluating, and using RWD and RWE. He noted that while the decision aids were basic, they could be used as a framework on which additional details could be added for specific stakeholders or applications. Workshop participants repeatedly emphasized that how one uses RWD and RWE depends on the research question, the clinical context, and the decision to be made (see Box 10-1). For example, if it is a regulatory decision for a new product, one needs the highest quality evidence available. If it is a regulatory decision for a new indication or population, the evidence can be slightly less robust, and if it is a decision about policy or insurance coverage, the evidence could be even less robust. McClellan said the decision aids could be adapted for more specific uses by individual stakeholders who are making these different types of decisions, leading to the idea that the value of RWE depends on the stakeholder and the specific questions each hopes to answer.

McClellan noted that, broadly, the goal of decision aids is to build a systematic framework for thinking about a particular topic; he highlighted the generation of evidence that is fit for purpose as an important part of utilizing RWE. The current system of evidence generation, as discussed in the first workshop, is expensive, time consuming, and cannot answer all of the questions about a product or intervention. Important questions could potentially be answered "cheaper and faster" with a systematic, validated framework for generating RWE, he said. The key to creating this framework will be clarity and specificity about when using RWE is

BOX 10-1
Key Messages Identified by Individual Speakers

- Lists of questions, such as those laid out in the decision aids discussed at the third workshop, are useful, particularly as a framework to add details for specific stakeholders or applications. (McClellan)
- Clarity and specificity about when real-world evidence (RWE) is appropriate, and which data sources and methods are appropriate to address different types of questions, is the key to developing a framework for generating relevant evidence. (McClellan)
- There is growing infrastructure for RWE; that, along with tools such as the decision aids, can promote systematic and predictable opportunities for its use. (McClellan)
- Delivering better health care is the ultimate goal of using RWE. (Simon)
- Patients are valuable sources of information about the lived experience of a disease and can provide insights into relevant endpoints, outcomes, and preferences. (Nowell, Terry)

appropriate, and appropriate data sources and methods for different types of questions.

Improving the quality and reliability of data will also be essential for moving forward. Simon noted that improving the quality of data will be useful not just for research, but also for patient care. He said "the primary goal of improving the information that is generated in health care is to deliver better health care." Referring to the phrase "the tail is wagging the dog," Simon said that research is the tail, not the dog: "Actually improving people's health is the dog, and [research] is just the tail that rides along." Improving the data that are generated in the course of real-world care will benefit everyone, he said, and several other participants pointed to better data as a foundational principle to the increased use of RWD and RWE. He noted that many issues that were identified as making data collection more difficult are issues that affect our health care system as a whole. For example, discontinuities in the health care system and fragmentation of care affect both the collection of quality data as well as the care that people receive.

Relatedly, several workshop participants emphasized the importance of including patients and the patient perspective when thinking about the collection and use of RWE. W. Benjamin Nowell and Sharon Terry noted that patients are important sources of information for understanding the lived experience of a disease, and can provide insights about appropriate and relevant endpoints and outcomes for research. Jennifer Graff suggested that the word "patient" be incorporated into the decision aids, because while FDA and other stakeholders use evidence to make decisions, so do patients and caregivers.

McClellan closed by saying that tools such as the decision aids are "intended to help us move from case by case to more systematic and predictable opportunities." He noted that there is growing infrastructure for RWE, from Sentinel to registries to NEST, and that FDA and other regulators were open to engaging with RWE and building the systems necessary to use RWE for decision making. These efforts, along with workshops such as this one, will contribute to the creation of "well-understood, well-curated, fit-for-purpose sources of data" that can be collected and used in routine practice, as well as the development of analytic methods and tools for turning RWD into evidence that can be used by a variety of stakeholders for making different types of decisions.

References

Alphs, L. D., and C. A. Bossie. 2016. ASPECT-R—A Tool to Rate the Pragmatic and Explanatory Characteristics of a Clinical Trial Design. *Innovations in Clinical Neuroscience* 13(1-2):15.

Alphs, L., C. Benson, K. Cheshire-Kinney, J. P. Lindenmayer, L. Mao, S. C. Rodriguez, and H. L. Starr. 2015. Real-world outcomes of paliperidone palmitate compared to daily oral antipsychotic therapy in schizophrenia: A randomized, open-label, review board-blinded 15-month study. *Journal of Clinical Psychiatry* 76(5):554-561.

Avalos, L. A., H. Chen, and D. K. Li. 2015. Antidepressant medication use, depression, and the risk of preeclampsia. *CNS Spectrums* 20(1):39-47.

Bai, Y., H. Deng, A. Shantsila, and G. Y. Lip. 2017. Rivaroxaban versus Dabigatran or warfarin in real-world studies of stroke prevention in atrial fibrillation: Systematic review and meta-analysis. *Stroke* 48(4):970-976.

Berger, M. L., H. Sox, R. J. Willke, D. L. Brixner, H. G. Eichler, W. Goettsch, D. Madigan, A. Makady, S. Schneeweiss, R. Tarricone, S. V. Wang, J. Watkins, and C. D. Mullins. 2017. 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 in Health* 20(8):1003-1008.

Chrischilles, E. A., J. J. Gagne, B. Fireman, J. Nelson, S. Toh, A. Shoaibi, M. E. Reichman, S. Wang, M. Nguyen, R. Zhang, R. Izem, M. R. Goulding, M. R. Southworth, D. J. Graham, C. Fuller, H. Katcoff, T. Woodworth, C. Rogers, R. Saliga, N. D. Lin, C. N. McMahill-Walraven, V. P. Nair, K. Haynes, and R. M. Carnahan. 2018. Prospective surveillance pilot of rivaroxaban safety within the US Food and Drug Administration Sentinel System. *Pharmacoepidemiology and Drug Safety* 27(3):263-271.

Connolly, S. J., M. D. Ezekowitz, S. Yusuf, J. Eikelboom, J. Oldgren, A. Parekh, J. Pogue, P. A. Reilly, E. Themeles, J. Varrone, and S. Wang. 2009. Dabigatran versus warfarin in patients with atrial fibrillation. *New England Journal of Medicine* 361(12):1139-1151.

Daniel, G. W., M. Menis, G. Sridhar, D. Scott, A. E. Wallace, M. V. Ovanesov, B. Golding, S. A. Anderson, J. Epstein, D. Martin, and R. Ball. 2012. Immune globulins and thrombotic adverse events as recorded in a large administrative database in 2008 through 2010. *Transfusion* 52(10):2113-2121.

DiMasi, J. A., H. G. Grabowski, and R. W. Hansen. 2016. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics* 47:20-33.

Evidation and Novartis. 2018. *Real-world use of wearable devices in a large multiple sclerosis cohort*. Poster presentation at the American Academy of Neurology Annual Meeting, Los Angeles, CA.

Fergusson, D. A., P. C. Hébert, C. D. Mazer, S. Fremes, C. MacAdams, J. M. Murkin, K. Teoh, P. C. Duke, R. Arellano, M. A. Blajchman, and J. S. Bussières. 2008. A comparison of aprotinin and lysine analogues in high-risk cardiac surgery. *New England Journal of Medicine* 358(22):2319-2331.

Fralick, M., S. Schneeweiss, and E. Patorno. 2017. Risk of diabetic ketoacidosis after initiation of an SGLT2 inhibitor. *New England Journal of Medicine* 376(23):2300-2302.

Franklin, J. M., and S. Schneeweiss. 2017. When and how can real world data analyses substitute for randomized controlled trials? *Clinical Pharmacology & Therapeutics* 102(6):924-933.

Galson, S., and G. Simon. 2016. *Real-World Evidence to Guide the Approval and Use of New Treatments*. NAM Perspectives. Discussion Paper, National Academy of Medicine, Washington, DC. doi: 10.31478/201610b.

Gambhir, S. S., T. J. Ge, O. Vermesh, and R. Spittler. 2018. Toward achieving precision health. *Science Translational Medicine* 10(430):eaao3612.

Giles, J. T., N. Sattar, S. E. Gabriel, P. M. Ridker, S. Gay, C. Warne, D. Musselman, L. Brockwell, E. Shittu, M. Klearman, and T. Fleming. 2016. Comparative cardiovascular safety of tocilizumab vs etanercept in rheumatoid arthritis: Results of a randomized, parallel-group, multicenter, noninferiority, phase 4 clinical trial. *Arthritis & Rheumatology* 68(Suppl 10).

Gill, J., P. Kanavos, B. Avouac, R. Duncombe, J. Hutton, K. Jahnz-Różyk, W. Schramm, F. Spandonaro, and M. Thomas. 2016. *The use of real world evidence in the European context: An analysis of key expert opinion*. http://www.lse.ac.uk/business-and-consultancy/consulting/assets/documents/the-use-of-real-world-evidence-in-the-european-context.pdf (accessed December 21, 2018).

Giugliano, R. P., C. T. Ruff, E. Braunwald, S. A. Murphy, S. D. Wiviott, J. L. Halperin, A. L. Waldo, M. D. Ezekowitz, J. I. Weitz, J. Špinar, and W. Ruzyllo. 2013. Edoxaban versus warfarin in patients with atrial fibrillation. *New England Journal of Medicine* 369(22):2093-2104.

Go, A. S., E. M. Hylek, L. H. Borowsky, K. A. Phillips, J. V. Selby, and D. E. Singer. 1999. Warfarin use among ambulatory patients with nonvalvular atrial fibrillation: The An-Ticoagulation and Risk Factors in Atrial Fibrillation (ATRIA) study. *Annals of Internal Medicine* 131(12):927-934.

Goettsch, W., and A. Makady. 2015. *Glossary of definitions of common terms. IMI GetReal*. https://www.imi-getreal.eu/Portals/1/Documents/01%20deliverables/D1.3%20-%20Revised%20GetReal%20glossary%20-%20FINAL%20updated%20version_25Oct16_webversion.pdf (accessed March 27, 2019).

Gökbuget, N., M. Kelsh, V. Chia, A. Advani, R. Bassan, H. Dombret, M. Doubek, A. K. Fielding, S. Giebel, V. Haddad, D. Hoelzer, C. Holland, N. Ifrah, A. Katz, T. Maniar, G. Martinelli, M. Morgades, S. O'Brien, J. M. Ribera, J. M. Rowe, A. Stein, M. Topp, M. Wadleigh, and H. Kantarjian. 2016. Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. *Blood Cancer Journal* 6:e473.

Granger, C. B., J. H. Alexander, J. J. McMurray, R. D. Lopes, E. M. Hylek, M. Hanna, H. R. Al-Khalidi, J. Ansell, D. Atar, A. Avezum, and M.C. Bahit. 2011. Apixaban versus warfarin in patients with atrial fibrillation. *New England Journal of Medicine* 365(11):981-992.

Greene, S. M., R. J. Reid, and E. B. Larson. 2012. Implementing the learning health system: From concept to action. *Annals of Internal Medicine* 157(3):207-210.

Huang, G., X. Sun, D. Liu, Y. Zhang, B. Zhang, G. Xiao, X. Li, X. Gao, C. Hu, M. Wang, H. Ren, and S. Qin. 2018. The efficacy and safety of anti-PD-1/PD-L1 antibody therapy versus docetaxel for pretreated advanced NSCLC: A meta-analysis. *Oncotarget* 9(3):4239-4248.

Izurieta, H. S., M. Wernecke, J. Kelman, S. Wong, R. Forshee, D. Pratt, Y. Lu, Q. Sun, C. Jankosky, P. Krause, and C. Worrall. 2017. Effectiveness and duration of protection provided by the live-attenuated herpes zoster vaccine in the Medicare population ages 65 years and older. *Clinical Infectious Diseases* 64(6):785-793.

Kim, S. C., D. H. Solomon, J. R. Rogers, S. Gale, M. Klearman, K. Sarsour, and S. Schneeweiss. 2017. Cardiovascular safety of tocilizumab versus tumor necrosis factor inhibitors in patients with rheumatoid arthritis: A multi-database cohort study. *Arthritis & Rheumatology* 69(6):1154-1164.

Kimmelman, J., and A. J. London. 2015. The structure of clinical translation. *Hastings Center Report* 45(2):27-39.

London, A. J. 2018. Learning health systems, clinical equipoise and the ethics of response adaptive randomization. *Journal of Medical Ethics* 44:409-415.

Madjid, M., C. C. Miller, V. V. Zarubaev, I. G. Marinich, O. I. Kiselev, Y. V. Lobzin, A. E. Filippov, and S. W. Casscells III. 2007. Influenza epidemics and acute respiratory disease activity are associated with a surge in autopsy-confirmed coronary heart disease death: Results from 8 years of autopsies in 34,892 subjects. *European Heart Journal* 28(10):1205-1210.

Moseley, J. B., K. O'Malley, N. J. Petersen, T. J. Menke, B. A. Brody, D. H. Kuykendall, J. C. Hollingsworth, C. M. Ashton, and N. P. Wray. 2002. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *New England Journal of Medicine* 347(2):81-88.

Neal, B., V. Perkovic, K. W. Mahaffey, D. De Zeeuw, G. Fulcher, N. Erondu, W. Shaw, G. Law, M. Desai, and D. R. Matthews. 2017. Canagliflozin and cardiovascular and renal events in type 2 diabetes. *New England Journal of Medicine* 377(7):644-657.

Palmsten, K., K. F. Huybrechts, K. B. Michels, P. L. Williams, H. Mogun, S. Setoguchi, and S. Hernández-Díaz. 2013. Antidepressant use and risk for preeclampsia. *Epidemiology (Cambridge, Mass.)* 24(5):682.

Patel, M. R., K. W. Mahaffey, J. Garg, G. Pan, D. E. Singer, W. Hacke, B. Breithardt, J. L. Halperin, G. J. Hankey, J. P. Piccini, R. C. Becker, C. C. Nessel, J. F. Paolini, S. D. Berkowitz, K. A. Fox, and R. M. Califf, for the ROCKET AF Investigators. 2011. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *New England Journal of Medicine* 365(10):883-891.

Patorno, E., A. B. Goldfine, S. Schneeweiss, B. M. Everett, R. J. Glynn, J. Liu, and S. C. Kim. 2018. Cardiovascular outcomes associated with canagliflozin versus other non-gliflozin antidiabetic drugs: Population based cohort study. *BMJ* 360:k119.

Pew Charitable Trusts. 2018. *Mobile fact sheet*. http://www.pewinternet.org/fact-sheet/mobile (accessed December 21, 2018).

Physicians Foundation. 2018. *2018 Survey of America's Physicians Practice Patterns and Perspectives.* https://physiciansfoundation.org/wp-content/uploads/2018/09/physicians-survey-results-final-2018.pdf (accessed December 21, 2018).

Rubin, D. and N. Thomas. 2000. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* 95(450):573-585.

Ruff, C. T., R. P. Giugliano, E. Braunwald, E. B. Hoffman, N. Deenadayalu, M. D. Ezekowitz, A. J. Camm, J. I. Weitz, B. S. Lewis, A. Parkhomenko, and T. Yamashita. 2014. Comparison of the efficacy and safety of new oral anticoagulants with warfarin in patients with atrial fibrillation: A meta-analysis of randomised trials. *The Lancet* 383(9921):955-962.

Schneeweiss, S., J. D. Seeger, J. Landon, and A. M. Walker. 2008. Aprotinin during coronary-artery bypass grafting and risk of death. *New England Journal of Medicine* 358(8):771-783.

Seeger, J. D., K. Bykov, D. B. Bartels, K. Huybrechts, K. Zint, and S. Schneeweiss. 2015. Safety and effectiveness of dabigatran and warfarin in routine care of patients with atrial fibrillation. *Thrombosis and Haemostasis* 114(06):1277-1289.

Sentinel. 2018. *Sentinel common data model*. https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/sentinel-common-data-model (accessed December 17, 2018).

Sherman, R. E., S. A. Anderson, G. J. Dal Pan, G. W. Gray, T. Gross, N. L. Hunter, L. LaVange, D. Marinac-Dabic, P. W. Marks, M. A. Robb, and J. Shuren. 2016. Real-world evidence—what is it and what can it tell us. *New England Journal of Medicine* 375(23):2293-2297.

Thompson, W. W., D. K. Shay, E. Weintraub, L. Brammer, N. Cox, L. J. Anderson, and K. Fukuda. 2003. Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA* 289(2):179-186.

Thompson, W. W., D. K. Shay, E. Weintraub, L. Brammer, C. B. Bridges, N. J. Cox, and K. Fukuda. 2004. Influenza-associated hospitalizations in the United States. *JAMA* 292(11):1333-1340.

Tseng, H. F., N. Smith, R. Harpaz, S. R. Bialek, L. S. Sy, and S. J. Jacobsen. 2011. Herpes zoster vaccine in older adults and the risk of subsequent herpes zoster disease. *JAMA* 305(2):160-166.

Udell, J. A., R. Zawi, D. L. Bhatt, M. Keshtkar-Jahromi, F. Gaughran, A. Phrommintikul, A. Ciszewski, H. Vakili, E. B. Hoffman, M. E. Farkouh, and C. P. Cannon. 2013. Association between influenza vaccination and cardiovascular outcomes in high-risk patients: A meta-analysis. *JAMA* 310(16):1711-1720.

Van Ermen, A., M. P. Hermanson, J. M. Moran, N. K. Sweitzer, M. R. Johnson, and O. Vardeny. 2013. Double dose vs. standard dose influenza vaccination in patients with heart failure: A pilot study. *European Journal of Heart Failure* 15(5):560-564.

Vardeny, O., N. K. Sweitzer, M. A. Detry, J. M. Moran, M. R. Johnson, and M. S. Hayney. 2009. Decreased immune responses to influenza vaccination in patients with heart failure. *Journal of Cardiac Failure* 15(4):368-373.

Waldo, A. L., R. C. Becker, V. F. Tapson, K. J. Colgan, and NABOR Steering Committee. 2005. Hospitalized patients with atrial fibrillation and a high risk of stroke are not being provided with adequate anticoagulation. *Journal of the American College of Cardiology* 46(9):1729-1736.

Yih, W. K., T. A. Lieu, M. Kulldorff, D. Martin, C. N. McMahill-Walraven, R. Platt, N. Selvam, M. Selvan, G. M. Lee, and M. Nguyen. 2014. Intussusception risk after rotavirus vaccination in U.S. infants. *New England Journal of Medicine* 370:503-512.

Zinman, B., C. Wanner, J. M. Lachin, D. Fitchett, E. Bluhmki, S. Hantel, M. Mattheus, T. Devins, O. E. Johansen, H. J. Woerle, and U. C. Broedl. 2015. Empagliflozin, cardiovascular outcomes, and mortality in type 2 diabetes. *New England Journal of Medicine* 373(22):2117-2128.

Appendix A

Related Resources

The references listed below are papers, projects, and guidance documents that supported the planning of and discussions during each workshop in this series. References are separated by relevance to each of the three workshops, and the Proceedings—in Brief from workshops 1 and 2 produced by the National Academies of Sciences, Engineering, and Medicine as part of this activity are included as well. These references are intended to provide readers and real-world evidence stakeholders with additional context for the discussions that occurred during the workshops.

Workshop 1: Incentives

Berger, M., G. Daniel, K. Frank, A. Hernandez, M. McClellan, S. Okun, M. Overhage, R. Platt, M. Romine, S. Tunius, and M. Wilson. 2017. *A framework for regulatory use of real-world evidence*. Washington, DC: Duke-Margolis Center for Health Policy.

CDRH (Center for Devices and Radiological Health). 2017. *Use of real-world evidence to support regulatory decision-making for medical devices: Guidance for industry and Food and Drug Administration staff*. Rockville, MD: U.S. Food and Drug Administration.

CTTI (Clinical Trials Transformation Initiative). 2018. *Project: Quality by Design*. https://www.ctti-clinicaltrials.org/projects/quality-design (accessed November 6, 2018).

CTTI. 2018. *Project: Real World Evidence*. https://www.ctti-clinicaltrials.org/projects/real-world-evidence (accessed November 6, 2018).

NASEM (National Academies of Sciences, Engineering, and Medicine). 2017. *Real-world evidence generation and evaluation of therapeutics: Proceedings of a workshop*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/24685.

NASEM. 2018. *Examining the impact of real-world evidence on medical product development: I. Incentives: Proceedings of a workshop—in brief*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/25024.

National Health Council. 2017. *Patient perspectives on real-world evidence: A roundtable to gather views, needs, and recommendations*. Washington, DC: National Health Council.

Sherman, R. E., S. A. Anderson, G. J. Dal Pan, G. W. Gray, T. Gross, N. L. Hunter, L. LaVange, D. Marinac-Dabic, P. W. Marks, M. A. Robb, J. Shuren, R. Temple, J. Woodcock, L. Q. Yue, and R. M. Califf. 2016. Real-world evidence—what is it and what can it tell us? *New England Journal of Medicine* 375(23):2293-2297.

Workshop 2: Practical Approaches

NASEM (National Academies of Sciences, Engineering, and Medicine). 2018. *Examining the impact of real-world evidence on medical product development: II. Practical approaches: Proceedings of a workshop—in brief*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/25176.

Workshop 3: Application

CTTI (Clinical Trials Transformation Initiative). 2018. *Project: Mobile Technologies*. https://www.ctti-clinicaltrials.org/projects/mobile-technologies (accessed November 6, 2018).

Daniel, G., C. Silcox, J. Bryan, M. McClellan, M. Romine, and K. Frank. 2018. *Characterizing RWD quality and relevancy for regulatory purposes*. Washington, DC: Duke-Margolis Center for Health Policy.

FDA (U.S. Food and Drug Administration). 2013. *Guidance for industry and FDA staff: Best practices for conducting and reporting pharmacoepidemiologic safety studies using electronic healthcare data*. Rockville, MD: FDA.

FDA. 2013. *Guidance for industry: Electronic source data in clinical investigations*. Rockville, MD: FDA.

FDA. 2016. *Use of electronic informed consent—questions and answers: Guidance for institutional review boards, investigators, and sponsors*. Rockville, MD: FDA.

FDA. 2018. *Use of electronic health record data in clinical investigations: Guidance for industry*. Rockville, MD: FDA.

FOCR (Friends of Cancer Research). 2018. *Establishing a framework to evaluate real-world endpoints*. Washington, DC: Friends of Cancer Research.

Innovative Medicines Initiative GetReal. 2018. *Deliverables*. http://www.imi-getreal.eu/Publications/Deliverables-and-reports (accessed November 6, 2018).

Izurieta, H. S., M. Wernecke, J. Kelman, S. Wong, R. Forshee, D. Pratt, Y. Lu, Q. Sun, C. Jankosky, P. Krause, C. Worrall, T. MaCurdy, and R. Harpaz. 2017. Effectiveness and duration of protection provided by the live-attenuated herpes zoster vaccines in the Medicare population ages 65 years and older. *Clinical Infectious Diseases* 64(6):785-793.

Kimmelman, J., and A. J. London. 2015. The structure of clinical translation: Efficiency, information, and ethics. *Hastings Center Report* 45(2):27-39.

UK Academy of Medical Sciences and the Association of the British Pharmaceutical Industry. 2015. *Real world evidence*. London, England: UK Academy of Medical Sciences and the Association of the British Pharmaceutical Industry.

UK Academy of Medical Sciences FORUM. 2018. *Next steps for using real world evidence*. London, England: UK Academy of Medical Sciences.

Appendix B

Workshop One Agenda

Workshop One: Incentives

September 19–20, 2017

National Academy of Sciences Building, Lecture Room
2101 Constitution Avenue, NW, Washington, DC 20418

The National Academies of Sciences, Engineering, and Medicine is convening a three-part workshop series examining how real-world evidence development and uptake can enhance medical product development and evaluation. The workshops will advance discussions and common knowledge about complex issues relating to the generation and usage of real-world evidence, including fostering development and implementation of the science and technology of real-world evidence generation and usage.

This first workshop will include discussions and background materials that address:

- Aligning incentives and addressing barriers to support collection and use of real-world evidence in health product review, payment, and delivery.

Workshops two and three will foster discussions that will:

- Illuminate what types of data are appropriate for what specific purposes and suggest approaches for data collection that match the right data to the right questions. (Q1 2018)

- Examine and suggest approaches for operationalizing the collection and use of real-world evidence. (Q3 2018)

DAY 1: SEPTEMBER 19, 2017

8:00 a.m. Breakfast Available Outside the Lecture Room

8:20 a.m. **Welcome and Opening Remarks**

GREGORY SIMON, *Workshop Series Co-Chair*
Investigator
Kaiser Permanente Washington Health Research Institute

Keynote Address

8:30 a.m. **Vision and Goals of a Collaborative, Practical, and Sustainable Real-World Evidence Program**

SCOTT GOTTLIEB
Commissioner
U.S. Food and Drug Administration

8:50 a.m. **Discussion with Audience**

Moderator: Mark McClellan, Duke-Margolis Center for Health Policy

Session I: Seeing Our Destination

Session Objectives:

- Explore what relevant facts the ultimate end users of evidence need to know in order to make informed decisions about using medical products.
- Discuss possible approaches to generating such fit-for-purpose evidence.

Moderator: Andy Bindman, University of California, San Francisco

9:00 a.m. **A Payer Perspective**

MICHAEL SHERMAN
Senior Vice President and Chief Medical Officer
Harvard Pilgrim Health Care

9:20 a.m. **Delivery System Perspective: Integrated Care Model at Kaiser Permanente**

MICHAEL HORBERG
Executive Director, Research, Community Benefit, and Medicaid Strategy
Executive Director, Mid-Atlantic Permanente Research Institute
Kaiser Permanente Mid-Atlantic Permanente Medical Group

9:40 a.m. **Delivery System Perspective: Academic Health System**

DANIEL FORD
Director, Institute for Clinical and Translational Research
Johns Hopkins University School of Medicine

10:00 a.m. **A Patient-Focused Perspective**

SHARON TERRY
President and Chief Executive Officer
Genetic Alliance

10:20 a.m. **Discussion with Audience**

Additional Invited Discussants:
JOANNE WALDSTREICHER
Chief Medical Officer
Johnson & Johnson

ELEANOR PERFETTO
Senior Vice President, Strategic Initiatives
National Health Council

11:10 a.m. **BREAK**

11:30 a.m. **Key Messages and Themes from the September 13 FDA/ Duke-Margolis Workshop: Generating Fit-for-Purpose Evidence**

MARK MCCLELLAN, *Workshop Series Co-Chair*
Director
Duke-Margolis Center for Health Policy

11:50 a.m. **Discussion with Audience**

12:00 p.m. **BREAK** (Lunch Available Outside the Lecture Room)

Session II: Learning from Success

Session Objectives:

- Highlight successful completed and ongoing initiatives that could potentially be examined for real-world evidence collection and use.
- Explore the features that led to the success in the given examples and how they could apply to future applications:
  - Conditions likely to make innovation successful; and
  - Potential ways to recreate those conditions to make real-world evidence use more routine.

Moderator: Gregory Simon, Kaiser Permanente Washington Health Research Institute

1:00 p.m. **Generalizing and Scaling the Salford Lung Studies**

MARTIN GIBSON
Chief Executive Officer
Northwest EHealth

MARIE KANE
Chief Operating Officer
Northwest EHealth

1:30 p.m. **Using Sentinel to Evaluate Effectiveness or Efficacy**

RICHARD PLATT
Professor and Chair, Department of Population Medicine
Harvard Medical School

1:50 p.m. **Applying Lessons Learned from Device Registries to Other Treatment Types**

Rachael Fleurence
Executive Director
National Evaluation System for Health Technology (NEST) Coordinating Center

2:10 p.m. **Discussion with Audience**

Additional Invited Discussants:
John Graham
Head, Value Evidence and Outcomes
GlaxoSmithKline

Rachel Sherman
Principal Deputy Commissioner
U.S. Food and Drug Administration

3:00 p.m. **BREAK**

Session III: Getting Unstuck: Aligning Incentives

Session Objectives:

- In a series of presentations, discuss with treatment developers and evidence generators:
  - Incentives maintaining the current data generation process; and
  - Disincentives and potential barriers to incorporation of real-world evidence.

Moderator: Petra Kaufmann, National Center for Advancing Translational Sciences, National Institutes of Health

3:20 p.m. **Contract Research Organization Perspective**

John Doyle
Senior Vice President and Managing Director
QuintilesIMS

3:40 p.m. **A Product Developer Perspective**

ELLIOTT LEVY
Senior Vice President, Global Development
Amgen Inc.

BRIAN D. BRADBURY
Executive Director, Center for Observational Research
Amgen Inc.

4:00 p.m. **An Academic Researcher Perspective**

DANIEL FORD
Director, Institute for Clinical and Translational Research
Johns Hopkins University School of Medicine

4:20 p.m. **Data Stewards: Organizations with Large Data Sources**

MARCUS WILSON
President
HealthCore, Inc.

4:40 p.m. **Discussion with Audience**

Additional Invited Discussants:
MICHAEL HORBERG
Executive Director, Research, Community Benefit, and Medicaid Strategy
Executive Director, Mid-Atlantic Permanente Research Institute
Kaiser Permanente Mid-Atlantic Permanente Medical Group

ANNA MCCOLLISTER-SLIPP
Chief Advocate for Participatory Research, Scripps Translational Science Institute
Founder, VitalCrowd
Co-Founder, Galileo Analytics

5:30 p.m. **ADJOURN WORKSHOP DAY 1**

DAY 2: SEPTEMBER 20, 2017

8:00 a.m. Breakfast Available Outside the Lecture Room

8:30 a.m. **Recap Day 1 and Discussion with Workshop Participants**

GREGORY SIMON, *Workshop Series Co-Chair*
Investigator
Kaiser Permanente Washington Health Research Institute

Keynote Address

9:00 a.m. **False Precision and Estimating the Reliability of Effects with the Traditional Evidence-Generating Process**

ROBERT CALIFF
Vice Chancellor, Health Data Science, Duke University
Scientific Advisor, Verily Life Sciences

Session IV: Getting Unstuck: Myth-Busting

Session Objective:

- Examine ideas—and misconceptions—about the necessity and acceptability of established evidence-generation practices.

Moderator: Robert Califf, Duke University and Verily Life Sciences

9:30 a.m. **Moving from "One Study at a Time" to "All by All" Analyses**

PATRICK RYAN
Senior Director and Head, Epidemiology Analytics
Janssen Research & Development

9:50 a.m. **A Medical Product Developer Perspective**

JOHN GRAHAM
Head, Value Evidence and Outcomes
GlaxoSmithKline

10:10 a.m. **BREAK**

10:30 a.m. **Evolve or Die: The Urgent Need to Streamline Randomized Trials**

Rory Collins
Head of Nuffield Department of Population Health
University of Oxford

10:50 a.m. **A Regulatory Perspective**

Janet Woodcock
Director
Center for Drug Evaluation and Research
U.S. Food and Drug Administration

11:10 a.m. **Discussion with Audience**

Additional Invited Discussant:
Deven McGraw
Deputy Director, Health Information Privacy
Office for Civil Rights
U.S. Department of Health and Human Services

12:30 p.m. **ADJOURN WORKSHOP DAY 2**

Appendix C

Workshop Two Agenda

Workshop Two: Practical Approaches

March 6–7, 2018

National Academy of Sciences Building, Room 120
2101 Constitution Avenue, NW, Washington, DC 20418

The National Academies of Sciences, Engineering, and Medicine is convening a three-part workshop series examining how real-world evidence development and uptake can enhance medical product development and evaluation. The workshops will advance discussions and common knowledge about complex issues relating to the generation and usage of real-world evidence, including fostering development and implementation of the science and technology of real-world evidence generation and usage.

- Workshop One (*September 19–20, 2017*) focused on how to align incentives to support collection and use of real-world evidence in health product review, payment, and delivery. Incentives need to address barriers impeding the uptake of real-world evidence, including barriers to transparency.
- **Workshop Two (*March 6–7, 2018*) will illuminate what types of data are appropriate for what specific purposes and suggest practical approaches for data collection and evidence use by developing and working through example use cases.**

- Workshop Three (*July 17–18, 2018*) will examine and suggest approaches for operationalizing the collection and use of real-world evidence.

DAY 1: MARCH 6, 2018

8:30 a.m. Breakfast Available Outside Room 120

8:40 a.m. **Welcome and Opening Remarks**

Gregory Simon, *Workshop Series Co-Chair*
Investigator
Kaiser Permanente Washington Health Research Institute

Session I: When Can We Rely on Real-World Data?

Session Discussion Questions:

- When can we have confidence in EHR data from real-world practice to accurately assess study eligibility, key prognostic factors, and study outcomes?
- When can we have confidence in data generated outside of clinical settings (e.g., mobile phones, connected glucometers, connected blood pressure monitors)?
- When does adjudication or other postprocessing of real-world data add value?

Moderator: Gregory Daniel, Duke-Margolis Center for Health Policy

Session Discussants
Jesse Berlin
Vice President and Global Head, Epidemiology
Johnson & Johnson

Andy Bindman
Professor of Medicine
University of California, San Francisco

9:00 a.m. **Introduction and Background to Inform the Discussion: Novel Oral Anticoagulants in Comparison with Warfarin**

Adrian Hernandez
Vice Dean for Clinical Research
Duke University School of Medicine

9:20 a.m. **Open Discussion with Audience**

- What questions can characterize the utility of any real-world data source and signal reliability before a study is performed (examples below)?
  - When is accuracy good enough to reasonably and consistently identify the right population?
  - When is accuracy good enough to reasonably and consistently assess the exposure or intervention?
  - When is accuracy good enough to reasonably and consistently assess the right outcome?
  - Are there any big safety issues that would be missed?
  - Are there concerns about data collection or entry, particularly in relation to creating systemic bias?
  - When is expert adjudication necessary to confirm that the recorded data are reliable and/or reasonably complete?
- What information is needed to answer such questions?

10:40 a.m. **BREAK** (*Workgroup Participants Gather to Synthesize Audience Feedback*)

11:00 a.m. **Workgroup Presents Synthesis of Audience Feedback**

Session II: When Can We Rely on Real-World Treatment?

Session Discussion Questions:

- When conducting research in a real-world setting, are there situations that would require special guidance, knowledge, or experience in order for clinicians to adequately monitor participant safety and respond appropriately to adverse events?
- When does variation between comparison groups (socioeconomic, demographic, etc.); in treatment fidelity; in provider behavior and preferences; or in adherence yield a valid signal about real-world effectiveness, and when is it just noise?

Moderator: Khaled Sarsour, Genentech

Session Discussants

Michael Horberg
Executive Director, Research, Community Benefit, and Medicaid Strategy
Executive Director, Mid-Atlantic Permanente Research Institute
Kaiser Permanente Mid-Atlantic Permanente Medical Group

Gregory Simon
Investigator
Kaiser Permanente Washington Health Research Institute

Robert Temple
Deputy Director for Clinical Science
Center for Drug Evaluation and Research
U.S. Food and Drug Administration

11:15 a.m. **Introduction and Presentation to Inform Discussion on Participant Monitoring: Study on Lithium for Suicidal Behavior in Mood Disorders**

Ira Katz
Senior Consultant for Program Evaluation
U.S. Department of Veterans Affairs Office of Mental Health and Suicide Prevention

11:35 a.m. **Open Discussion with Audience**

- What conditions make self-monitoring and reporting acceptable?
- Does this vary for treatments at different stages of product development or with different baseline knowledge about use in varied patient types and treatment conditions?
- Can we draw any generalizable lessons about cases in which self-monitoring is acceptable and safe?

12:15 p.m. **Introduction and Presentation to Inform Discussion on Signal Detection: Novel Oral Anticoagulants in Comparison with Warfarin**

12:30 p.m. **Open Discussion with Audience**

- What conditions and training prepare clinical care providers to monitor patient safety outside a tightly controlled environment?
- How does this vary for treatments at different stages of product development or with different baseline knowledge about use in varied patient types and treatment conditions?
- How do you decide which variables require strict adherence to "protocol" and which can be allowed to vary?

1:00 p.m. **BREAK** (Lunch Available Outside Room 120)
(*Workgroup participants gather to synthesize audience feedback*)

2:00 p.m. **Workgroup Presents Synthesis of Audience Feedback**

Session III: When Can We Learn from Real-World Treatment Assignment

Session Discussion Questions:

- When can we have confidence in inference from cluster-randomized or stepped-wedge study designs?
- Under what conditions can we trust inference from observational or naturalistic comparisons?
- How could we judge the validity of observational comparisons in advance, rather than waiting until we have observed the result?

Moderator: Richard Platt, Harvard Medical School

Session Discussants

Robert Califf
Vice Chancellor, Health Data Science, Duke University
Scientific Advisor, Verily Life Sciences

David Madigan
Professor of Statistics
Dean, Faculty of Arts and Sciences
Columbia University

David Martin
Associate Director for Real-World Evidence Analytics
U.S. Food and Drug Administration

2:20 p.m. **Introduction and Presentation to Inform the Discussion: Health Care Database Analyses of Medical Products for Regulatory Decision Making**

Sebastian Schneeweiss
Professor of Medicine and Epidemiology
Harvard Medical School
Brigham & Women's Hospital

2:50 p.m. **Open Discussion with Audience**

- Random assignment is always preferable, but when is the cost (in time, money, infrastructure, patient exposure) truly necessary?
- How can we know that the effects from unmeasured confounders are not so large that they would change a decision based on information from an observational study?
- What are some of the conditions under which there is more confidence in inference from non-randomized comparisons (*examples of some conditions below*)?
  - Expectation of large effects
  - Outcome proximal to treatment
  - High degree of similarity between comparison groups
  - Pathway from treatment to outcome is relatively clear, and without lots of complexity or reciprocal effects
  - Treatment allocation method is relatively transparent

3:40 p.m. **BREAK**

4:00 p.m. **Open Discussion with Audience and Reflections from Panelists**

5:00 p.m. **ADJOURN WORKSHOP DAY 1**

DAY 2: MARCH 7, 2018

8:30 a.m. Breakfast Available Outside Room 120

Session IV: Synthesizing the Use Cases

Session Objectives:

- Discuss key considerations presented in each session on Day 1
- Consider components of a potential "checklist" for using real-world evidence

9:00 a.m. **Welcome and Recap of Day 1**

Mark McClellan, *Workshop Series Co-Chair*
Director
Duke-Margolis Center for Health Policy

Gregory Simon, *Workshop Series Co-Chair*
Investigator
Kaiser Permanente Washington Health Research Institute

9:20 a.m. **Open Discussion with Audience of Outputs from Day 1 and Potential Components to a "Checklist" for Using Real-World Evidence**

10:40 a.m. **BREAK**

11:00 a.m. **Open Discussion with Audience of Outputs from Day 1 and Potential Components to a "Checklist" for Using Real-World Evidence**

12:30 p.m. **ADJOURN WORKSHOP DAY 2**

Future Workshop Objectives

WORKSHOP THREE: Examine and suggest approaches for operationalizing the collection and use of real-world evidence (July 17–18, 2018, Washington, DC)

- Applications for using real-world evidence to supplement traditional clinical trials, pragmatic/effectiveness trials, or routine clinical application.
- Mechanisms for determining which discrete types of real-world evidence could support regulatory decisions.
- Operational challenges and barriers for generating and incorporating real-world evidence in the context of a learning health system and how clinicians can best be involved in the collection and usage/utilization of real-world evidence.

Appendix D

Workshop Three Agenda

Workshop Three: Application

July 17–18, 2018

National Academy of Sciences Building, Lecture Room
2101 Constitution Avenue, NW, Washington, DC 20418

The National Academies of Sciences, Engineering, and Medicine is convening a three-part workshop series, sponsored by the U.S. Food and Drug Administration, examining how real-world evidence development and uptake can enhance medical product development and evaluation. The workshops will advance discussions and common knowledge about complex issues relating to the generation and usage of real-world evidence, including fostering development and implementation of the science and technology of real-world evidence generation and usage.

- Workshop One (*September 19–20, 2017*) focused on how to align incentives to support collection and use of real-world evidence in health product review, payment, and delivery. Incentives need to address barriers impeding the uptake of real-world evidence, including barriers to transparency.
- Workshop Two (*March 6–7, 2018*) illuminated what types of data are appropriate for what specific purposes and suggested practical approaches for data collection and evidence use by developing and working through example use cases.

- **Workshop Three (*July 17–18, 2018*) examined and suggested approaches for operationalizing the collection and use of real-world evidence through discussing "decision aids" about specific topics in study design. The decision aids are question lists developed to inform discussion around specific topics addressed at the third workshop. These discussions may help inform workshop attendees and other stakeholders about study design choices, including potential risks, costs, and reporting/transparency expectations.**

DAY 1: JULY 17, 2018

8:00 a.m. Breakfast Available Outside the Lecture Room

8:15 a.m. **Welcome and Opening Remarks**

Mark McClellan, *Workshop Series Co-Chair*
Director
Duke-Margolis Center for Health Policy

Gregory Simon, *Workshop Series Co-Chair*
Investigator
Kaiser Permanente Washington Health Research Institute

Session I: Key Considerations for Real-World Evidence Application

Session Objectives:

- Examine how some organizations are currently considering traditional and real-world evidence.
- Discuss factors that may be influencing overall cost and time investment required by traditional evidence generation.
- Consider when non-traditional data sources may be beneficial to assess outcomes.

8:45 a.m. **Update on the Innovative Medicines Initiative's GetReal and View from the National Institute for Health and Care Excellence, United Kingdom**

Pall Jonsson
Associate Director, Research and Development
National Institute for Health and Care Excellence

9:05 a.m. **Drivers of Expense and Delay**

ELLIOTT LEVY
Senior Vice President, Global Development
Amgen Inc.

9:25 a.m. **Patient-Collected and -Owned Data**

KOMATHI STEM
Chief Executive Officer and Founder
monARC Bionetworks

9:45 a.m. **BREAK**

Session II: When Is a Real-World Data Element Fit for Assessment of Eligibility, Treatment Exposure, or Outcomes?

Session Objectives:

- Discuss potential bias-introducing steps in evidence generation from real-world data.
- Suggest key considerations in the data collection and evidence-generation processes that influence reliability of real-world data.
- Discuss how a decision aid laying out key questions and considerations might help inform current and future studies.

10:05 a.m. **Introduction: A Proposed Framework for a Decision Aid**

PALL JONSSON, *Session Moderator*
Associate Director, Research and Development
National Institute for Health and Care Excellence

10:15 a.m. **Looking Back: How Might a Decision Aid Inform a Real-World Example?**

JEFF ALLEN
President and Chief Executive Officer
Friends of Cancer Research

10:35 a.m. **Looking Forward: How Decision Aid Might Apply to Future Studies**
Panel Discussion and Audience Q&A

Aylin Altan
Senior Vice President of Research
OptumLabs

Robert Ball
Deputy Director, Office of Surveillance and Epidemiology
Center for Drug Evaluation and Research
U.S. Food and Drug Administration

Luca Foschini
Co-Founder and Chief Data Scientist
Evidation Health

Brande Yaist
Senior Director, Global Patient Outcomes and Real-World Evidence
Eli Lilly and Company

12:00 p.m. **BREAK** (Lunch Available Outside the Lecture Room)

Session III: Obscuring Intervention Allocation in Trials to Generate Real-World Evidence: Why, Who, and When?

Session Objectives:

- Discuss how variability in knowledge of treatment assignment group affects:
  - Provider and patient adherence and outcomes.
  - Study cost and reliability.
- Suggest key factors that could affect decisions to obscure intervention allocation.
- Discuss how a decision aid laying out key questions and considerations might help inform current and future studies.

1:00 p.m. **Introduction: A Proposed Framework for a Decision Aid**

JONATHAN WATANABE, *Session Moderator*
Associate Professor of Clinical Pharmacy
National Academy of Medicine Anniversary Fellow in Pharmacy
University of California, San Diego

1:10 p.m. **Looking Back: How Might a Decision Aid Inform a Real-World Example?**

JOHN GRAHAM
Head, Value Evidence and Outcomes
GlaxoSmithKline

ORLY VARDENY
Minneapolis Department of Veterans Affairs Center for Chronic Disease Outcomes Research
Associate Professor of Medicine
University of Minnesota

1:30 p.m. **Looking Forward: How a Decision Aid Might Apply to Future Studies**
Panel Discussion and Audience Q&A

CATHY CRITCHLOW
Vice President, Center for Observational Research
Amgen Inc.

NANCY DREYER
Chief Scientific Officer
IQVIA

ALEX JOHN LONDON
Clara L. West Professor of Ethics and Philosophy
Carnegie Mellon University

JAMES P. SMITH
Deputy Director, Division of Metabolism and Endocrinology Products
Center for Drug Evaluation and Research
U.S. Food and Drug Administration

2:50 p.m. **BREAK**

Session IV: How Tightly Should Investigators Attempt to Control or Restrict Treatment Quality in a Pragmatic or Real-World Trial?

Session Objectives:

- Discuss how variability in treatment delivery and adherence can affect results, including
  - Potential influence of variation in standard treatment practice, and
  - Considerations for balancing participant autonomy and safety.
- Suggest key factors that could help determine the base comparison and level of control suited to a particular trial.
- Discuss how a decision aid laying out key questions and considerations might help inform current and future studies.

3:10 p.m. **Introduction: A Proposed Framework for a Decision Aid**

Jennifer Graff, *Session Moderator*
Vice President of Comparative Effectiveness Research
National Pharmaceutical Council

3:20 p.m. **Looking Back: How Might a Decision Aid Inform a Real-World Example?**

Larry Alphs
Deputy Chief Medical Officer
Newron Pharmaceuticals

3:40 p.m. **Looking Forward: How a Decision Aid Might Apply to Future Studies**
Panel Discussion and Audience Q&A

Judith Carrithers
Director of Regulatory Services
Advarra

W. Benjamin Nowell
Director, Patient-Centered Research
Global Healthy Living Foundation
Co–Principal Investigator, ArthritisPower Patient Powered Research Network

PETER STEIN
Deputy Director, Office of New Drugs
Center for Drug Evaluation and Research
U.S. Food and Drug Administration

4:50 p.m. **Day 1 Wrap-Up and Concluding Thoughts/Discussion with Audience**

5:00 p.m. **ADJOURN WORKSHOP DAY 1**

DAY 2: JULY 18, 2018

8:00 a.m. **Welcome**

MARK MCCLELLAN, *Workshop Series Co-Chair*
Director
Duke-Margolis Center for Health Policy

GREGORY SIMON, *Workshop Series Co-Chair*
Investigator
Kaiser Permanente Washington Health Research Institute

Session V: How Can Bias in Observational Comparisons Be Assessed and Minimized?

Session Objectives:

- Discuss methods to assess presence of and optimally reduce bias from unmeasured confounding.
- Suggest key considerations for assessing—and communicating—uncertainty in observational studies.
- Discuss how a decision aid laying out key questions and considerations might help inform current and future studies.

8:10 a.m. **Introduction: A Proposed Framework for a Decision Aid**

DAVID MARTIN
Associate Director for Real-World Evidence Analytics
U.S. Food and Drug Administration

8:20 a.m. **Looking Back: How Might a Decision Aid Inform a Real-World Example?**

HECTOR IZURIETA
Epidemiologist, Office of Biostatistics and Epidemiology
Center for Biologics Evaluation and Research
U.S. Food and Drug Administration

8:40 a.m. **Looking Forward: How a Decision Aid Might Apply to Future Studies**
Panel Discussion and Audience Q&A

GREGORY DANIEL, *Session Moderator*
Deputy Director
Duke-Margolis Center for Health Policy

JESSICA FRANKLIN
Assistant Professor of Medicine
Harvard Medical School

NICOLE GORMLEY
Team Lead, Division of Hematologic Products
Office of Hematology and Oncology Products
Center for Drug Evaluation and Research
U.S. Food and Drug Administration

JAVIER JIMENEZ
Vice President and Global Head for Real-World Evidence and Clinical Outcomes
Sanofi

HENG LI
Mathematical Statistician
Center for Devices and Radiological Health
U.S. Food and Drug Administration

MARK VAN DER LAAN
Professor, Biostatistics and Statistics
University of California, Berkeley

10:00 a.m. **BREAK**

Session VI: FDA Panel

Session Objectives:

- Hear updates and perspective of current thinking about real-world evidence in Europe.
- Discuss challenges, opportunities, and remaining gaps for moving forward with real-world evidence application.

10:15 a.m. **A European Perspective**

ALASDAIR BRECKENRIDGE, *Session Moderator*
Emeritus Professor of Clinical Pharmacology
University of Liverpool

10:30 a.m. **Reflections from FDA**

STEVEN ANDERSON
Director, Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research
U.S. Food and Drug Administration

JACQUELINE CORRIGAN-CURAY
Director, Office of Medical Policy, Center for Drug Evaluation and Research
U.S. Food and Drug Administration

JEFF SHUREN
Director, Center for Devices and Radiological Health
U.S. Food and Drug Administration

11:15 a.m. **Panel Discussion with Audience**

11:50 a.m. **Synthesis of Workshop Discussions**

MARK MCCLELLAN, *Workshop Series Co-Chair*
Director
Duke-Margolis Center for Health Policy

GREGORY SIMON, *Workshop Series Co-Chair*
Investigator
Kaiser Permanente Washington Health Research Institute

12:00 p.m. **ADJOURN WORKSHOP DAY 2**