Green Park Collaborative

A partnership for innovation and effectiveness

RWE Decoder Framework

A Practical Tool for Assessing Relevance and Rigor of Real World Evidence

A White Paper from the Green Park Collaborative

The Green Park Collaborative is a major initiative of the Center for Medical Technology Policy

February 7, 2017 | Version 1.0

BETTER EVIDENCE.
BETTER DECISIONS.
BETTER HEALTH.

Center for Medical Technology Policy
World Trade Center Baltimore
401 E Pratt Street, Suite 631
Baltimore, Maryland 21202

410.547.CMTP (P)
410.547.5088 (F)
info@cmtpnet.org (E)
@cmtp_baltimore (Tw)

www.cmtpnet.org

TABLE OF CONTENTS

EXECUTIVE SUMMARY

With the growth of electronic medical records and large data repositories, and increasing efforts to systematically collect information from routine clinical encounters, the potential for development and use of Real World Evidence (RWE) has been exploding. RWE is typically derived from data that are generated from clinical care or patient experience, as opposed to formal research settings. RWE studies can answer, for example, questions of comparative effectiveness, total costs of care, or patient-centered outcomes research. In some contexts, RWE may be more directly relevant and in that respect, more useful than a well-designed randomized clinical trial (RCT). It can also provide useful information to supplement or enrich evidence from RCTs or existing research with data collected during routine patient care.

However, not all RWE is created equal. Concerns prevail regarding the appropriateness of applying RWE of unknown rigor to decision-making for patients and populations. Despite the rapid growth and availability of real world data (RWD) and RWE generated from multiple data sources, there are still very common and in some cases, well-founded concerns about the trustworthiness of data and the quality and credibility of conclusions from RWE using these data, particularly among decision makers who must act with imperfect evidence and incomplete access to the details of real world studies. On the other hand, many decision makers and researchers alike recognize that good quality RWE exists and can be highly relevant for many health care decision contexts.

In recognition of this challenge, some groups have developed checklists, hierarchies, and other frameworks to assess the quality of studies. While expertly developed and reported, many of these tools are difficult for some of the real world decision makers[1] who would use them. In addition, they typically do not devote significant consideration to the individual user's context of decision-making. No single framework has to date seen broad adoption among the variety of stakeholder groups representing users of RWE for decision-making. There is not yet any degree of consensus among decision makers about a framework for assessing RWE that accurately discriminates between studies that deserve greater consideration in decision-making and those with significant flaws rendering them inappropriate or irrelevant for use. As a result, the use of RWE for many decision makers is still timid and inconsistent, thus reducing the enthusiasm of RWE creators who would generate this type of evidence. Hence, in striking contrast to the rapidly growing availability of RWE, its application in many areas of health related decision-making has evolved at a much slower pace.

In recognition of this need, the Center for Medical Technology Policy (CMTP) and the Green Park Collaborative (GPC) launched an initiative to develop a framework to help decision makers more confidently and consistently assess RWE for their unique decision-making needs. Our approach for this project, which was launched in January 2016, was to gather and incorporate decision maker perspectives representing the broader universe of potential RWE users, together with those who generate RWE, at multiple points of the framework development.

As described in this white paper, our background and engagement work underscored that existing evidence frameworks require a high level of sophistication and a great deal of time to use, meanwhile actual users have varying levels of methodological sophistication and often limited time. With our workgroups, we sought to leverage the strengths of these existing efforts while engaging the broader perspectives of RWE users to develop a framework (and tools) that can help decision makers to more

[1] Post-regulatory decision makers, such as payers, clinical guideline developers, health systems, health technology assessment groups, accountable care organizations, and etc.

confidently and consistently use RWE appropriate for their decision-making purposes. The resulting framework is a spreadsheet-based tool to review existing published studies by independently assessing both the rigor and relevance of available studies for a given decision. The framework guides users towards an "informed judgment" of study quality, informed by stepwise considerations of domains within both relevance and rigor, and provides a visual summary by which a user can gauge as a whole, an applicable body of evidence.

As most existing tools and best practices require a significant level of sophistication regarding methods for assessing published clinical studies we aimed to develop a framework and approach that could be applied with limited training, and in a reasonable amount of time. Our intention was to create something that included sufficient technical detail to accurately differentiate those RWE studies that are high quality form those that are low quality. Thus, the GPC RWE Framework (RWE Decoder) presents a formalized but easy to follow approach for the broader universe of U.S. health care decision makers who already or are likely to use RWE to help guide their decision-making. Our working group deliberated extensively over the right balance between usability and accuracy in the tool. This balance is ultimately a matter of judgement; we will continue to improve the tool with additional pilot testing and feedback.

One important value of such a tool is to overcome the temptation of some decision makers to discard all RWE studies, perhaps in part reflecting the concern that they have no reliable way to know which ones should at least be considered. However, like any cultural shift in research, we look to achieve stepwise progress instead of overnight change. We are confident that these tools are a step in the right direction.

CMTP WRITING AND EDITORIAL TEAM

Rachael Moloney, MHS
Research Manager, Center for Medical Technology Policy

Jennifer Al Naber, MS, MSPH
Program Manager, Green Park Collaborative

Donna Messner, PhD
Senior Vice President, Center for Medical Technology Policy

Michael Stoto, PhD
Professor of Health Systems Administration and Population Healthy, Georgetown University

FUNDING

Support for this research initiative was provided by Astellas Pharma US, Eli Lilly and Company, EMD Serono, Genentech, Inc., Janssen Pharmaceuticals, Inc., Merck & Co., Inc., National Pharmaceutical Council, Pfizer Inc., Sanofi US, and UCB.

ACKNOWLEDGEMENTS

CMTP is grateful to all of the individuals who served on the Advisory Committee, Methods and Dissemination Workgroups for this project.[2] Their generous donation of time and expertise contributed greatly to the overall success of this initiative.

[2] See Appendix 1 for a full listing of participants and working group members.

CMTP would also like to acknowledge the following staff members for their dedication and hard work over the course of the past year: Julie Simmons, CMP, Manager, Marketing and Communications; and Janelle King, Executive Assistant.

Finally, the writing and editorial team would like to especially acknowledge Sean Tunis, MD, MSc, President and Chief Executive Officer of CMTP, for his leadership in the conceptualization of the GPC RWE Initiative and his ongoing guidance over the course of the project.

BACKGROUND: INITIATIVE AIMS AND APPROACH

The Center for Medical Technology Policy (CMTP), through its Green Park Collaborative (GPC), undertook a multifaceted multi-stakeholder initiative with the aim of facilitating the use of high-quality real world evidence (RWE) – generally defined as health care related information derived from a variety of sources outside typical clinical research settings (e.g. administrative claims, electronic health records, etc.) (Sherman et al) – for post-regulatory decision-making by users in health systems, guideline developers, payers and others, and of helping to create better consistency and transparency in the way such evidence is used. A primary motivation for the GPC RWE Initiative is to reduce the inclination of taking a "randomized or bust" approach when ascertaining the quality of a study[3] that has generated RWE by leveraging data collected as a by-product of clinical care, or for a purpose other than research. This dichotomous lens can lead to missed opportunities to inform decision-making with high quality RWE – which does exist but which can be difficult for the average decision maker to identify and assess. While a number of high-quality evidence evaluation frameworks exist, there is no broad consensus over the standards that should go into these frameworks, many of which tend to prioritize randomized evidence over all other types. Yet for some decisions, high-quality non-randomized RWE, or evidence from pragmatic trials utilizing real world data to assess primary outcomes may be more relevant and useful than available randomized studies or epidemiologic cohort studies.

For these reasons, in January 2016 GPC launched an initiative comprised of four major parts (illustrated in Figure 1).

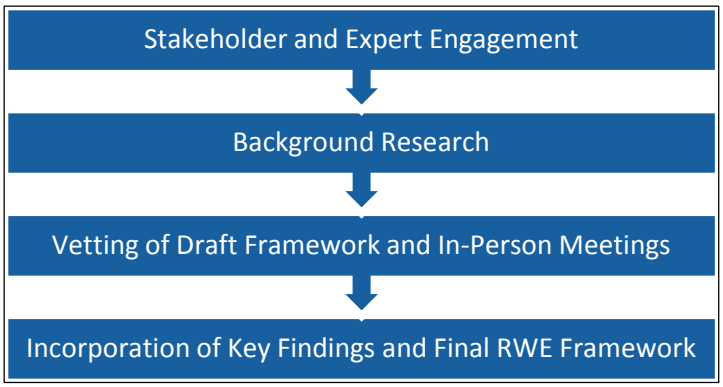

Figure 1. Approach to Developing RWE Decoder Framework

STAKEHOLDER AND EXPERT ENGAGEMENT

Three expert multi-stakeholder workgroups were convened as follows:

- Advisory Committee: senior representatives from across a range of different stakeholder organizations to guide the overall initiative approach, provide nominations for the initiative workgroups (dissemination and methods), and input on all other workgroup activities and products.
- Methods Workgroup: a group of academic and other methodologists tasked with taking the theoretical objective of the GPC RWE Initiative and conceptualizing an implementable framework, with tools that are feasible and helpful for decision makers to use.

[3] Here, and for the purposes of the GPC RWE Initiative we define "study" loosely to be any systematic scientific or statistical approach undertaken to answer a health related query.

- Dissemination Workgroup: tasked with developing and implementing a communication plan for the short- and long-term dissemination and uptake of the RWE Framework. This group worked closely with the CMTP Marketing and Communications Committee.

Project staff provided channels for communication between these groups, as needed, over the course of the framework development. Lists of all workgroup participants are provided in Appendix 1.

BACKGROUND RESEARCH

GPC staff conducted background research to help refine the overall initiative approach and inform the deliberations of the workgroups. This included an environmental scan of existing methods guidance, best practices, reporting guidelines or other evidence assessment frameworks directly or indirectly related to real world evidence. The purpose of the scan was to describe existing work and help inform the early conceptualization and development of the GPC RWE Decoder Framework. Related methods guidelines were characterized and compared, including how they were developed and by whom (e.g. expert consensus, technical working groups, etc.), intended users, and published tools. Initiatives were identified through directed online searching as well as expert input from members of the Advisory Committee and the Methods Workgroup.

The environmental scan (Appendix 2) reviewed the following initiatives: GRADE, STROBE Statement, RECORD Statement, PCORI Methodology Report, AHRQ Effective Health Care Program, GRACE Principles and Checklist, Cochrane Risk of Bias, FDA CDER Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment, ISPE Guidelines (Good Pharmacoepidemiology Practices and Good Database Selection and Use in Pharmacoepidemiologic Research), CER Collaborative, IMI GetReal, ISPOR Best Practices Task Force for Comparative Effectiveness Research (Prospective Observational Studies, Retrospective Database Analysis Parts I-III, and Real World Data), and European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). (FDA Draft Guidance and PCOR Data Quality and Transparency Standards project (Kahn et al, 2015) were later incorporated). Appendix 2 compares these initiatives with the GPC RWE Decoder Framework.

These initiatives did not all produce guidelines or best practices explicitly for RWE, however they covered different study designs and analyses that might be utilized to generate RWE. For example, ISPE guidelines are specific to pharmacoepidemiology research and focus on the appropriate use of healthcare databases (Public Policy Committee, ISPE, 2016). The collection of standards, guidelines and frameworks disseminated by these related initiatives were also written for a variety of audiences with different areas of expertise (ranging from researchers in academia to pharmacovigilance specialists in the industry), and thus did not entirely overlap on topics and criteria.

In addition, telephone interviews were conducted with 12 key informants and members of the Advisory Committee, to gather a range of stakeholder perspectives (representing RWE users), regarding: their evidence needs; policy challenges unique to their organizations; ways in which RWE is currently incorporated into decision-making; and contexts in which a higher threshold of uncertainty might be more or less acceptable. Additional key informants from outside the Advisory Committee were also recruited to speak to the perspectives of accountable care organizations and integrated health delivery systems. Interviews were recorded and transcribed. Themes were identified, discussed among staff, and reported to GPC RWE Initiative stakeholders in a Stakeholder Briefing Document, circulated as pre-read material prior to the in-person stakeholder meeting. Preliminary findings were also shared early on with the Methods Workgroup to aid discussions regarding key elements of the draft GPC RWE Decoder Framework.

CORE CONCEPTS OF AN RWE ASSESSMENT

Several core concepts of the draft RWE Decoder Framework emerged early in Methods Workgroup discussions, refined by feedback from the multi-stakeholder Advisory Committee, and later discussed with a broader stakeholder audience. These core concepts are listed below:

1. The value of RWE, within a given decision-making context, can be broken down into two distinct but equally important attributes: **Relevance** and **Rigor**.[4,5]
2. The RWE Decoder Framework can facilitate better understanding of the relevance and rigor of RWE and improve the way it is understood and used by decision makers, ultimately leading to better decisions for patients and populations.
3. Clear articulation of the decision-making context and research question is important for assessing the relevance of RWE. It can also improve the dialogue between decision makers and people who are primarily researchers or technical evidence assessors when considering methodological tradeoffs and thresholds for rigor.
4. The RWE Decoder Framework should be user-friendly, its output easy to communicate between technical and non-technical stakeholders.
5. Declarations of relevance are context-dependent and cannot be assessed using a "one-size-fits-all" scale. However, the *approach* with which relevance is assessed *can* be standardized.
6. Rigor is less subjective, however priorities, and acceptable tradeoffs in rigor, may vary.

IN-PERSON STAKEHOLDER MEETING

The draft framework developed by the Methods Workgroup was first presented to the Advisory Committee via webinar for initial reactions, and then formed the basis of a full day, multi-stakeholder in-person meeting. Forty-six stakeholder attendees represented a broad array of perspectives, including payers, health technology assessment organizations, patient advocacy organizations, professional societies and practice guidelines developers, life sciences companies, academic and proprietary research, regulatory decision makers, health care systems and accountable care organizations, comparative effectiveness, epidemiology, and outcomes research methodologists. An overview of the draft RWE Decoder Framework was provided to meeting attendees ahead of time, also as pre-read material within the Stakeholder Briefing Document. The agenda included an early multi-stakeholder panel discussion, small group breakout activities in which participants walked through the framework itself with the guide of hypothetical use cases (likely scenarios for RWE use), and larger facilitated discussions in which participants compared their experiences during the breakout activity and exchanged ideas to modify or improve the RWE framework.

INCORPORATION OF KEY FINDINGS AND FINAL FRAMEWORK

The Methods Workgroup continued to convene on a monthly basis and incorporate stakeholder feedback though multiple iterations of the RWE Decoder Framework. A revised draft framework was then made available to all project participants for a two-week period of open comment. Written comments were invited and incorporated into this final document.

[4] The methodological RIGOR, which impacts the trustworthiness of results, and RELEVANCE, or the direct relation of research findings to decision makers' questions.

[5] Generally, there are tradeoffs between the two, however, this is not always the case; a study may be highly relevance and highly rigorous, or it may be minimally relevant and minimally rigorous.

DESCRIBING REAL WORLD EVIDENCE AND ITS USES

The U.S. Food and Drug Administration (FDA) defines RWE as "information on health care that is derived from multiple sources outside typical clinical research settings, including electronic health records (EHRs), claims and billing data, product and disease registries, and data gathered through personal devices and health applications" (U.S. DHHS FDA, 2016; Sherman et al 2016). If we consider real world data (RWD) to be the elements from which RWE is derived, these data typically originate from sources other than traditional research or clinical trial settings; they are primarily collected as by-products of clinical care. The essence of RWE (previously discussed in draft FDA guidance on the use of RWE to support regulatory decision-making for devices), is that the collection of RWD and the production of RWE occur without causing any interference or course changes to normal clinical care and treatment choices for patients (U.S. DHHS FDA, 2016). In this way, RWE strives to truly represent the untampered actions and choices of patients, their doctors and health care providers.

Box 1 presents an independent working definition of RWE developed by the Methods Workgroup, for the purposes of guiding their work on this project and the development of the RWE Decoder Framework. Of note, RWE may include evidence from observational studies, but is not synonymous with observational research. RWE may come from multiple types of studies, including pragmatic randomized trials, for example.

Box 1. Real World Evidence Working Definition

Part 1: ***Evidence.*** Health services researchers produce many kinds of information that might be considered "evidence." This includes: (1) descriptive information on the burden of illness, provider performance, the rate of adverse events, costs, and utilization; (2) evidence about whether interventions "work," that is they effect outcomes of interest (positively or adversely), as well as for whom and in what contexts; and (3) how and why the intervention works, and how a model can be amended to work in new settings. The primary focus of the RWE Initiative is (2) and (3). Note: sometimes evidence isn't sufficiently rigorous to be sure that the intervention and the outcome are causally related, but that causal relationships are implicit in "works" or "effect outcomes of interest."

Part 2. ***Real World***. This implies that the evidence was generated during the delivery of healthcare in realistic settings. Typically, observational evidence, i.e. not randomized, but PCTs and stepped-wedge randomized cluster designs conducted in real practice may also be considered RWE.

There has been growing interest in understanding the "real-world" safety and effectiveness of drug therapies, medical product developers and other clinical interventions. Researchers increasingly use data collected outside the traditional confines of controlled research studies (i.e., RWD) to evaluate product performance under typical clinical conditions. This trend, driven in part by the infeasibility of conducting traditional RCTs for most important topics, is accelerating with the rapid growth in large integrated datasets that glean information from electronic health records, administrative data, laboratory results, and other sources (Kahn et al, 2015).

There is also a growing literature of comparative effectiveness, policy, and other contemporary research questions that can be well addressed by RWE. For example, Dacks et al (2016) call out the need for research that demonstrates how choices in the clinical care and treatment of common chronic diseases, such as diabetes or hypertension, may protect or accelerate cognitive decline and dementia in Alzheimer's patients. RWE can also play a pivotal role in coverage and reimbursement decisions, for example, collecting and analyzing data on expensive specialty drugs such as clotting factor concentrates for hemophilia care, ultimately to influence understanding of potential ways to reduce adverse events, improve outcomes and achieve future cost savings (Berger et al, 2016). In therapeutic areas with rapidly increasing spending such as oncology, RWE can help assess the effectiveness of new specialty pharmacy programs to improve patient adherence to treatments and reduce costs (Molina et al, 2016).

With many potential applications for health care decision-making, RWE can, in some cases, be more directly relevant and informative than an RCT, or other high quality study such as an epidemiology cohort study, using data collected explicitly for research purposes. RWE can answer questions of real world effectiveness, value and net costs of care, heterogeneity of treatment effects, and other patient-important outcomes such as tolerability of treatments (Galson and Simon, 2016). However, the traditional evidence hierarchy that many evidence assessors adhere to gives heavy preference for RCTs regardless of relevance, and inconsistent or low consideration of RWE. This prevailing prioritization of evidence from RCTs and de-prioritization of RWE can limit the consideration of high quality evidence that could aid decision-making.[6,7]

The quality of RWE is often difficult to ascertain, in part because of the "unknowns" associated with the results produced in settings or circumstances not as carefully monitored or as clearly understood as traditional research settings, and confounded by a variety of factors. Currently, decision makers consider RWE in highly variable ways that are not always transparent. Many rely, or have relied, on using a traditional evidence hierarchy to assess the quality of RWE. Several initiatives have begun to address this challenge from the bottom up, or to expand existing guidance or standards for the design, reporting and assessment of bodies of evidence, to RWE. Among the initiatives we reviewed no single framework for assessing the quality of RWE has penetrated the broad national population of RWE users or demonstrated systematic and consistent adoption by multiple stakeholder organizations for assessing RWE.

As yet unanswered are the following questions:

1. How frequently decision makers are utilizing these guidelines and best practices; and
2. Whether the broader universe of end users actually use any of these resources consistently to asses RWE.

In addition, the majority of tools described above were developed by committees or working groups comprised primarily of academic or industry experts, with limited participation from decision makers. As such, much of the content and many of the accompanying checklists, tools, etc. assume a certain level of technical expertise of the user and can also be time-intensive to use.

CHARACTERISTICS OF RWE USERS

Our review of potential users of RWE include clinical guideline developers, health plan formulary and medical policy committees, health technology assessment (HTA) groups, and other evidence review organizations that support these decision makers. Risk-bearing provider groups and delivery systems, including accountable care organizations, are also becoming increasingly important users of studies on comparative effectiveness and value. Regulatory stakeholders have also begun to articulate the potential application of RWE to certain types of regulatory decisions (U.S. DHHS FDA, 2016; Sherman et al 2016; Aleymayehu and Berger, 2016).

[6] Usually, within a GRADE process for assessing strength of evidence, RCTs initially start with a provisional grade of "high" strength of evidence, which can then be modified or lowered based on assessment of study limitations; observational studies typically start with a provisional grade of low, or moderate for certain circumstances, but can then be adjusted or raised based on assessment of study limitations (Guyatt et al 2008; Berkman et al, 2015).
[7] The RWE Decoder is not setting itself up as a tool to judge RWE "versus" RCTs. High quality observational cohort or other epidemiological studies are also often considered above RWE by decision makers. The objective is to shift thinking toward a more open-minded and transparent assessment of RWE that considers both relevance and rigor.

In our key informant interviews with users, we found variation in the level of sophistication and uses of RWE. Respondents reflected on using RWE in clinical policy decisions, economic decision-making, and benefit design. One respondent noted the application of RWE in determining if best practices are utilized at the point of care; another observed its usefulness in identifying non-responders to first line therapies.

Other RWE uses included:

- Evaluating total cost of care for a specific condition;
- Informing policy or position statements in cases where practice guidelines don't extend;
- Understanding the "clinical experience of our membership", and using that information as the basis for policy and programmatic decisions; and
- Supplementing evidence of safety.

Some noted that the acceptability of RWE might depend on the therapeutic area and the "seriousness of consequences" of decisions based on RWE. One respondent remarked that RCTs provide only a limited representation of patient experiences, and observed that RWE can encompass the experiences of many more individuals. Another noted that RWE is useful in the context of shifting from a "volume-based to a value-based" healthcare delivery system.

When assessing the quality of RWE, several key informants noted that they and their colleagues utilized or adapted a modified version of the GRADE evidence framework. However, some felt that within the GRADE context, RWE is generally and preemptively regarded as lower quality, and that this particular framework might not be the best fit for assessing RWE.[8] Overall, however, respondents gave examples related to the quality of the data and the credibility of the data source. Their responses also reflected consideration of the size and representativeness of the population being studied, the means of validating data, and whether results have been demonstrated over an adequate period of time and under multiple circumstances. For some, an important factor was determining if the evidence generated was hypothesis-driven, as opposed to being established post hoc, or based on data "dredging". Respondents also described various challenges to the use of RWE for decision-making. Several respondents noticed a bias in favor of published, peer-reviewed evidence. Others commented on the timeliness of RWE, and the fact that decisions sometimes must be made before high quality or relevant evidence is available. One respondent noted that evaluating the impact of policies and programs developed based on RWE rarely involves a "quick turnaround", and that measuring the impact of RWE in assessing a clinical or cost-related question requires a substantial investment of time. The lack of a "common language" in electronic health records and other data sources was another noted challenge.

Those interviewed appear to fall into three distinct types of RWE Users:

1. Those who want to "dig in" to examine the methodology, understand the dataset, and determine potential sources of bias and the limitations of the findings. These groups might assess a body of evidence to draft or publish clinical practice guidelines, or might form a committee within a payer organization, to assess RWE to inform a new national policy. More

[8] Technically, within the GRADE approach, different study types start at different default levels of quality: randomized trials start at a high quality and observational studies at low quality, through the GRADE assessment these levels can ultimately change (Guyatt et al 2008). We should remind readers that the definition of RWE developed within this GPC RWE project does not exclude randomized trials from potential sources of RWE.

broadly, these groups are publicly accountable to a larger membership, need to be very confident in their decisions and transparent in how they go about making them.

2. Those juggling multiple competing priorities, who may not have the time or interest in reading a full article unless they have some sense of the value they will derive from it. Instead they take a cursory look, perhaps read an abstract, and ultimately want to know, "is this something I should look into further?"
3. Those with limited technical knowledge beyond a basic understanding of statistics (e.g. p-values and confidence intervals), who feel less comfortable assessing the evidence themselves. This group looks for supporting guidance or easy to follow resource to explain how available RWE supports or opposes particular types of decisions and questions.

Any evidence framework designed to meet the needs of these users must balance a series of tradeoffs, which was the task undertaken by the Methods Workgroup with input from other stakeholders.

CREATING A FRAMEWORK

In our engagement with stakeholders in the full-day in-person meeting, there was broad agreement that the type of framework discussed below would help facilitate conversations between decision makers, or between decision makers and their evidence-assessing colleagues, regarding the merits of RWE available for consideration during decision-making. In addition, it could contribute to a growing awareness among decision makers that relevance of RWE should be considered independently and as seriously as methodological rigor. It is also hoped that the framework can enable greater understanding among decision makers of their own needs, in turn leading to greater communication and translation between stakeholder questions and formatted research questions that can be answered by scientific query. Though utilizing a Likert-type scale to assess both relevance and rigor, the framework ultimately encourages deliberate consideration of evidence followed by an “informed judgment”; it thereby avoids imposing a non-validated rating scale or assigning numeric assessment of RWE value that may not fit all situations. The visual summary (below) as well as inclusion of effect size (represented by the size (area) of the “bubble”) is also novel and important to decision makers.

The final version of the framework is presented in Table 1. A Decision Maker’s Framework for Assessing Real World Evidence. Below we walk through this framework and provide additional details regarding the intentions and content within each module, as well as key points of deliberation. Throughout this section we refer to the “user” to indicate the decision maker, evidence assessor or any other stakeholder who uses RWE to guide decision-making.

Table 1. A Decision Maker’s Framework for Assessing Real World Evidence.

MODULE 1: ARTICULATING THE QUESTION	
Q1. What is the nature of your decision? Q2. What do you want to know? Q3. What do patients want to know (e.g. which outcomes are most important to patients?) Q4. What is your research question, rephrased following PICOTS format?	
MODULE 2A: ASSESSING THE RELEVANCE OF EACH RWE STUDY	
PICOTS: Population – **I**ntervention – **C**omparator – **O**utcomes (Primary, Secondary) – **T**iming – **S**etting	
MODULE 2B: ASSESSING THE RIGOR OF EACH RWE STUDY	
Quality of Research Question	• PICOTS Stated • Appropriateness of study design
Potential for Bias	• Confounding bias • Selection bias

	• Bias from classification of intervention • Bias from deviation from intended intervention • Bias from missing data • Bias from measurement outcomes • Bias in results reporting Additional sources (randomized studies): • Performance Bias • Detection Bias • Attrition Bias • Other (e.g. contamination, recruitment bias)
Precision	• Confidence interval
Data Integrity	• Data source & intention • Completeness (absence of missing data) • Fidelity (e.g. a female is coded as a female) • Plausibility (e.g. are the data believable) • Cohort construction & linkage
MODULE 2C: MAGNITUDE AND DIRECTION OF EFFECT*	
Effect Size (primary outcome)	Direction (-, +, or no difference)
MODULE 3: RWE FRAMEWORK VISUAL SUMMARY	
Data point (one per study): (Relevance [X], Rigor [Y], Effect Size [bubble size], Direction [bubble color])	

MODULE 1. ARTICULATING THE QUESTION

One challenge identified early in background research and amplified by stakeholder discussion is that the questions of decision makers are not always well articulated or explained in a way that is directly addressable by research. The intention of Module 1 is to encourage the user to identify his or her needs, articulate a question and translate that question into a format that can be communicated to a research perspective. Module 1 was also motivated by the assumption that context matters when assessing RWE for decision-making, and it is helpful to encourage users to identify up front what their particular decision-making challenge and contextual factors that may shape their expectations for RWE.

Module 1 instructs users to express what it is they want to know and reshape their question into a structured research question using the PICOTS (Population, Intervention, Comparison, Outcomes, Timing, Setting) format (Velentgas et al, 2013). Given the prior existence of the PICO(TS) framework, and its purpose and familiarity among many in the research community, its domains make a natural template for assessing relevance in the RWE Decoder Framework. For example, if you are a Hispanic woman who has just been diagnosed with high blood pressure, non-smoking, and 55 years old with a family history of [x, y, or z], your question may need to be translated into a population-level question that can be answered for you and other patients who are similar to you. You (the patient) may also be more interested in particular outcomes or side effects than your doctor, or health insurance provider, who are likely more interested in specific clinical outcomes of disease progression or outcomes of resource utilization. Your priorities and the outcomes that matter may also shape a question. As a patient, your question of “which blood pressure drug is the best option for me” in a conversation between you and your doctor, might, in a different conversation between you and a researcher, translate into something more structured and specific, such as:

> "Among Hispanic women between the ages of 45-60 with a family history of [x, y, or z], who are recently diagnosed and have never previously been treated for high blood pressure, is [Drug X] more effective than [Drug Y] at reducing [Outcomes]?"

The importance of facilitating understanding between decision maker and researcher or between decision maker and technical evidence assessor was very prevalent in the multi-stakeholder discussion on June 29th. During that discussion, stakeholders suggested adding an explicit question in Module 1 that would bring in the patient perspective by, at minimum, causing users of the framework to pause and consider whether there were patient-important outcomes and whether their question considered the needs and priorities of patients. Thus, Module 1 was modified to ask the user to specifically highlight patient-important aspects of the question so as to keep the patient perspective near the forefront of conversation about the merit of studies generating RWE.

MODULE 2A. ASSESSING THE RELEVANCE OF EACH RWE STUDY

Module 2A is an exercise to assess the relevance of a real world study or analysis to the question or circumstance faced by the user. The PICOTS formatted question articulated in Module 1 forms the baseline or ideal against which relevance is assessed. The tabular formatted research question in Module 1 is copied into Module 2 as reference. Continuing the example from the previous section, this might look like:

Population	Hispanic women between the ages of 45-60 with a family history of [x, y, or z], who are recently diagnosed and have never previously been treated for high blood pressure
Intervention	Drug X
Comparator	Drug Y
Primary Outcome(s)	Blood Pressure (measured in mm/Hg)
Secondary Outcome(s)	Patient-important side effects (e.g. drowsiness, dizziness, constipation, etc.)
Timing	2-4 weeks
Setting	Usual care

Considering each of the PICOTS domains and then using a Likert-type scale, the user 'rates' the overall relevance of a study to their own unique question.

The assessment of relevance is ultimately unique to the user and highly subjective. The intention of Module 2A is not to fit a one-size scale for rating relevance, but rather to standardize the way decision makers approach assessing the relevance of studies generating RWE. In doing so, there is an opportunity to create: a) better understanding of evidence needs; b) transparency; and c) more meaningful discussions between decision makers, evidence assessors, and researchers regarding the priority and relative value of different aspects of evidence quality.

MODULE 2B. ASSESSING THE RIGOR OF EACH RWE STUDY

The Methods Workgroup converged on four major domains of rigor for those assessing RWE for decision-making: 1) Quality of the Research Question; 2) Risk of Bias; 3) Precision; and 4) Data Integrity. Within these four, additional factors or elements of rigor, common across multiple method standards,

tools and reporting requirements,[9] can be grouped. Below we provide details and discuss the rationale for including each domain in the RWE Decoder Framework.

Quality of Research Question

There is often skepticism of RWE and those less familiar or comfortable with understanding rigor may understandably question the motives behind study or analysis design decisions. Knowing that a specific question was stated (or published) up front, and which directly and transparently influenced the selection of the most appropriate or feasible study and analytic designs, adds a level of confidence to those decision makers and evidence assessors who might otherwise question the importance of a study to a given clinical or health care decision. Users are instructed to consider, in particular, 1) whether and how clearly the PICO(TS) elements of the research question are articulated, and 2) the appropriateness of the study and analytic design specific to the study aims. Examples are provided in the ***User's Guide***.

Potential for Bias

There are multiple tools for assessing bias in randomized and nonrandomized studies. Bias is a major domain of concern to methodologists and evidence assessors, however understanding its sources and the different types of bias that exist can require detailed knowledge and complexity, which many decision makers may lack. The challenge of bias is understanding potential problem areas, sources of bias that may or may not be reasonably addressed, *and*, whether or not the existence or potential for bias influences the usability of study findings. These considerations often require discussion between a decision maker and more technologically savvy evidence assessor. For the RWE Framework, we adopted the considerations of bias as developed by a Cochrane Collaboration group (Sterne et al, 2016; Higgins et al 2011). These potential sources of bias have been identified and vetted by foremost methodological experts and overlap with other leading initiatives for assessing the rigor or credibility of observational research findings. For nonrandomized studies, potential sources of bias include: confounding bias, selection bias, bias from classification of the intervention, bias from deviation from the intended intervention, bias from missing data, bias from the measurement outcomes, and bias in results reporting (Sterne et al, 2016). Potential sources of bias for randomized studies include: selection bias, performance bias, detection bias, attrition bias, reporting bias, and other biases (unique to particular designs or situations such as cluster randomization or case-crossover designs) (Higgins et al, 2011).[10] Examples are provided in the User's Guide, which is written to address most directly those who are perhaps practiced users of research but less familiar or trained in the critical appraisal of methodological rigor.

Precision

Precision is a key consideration for decision makers, whose criteria for RWE go beyond just whether the study was "done well". This domain came up when thinking about the tradeoffs one might accept for less formal or more time-pressed decisions. Decision makers must decide at what threshold a change in care or policy is warranted, and want to feel confident that a change will result in meaningful and observable differences for their patients and populations. There were no disagreements from the general audience of the in-person stakeholder meeting regarding keeping Precision as a domain of Rigor in the RWE Decoder Framework. In the initial draft, Effect Size was included in Module 2B; it was recommended and agreed upon by the Methods Workgroup, that Effect Size be pulled from Module 2B

[9] See Appendix 2.

[10] The most current version is available online at: [http://methods.cochrane.org/bias/assessing-risk-bias-included-studies]

but given its own consideration (Module 2C) and included in the final visual summary of Module 3 as a function of the bubble size (area).

Data Integrity

Despite numerous sources and growing availability of RWD, challenges in terms of assessing their integrity and quality also amount. Key informants expressed a real discomfort with trusting evidence from unfamiliar or new RWD sources. They want to see clear parameters of integrity, yet in most cases, published research reports lack that level of detail. Box 2 lists key words or phrases related to data integrity that were heard during key informant interviews, and which shaped Module 2B. Given the lack of published information, other possible criteria for assessing data integrity might include: Are there other published studies that utilized the same dataset? Did the researchers conduct double extractions so as to, at minimum, eliminate human error in the extraction process? Do rates within the study sample make sense and resemble rates in other, similar populations?

Box 2. Key Words and Phrases for Data Integrity

- How the data were curated [e.g., how data were managed through their lifecycle, from creation, to storage, to retrieval for use]
- Missing Data
- Validating information with other sources
- Data contributors involved with researchers
- Nuance of the data
- [Gaps in data] preclude inferences about causality
- Quality control checks
- Large enough population
- Double data extraction
- Hypothesis driven analysis versus phishing expedition (goes to first subdomain)
- Collaborativeness of work
- Data taken out of context

Initial elements of data integrity in the RWE Decoder Framework included: data source and coding, data reliability/validity, and cohort construction. However, this particular domain of rigor remained the furthest from consensus during stakeholder discussions, the criteria being quite technical and assuming access to detailed information often excluded from published reports. These discussions shed light on the dilemma of disseminating a detailed, methods-minded assessment of data quality when in reality that information is hardly available. The choice facing the GPC RWE Initiative thus became whether to include specific (and more ideal) elements of data integrity in the framework, and anticipate that many decision makers will have difficulty completing that particular assessment due to lack of information, or, propose alternate considerations of data integrity that, though less stringent, are more likely to be assessable.[11]

Key Considerations from Other Initiatives

Kahn et al (2015) and others note that data quality is often context dependent. The same data sources or data elements may be high quality for one particular use while deemed low quality for another; this is because different applications for data call on different variables over others. The term “fitness for use” often describes the context dependent nature of data integrity. The *intention* of the data collected is a repeated theme and high level consideration for FDA acceptance of RWE for regulatory decision making for devices (U.S. DHHS FDA, 2016). The RWE Decoder Framework also recommends decision makers consider the data source and its original intent when assessing available RWE.

[11] One could argue that demanding more information up front will urge greater transparency and reporting by researchers and producers of RWE in the future. Of note, Basu et al point to a growing availability of online annexes to published studies (Basu et al, 2016).

MODULE 2C. EFFECT SIZE AND DIRECTION

In early drafts, the study sample size was included as a measure of precision and then incorporated as the "bubble size" or size of the data point in the visual summary (Module 3), however most stakeholders ultimately felt that sample size was not as meaningful to them as effect size and effect size thus replaced sample size.

Key informants emphasized the need not only for high quality evidence but evidence demonstrating clinical differences, or meaningful differences in effect in order to justify many of their decisions.

> "Rather than focusing on statistical difference, we're trying to take the tact of, 'how big does the difference have to be before you should be worried?' It's not: do A and B differ? It's: is A different enough from B that you should pause and say, 'Houston, we have a problem.'"

This preference was confirmed by attendees at the stakeholder meeting. The way users of the RWE Decoder Framework are now instructed to describe effect size in Module 2C, they make a context-dependent, subjective indication of the meaningfulness of the given effect estimate on the primary outcome variable. Though initially most agreed they were less likely to attribute larger effects to bias, several attendees challenged the group to consider what ought to be done when a well-designed rigorous study, such as a pragmatic trial, does not result in a large estimate: "…a small but meaningful effect can be very expensive, or a big cost savings." The RWE Decoder Framework attempts to resolve this issue in Module 2C by having users of the RWE Decoder Tool indicate the meaningfulness of the effect size, as opposed to simply recording the value. The ***User's Guide*** further encourages decision makers to consider the size of the study population as well as the size of the target population for which a decision might have an impact, before indicating the meaningfulness of the effect size. In Module C, effect size is then explicitly recorded and clearly incorporated into the visual summary as a function of bubble size (area), after the user assesses rigor, in an attempt to reduce bias in ratings due to (anticipated) rater preference for larger effect sizes.

MODULE 3. VISUAL SUMMARY OF RWE ASSESSMENT

Module 3 provides a visual depiction of available evidence, the data point for each study being plotted along a continuum for Relevance (x-axis) and Rigor (y-axis). The idea and novelty of Module 3, a visual summary of the information generated by the user's assessment of available RWE, was well received by stakeholders across different aspects of the project, including Advisory Committee, Dissemination Workgroup, and in-person meeting attendees.

In the display below, the size (area) of the data point corresponds to the meaningfulness of the effect size. Green bubbles indicate positive effect on the primary outcome, white bubbles indicate a negative effect, or no difference. As data points move toward the upper right-hand corner of the plot area, they are more relevant and more rigorous. In this example, Study 7 is minimally rigorous, only somewhat relevant, and the effect size is minimally meaningful.

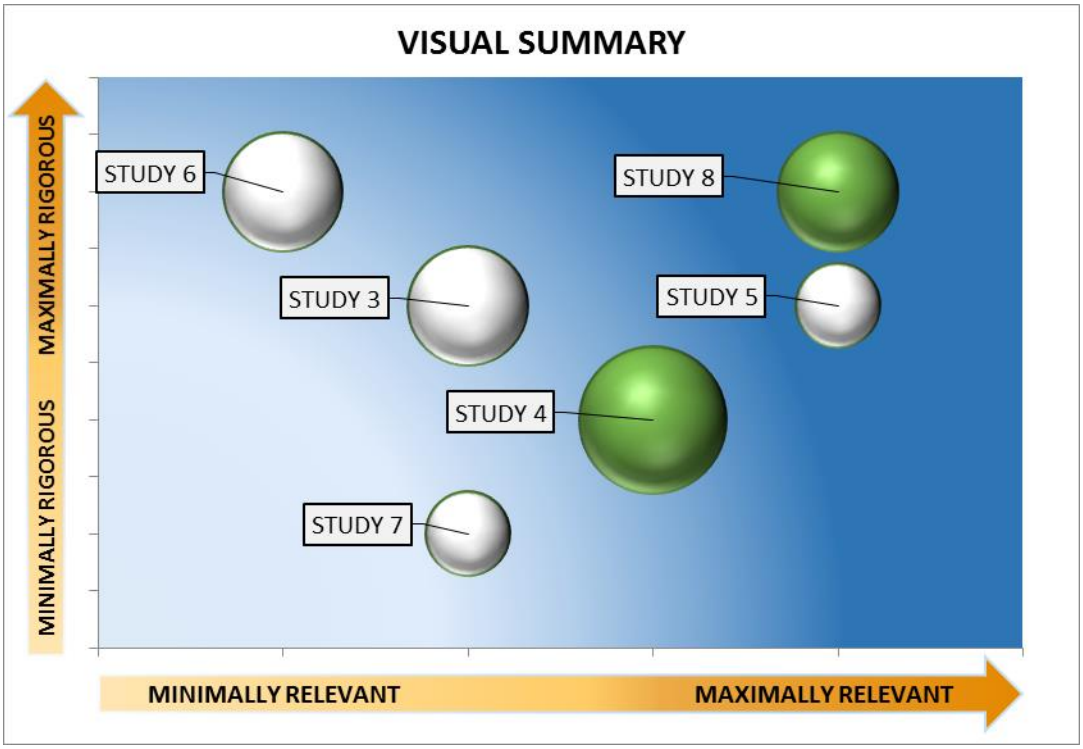

Figure 2. GPC RWE Decoder Tool Visual Summary (Module 3)

Currently, users check boxes in the RWE Decoder Tool to indicate domains and subdomains for relevance and rigor (Modules 2A and 2B, respectively) for which information was reported or available for them to assess. If additional tools are created using other media, such as desktop software, the number of boxes checked can correspond to darker or lighter shading within the bubble to indicate how much or how little information was incorporated into the assessments of relevance and rigor. This concept arose from a request at the stakeholder meeting to find some way to visually indicate to the user how confident one might be in the final assessment and plotting of each study.[12]

Version 1.0 of the RWE Decoder Tool also builds in "quick fail" conditions allowing users to tag studies which lack essential aspects, rendering them unusable to the decision maker. Studies given "0" in any domain (of relevance or rigor) are dropped from the final visual summary. This approach avoids forcing a full and unnecessary assessment of studies that won't be informative to a particular decision. Users of the framework are instructed to only give a "0" to a domain if the study misses or fails to address a critical element or high priority domain, thus rendering it unusable by the decision maker. However, before continuing onto the next module, the user must select the particular criteria for which the "0" applies. This suggestion was posed during the in-person stakeholder meeting, and discussed as an alternate to a more intensive process of ranking and weighting each domain. It is, in fact, a tool also employed by the CER Collaborative.

[12] Appendix 3 provides an example application of the GPC RWE Framework to a hypothetical use case.

DISCUSSION

The GPC RWE Decoder Framework represents a compromise between the need for sufficient detail in evidence assessment and the struggle to simplify what is, to many, a prohibitively cumbersome process. It was developed, not with the intention to generate formal analyses or quantifiable assessment, but rather to guide decision makers towards an "informed judgment" and approximation of the quality of RWE available.[13] Box 3 proposes likely scenarios for use of the RWE Decoder Framework, identified during the course of the project. Unlike most existing evidence assessment frameworks, the purport of the RWE Decoder Framework is not a "by methodologists, for methodologists" tool. Its value may thus be challenged under this particular lens. However, before methodological standards catch up to the growth and proliferation of RWE, innumerable decision makers make, and will continue to make out of necessity, important decisions without any clear framework for guidance. The RWE Decoder Framework can fill this gap and help drive a paradigm shift in the way RWE is considered and used by decision makers, from highly-variable, untrained and inconsistent to a more transparent, confident, and shared approach.

Box 3. Scenarios for Use of the GPC RWE Framework

- Communication between decision maker and analyst, epidemiologist or health technology assessor
- Internal deliberation among decision makers & staff
- Assessing RWE to inform formulary or policy committee
- Communication between researchers & their audience
- By an individual facing important short term decisions without support staff
- To identify gaps in existing evidence

The RWE Decoder Framework assumes that rigor and relevance are the two primary domains with which to assess RWE for decision-making. These assumptions were agreed upon or at minimum, unchallenged in stakeholder and Methods Workgroup discussions, though some debate persisted over wording. We believe that the definition and distinction of both concepts will be intuitive to decision makers, including those with less methodological training. Encouragingly, though their audiences are not identical to the GPC RWE Decoder Framework audience, two initiatives did in fact converge on a similar assumption of rigor and relevance as primary domains, through parallel and independent work. When describing the quality of RWD and in turn, RWE, the FDA articulates two aspects: relevance and reliability of the data source and its elements (U.S. DHHS FDA, 2016). CER Collaborative takes a similar two-dimensional approach (Berger et al, 2014), with a standard questionnaire for relevance across all study types, to accompany detailed assessments of credibility (instead of "rigor") for different types of study designs.

BENEFITS OF THE GPC RWE DECODER FRAMEWORK

Although one may argue to promote thorough assessment of evidence across the board, the reality is that decisions are made daily without sufficient evidence, or without a clear understanding of what a body of evidence might suggest. The GPC RWE Decoder Framework offers a manageable approach for decision makers who currently lack the means to systematically assess RWE for decision-making, or who otherwise have limited training, staff, or time dedicated to the regular assessment of research evidence for decision-making.

[13] At the stakeholder meeting, attendees exchanged ideas related to the potential value of the RWE Framework, noting gaps or needs the project and deliverables might help to address. The RWE Decoder Framework can get a decision maker "in the ballpark" of whether a study is contributing more or less to the body of evidence. One participant noted that in addition to published studies, the framework could also be useful for evaluating "those data mining or ad hoc kinds of queries of a database where you are trying to answer a specific question."

While the "rules of rigor" may be established, we assume stakeholder tolerance for different levels of rigor is variable. Through our background research, we must infer that some stakeholders may not have sufficient technical knowledge to be able to articulate their own context-specific tolerance for uncertainty. However, like any cultural shift in research, we look to achieve stepwise progress instead of overnight change. The current need and potential benefit of this framework is not to fortify overnight total consensus on methodological standards for RWE assessment. The benefit of this framework is to create an entrée into understanding and better use of RWE for the broader universe of stakeholders and decision makers who currently, or will soon, make informed decisions based in part on RWE. Users of RWE look to evidence to provide support for timely decisions that will have tangible consequences for their patients and providers. The term “timely and actionable evidence” is used often in discussions of learning health systems and the use of routinely collected health information to generate evidence to help inform care decisions (Brown et al, 2013; Olsen, 2007; Young et al, 2010; IOM 2011). Hubbard and Paradis (2016) call for RWE that is “fit for purpose” and a fostering environment in which stakeholders develop and communicate expectations for how RWE will be used in decision-making. The GPC RWE Decoder Framework can help to facilitate such a paradigm shift toward a “culture of high quality RWE.”

Furthermore, since RWE often complements other types of evidence, other studies can be added to the RWE Decoder Tool and plotted together with RWE studies to map out a more comprehensive picture of available evidence for decision-making. This is possible because an overall assessment of relevance and rigor could be made of any available evidence. However, the content of the modules was developed with RWE in mind. Module 2B (assessing rigor) in particular, may require additional knowledge of or review of alternate frameworks to guide judgment of methodological rigor. In this way it may be possible to assess RWE and consider it within a broader context of available research, thus seeing where and how RWE complements other types of available evidence.

RWE FOR COVERAGE AND REIMBURSEMENT DECISIONS

Many argue that articulating and integrating the needs of payers into the pharmaceutical development process will improve the likelihood of success while also facilitating timeline access to new products for patients (Dunlop et al, 2016; Epstein et al 2012). It is our belief that, in addition to improving the confidence, consistency, and transparency in which decision makers approach RWE for decision-making, broad adoption of the RWE Decoder Framework will lead to greater clarity among the developers and generators of RWE regarding the needs of decision makers and how the evidence they bring to the table will be assessed and utilized. Although many groups have sophisticated methodological training and standards in place for assessing the quality of evidence, there is a much broader universe of decision makers, from ACOs to health systems, small payers, provider groups, etc., with different levels of expertise, time and other available resources to be as thorough. Furthermore, the amount of training required to utilize such frameworks can be prohibitive.

The GPC RWE Decoder Framework will encourage greater transparency in the way RWE is utilized and evaluated by decision makers. This can in turn benefit for decision makers looking to build trust among their covered populations and consistency in their evidence assessments and provide better clarity for developers of RWE who will come to better understand expectations and standards for RWE, including how the evidence they develop will be assessed or used by decision makers.

PATHWAY TO STANDARDS

It is hoped that in time, studies will be developed to improve the confidence stakeholders have in relying on RWE for decisions that will have impact. The question of whether the GPC RWE Decoder Framework could eventually lead to the development of standards for RWE was raised during the project. Its adoption might, for example, help to standardize the articulation of questions, helping to identify the factors or domains that are most important to decision makers, which can in turn help to better understand rigor and relevance. Though many published studies don't articulate PICO(TS) explicitly, this framework may help drive more structure in the way relevance is presented in published studies. As decision makers adopt this framework, evidence developers and researchers writing publications might also use the framework to communicate their study's usefulness for a given application or specific type of decision.

LIMITATIONS OF THE GPC RWE DECODER FRAMEWORK

The GPC RWE Decoder Framework is an imperfect approach to assessing the value of RWE, and not as rigorous as many existing evidence assessment frameworks. It is not intended to be an exact and numeric assessment. Rather, it is intended to provide an informed approximation of rigor (and relevance) and a way for decision makers to identify where they perceive potential problem areas in using a piece of evidence.

Another debate within the RWE Initiative is the extent to which the RWE Decoder Tool might be applied to unpublished or non-peer-reviewed studies. In *principle*, the RWE Decoder can apply to any kind of study. And since much evidence of interest is not or not yet published at the time of decision-making, the RWE Decoder may need to be used more inclusively. The challenge, on the one hand, is that much unpublished research lacks the information needed to situate a study in the rigor-relevance landscape. On the other hand, that may be true for some published studies as well.

In Version 1.0 of the RWE Decoder Tool, we therefore propose the following solution, regardless of publication status:

1) In the Rigor dimension, do not give a study credit for having done something that would increase its Rigor score, unless there is evidence that the authors actually did it; and
2) In the Relevance dimension, if the study is of no value on a critical dimension of PICOTS, for instance, assume that it is not relevant to the question at hand. (On the tool, this correlates to assigning a "0" to that particular domain of Relevance, thereby excluding that study from the final module.)

NEXT STEPS

Ultimately, it is our intention that the GPC RWE Decoder brings clarity to decision makers regarding the value of RWE and helps decision makers and other assessors of RWE become more confident in their understanding and utilization of RWE for decision-making. As a testament of our commitment to this goal, we have identified additional follow-up work streams, described below, which can be feasibly implemented in the following calendar year.

TESTING AND DEMONSTRATION

The Green Park Collaborative has developed a free and downloadable Excel Tool (RWE Decoder, Version 1.0) for applying the GPC RWE Decoder Framework, available on the CMTP Website. We will continue to vet and test the framework for usefulness, relevance, reliability and interpretation and build up empirical evidence to support a next version or iteration. This should include, but is not limited to: growing a library of use cases, video demonstrations for users, online evaluation form with Version 1.0,

additional training tools based on feedback. Broader testing may also lead to adoption, expansion, or alternate versions in discrete areas such as assessing RWE on devices or diagnostics.

In addition, GPC has posted an open invitation to the release page of the RWE Decoder, recruiting interested individuals and organizations from different stakeholder perspectives to join a User's Group to contribute data from their own experiences using the RWE Decoder, help develop evaluation strategies to better understand how users apply the RWE Decoder, identify additional educational tools or dissemination strategies to increase uptake, and participate in follow-up pilot activities and beta testing of potential software solutions.

The RWE Decoder Tool, additional supporting documents, and a sign-up for the User's Group or additional information is all available on the CMTP Website.

ADDITIONAL TOOLS

GPC has identified supplemental tasks to pilot the RWE Decoder Framework and develop additional tools for users. These include, at minimum, focus groups and early adopter survey to inform the next iteration of the RWE Decoder Tool (Version 2.0), as well as the development of a desktop software solution. Pending adoption and feedback from early users, other avenues, such as a mobile application, will be explored.

REFERENCES

Aleymaychu D, Berger ML. Big Data: transforming drug development and health policy decision making. Health Serv Outcomes Res Methods 2016; 16:92-102.

Basu A, Axelsen K, Grabowski DC, et al. Real world data - policy issues regarding their access and use. Medical Care 2016; 54(12): 1038-1044.

Berkman ND, Lohr KN, Ansari MT, et al. Grading the strength of a body of evidence when assessing health care interventions: an EPC update. J Clin Epidemiol 2015; 68: 1312-24.

Berger KC, Feldman BM, Wasserman J, Schramm W, Blanchette V, Fischer K. Securing Reimbursement For Patient Centered Haemophilia Care: Major Collaborative Efforts Are Needed. Haematologica 2016; 101: 266-268.

Berger ML, Martin BC, Husereau D, et al. A questionnaire to assess the relevance and credibility of observational studies to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force Report. Value in Health 2014; 17: 143-156.

Brown J, Khan M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care 2013; 51 (803) S22-S29.

Dacks PA, Armstrong JJ, Brannan SK, et a. A call for comparative effectiveness research to learn whether routine clinical care decisions can protect from dementia and cognitive decline. Alzheimer's Research & Therapy 2016;8:33.

Dreyer NA, Bryant A, Velentgas P. The GRACE Checklist: a validated assessment tool for high quality observational studies of comparative effectiveness. J Manag Care Pharm 2016; 22(10): 1107-13.

Dunlop WCN, Mullins CD, Pirk O, et al. BEACON: a summary framework to overcome potential reimbursement hurdles. Pharmacoeconomics 2016; 34: 1051-1065.

Epstein RS, Sidorov J, Lehner J, Salimi T. Integrating scientific and real world evidence within and beyond the drug development process. J Comp Effec Res 2012; 1(1): 9-13.

Galson S, Simon G. Real-world evidence to guide the approval and use of new treatments. Discussion paper, National Academy of Medicine, Washington, DC. Available online at [https://nam.edu/wp-content/uploads/2016/10/Real-World-Evidence-to-Guide-the-Approval-and-Use-of-New-Treatments.pdf]

Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008; 336:924.

Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ. What is "quality of evidence" and why is it important to clinicians? BMJ 2008; 336:995.

Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomized studies. BMJ 2011; 343: d5928.

Hubbard TE, Paradis R. Maximizing the potential of real world evidence to support health care innovation. Issue Brief 2016 (June). Available online at [http://www.nehi.net/publications/70-maximizing-the-potential-of-real-world-evidence-to-support-health-care-innovation/view]

Institute of Medicine. Digital infrastructure for the learning health system: the foundation for continuous improvement in health and health care: Workshop series summary. The National Academies Press; 2011.

Kahn MG, Brown JS, Chun AT, Davidson BN. Transparent reporting of data quality in distributed data networks. eGEMs 2015; 3(1): Art 7.

Molina I, Bongero D, Edillor F, et al. Impact of an oral oncology program in specialty pharmacy. In: Ostrovsky L. Implications of real world data and pharmacoeconomics for managed care. Am Health & Drug Benefits 2016; 9(3):151-155.

Morton SC, Costlow MR, Graf JS, Dubois RW. Standards and guidelines for observational studies: quality is in the eye of the beholder. Journal of Clinical Epidemiology,2015; 71: 3-10.
Olsen L. IOM Roundtable on evidence-based medicine. The Learning Healthcare System: Workshop Summary. Washington, DC: 2007.

Public Policy Committee, International Society of Pharmacoepidemiology. Guidelines for good pharmacoepidemiology practice (GPP). Pharmacoepidemiology and drug safety 2016; 25: 2-10.
Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-world evidence - what is it and what can it tell us? N Engl J Med 2016;375(23):2293-2297.

Sterne JAC, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016; 355:i4919.

U.S. Department of Health and Human Services Food and Drug Administration. Use of real world evidence to support regulatory decision-making for medical devices. Draft guidance for industry and Food and Drug Administration staff. July 27, 2016. Available online at [http://www.fda.gov/ucm/groups/fdagov-public/@fdagov-meddev-gen/documents/document/ucm513027.pdf]

Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM, eds. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. AHRQ Publication No. 12(13)-EHC099. Rockville, MD: Agency for Healthcare Research and Quality; January 2013. www.effectivehealthcare.ahrq.gov/Methods-OCER.cfm.

Young PL, Olsen L, McGinnis JM. Value in Health Care: accounting for cost, quality, safety, outcomes and innovation: workshop summary. The National Academies Press; 2010.

APPENDIX 1: PARTICIPANTS

ORGANIZATIONS

Aetna [0, 4, 5]	Agency for Healthcare Research and Quality (AHRQ) [4]
American Society of Clinical Oncology (ASCO) [0, 1, 5]	Anthem, Inc. [0, 5]
Astellas Pharma US [0, 1, 2, 3, 4]	Blue Cross Blue Shield Association (BSCBA) [4]
Centers for Medicare and Medicaid Services (CMS) [0]	Duke University [0]
ECRI Institute [4]	Eli Lilly & Company [0, 1, 2, 3, 4, 5]
EMD Serono, Inc. [0, 2, 4]	FasterCures [4]
Flatiron [1, 4]	US Food and Drug Administration (FDA) [0, 1, 4]
Fred Hutchinson Cancer Research Center (FHCRC) [1]	Genentech, Inc. [0, 3, 4]
Georgetown University, School of Nursing [1]	Greater Baltimore Health Alliance [0, 4,5]
Hayes [4]	Healthcore [0, 2, 4, 5]
Institute for Clinical and Economical Review (ICER) [4]	Institute for Policy Advancement Ltd. [4]
Intermountain Healthcare [0, 4]	Janssen Pharmaceuticals, Inc. [0, 1, 2, 3, 4]
Johns Hopkins Bloomberg School of Public Health [4]	Kaiser Permanente [0, 4, 5]
Maine Health ACO [4]	McMaster University [1]
MeYou Health [1]	Milliman [4]
Merck & Co., Inc. [3]	Mount Sinai Health Partners [4]
National Comprehensive Cancer Network (NCCN) [2, 4]	National Health Council (NHC) [4, 5]
National Pharmaceutical Council (NPC) [0, 1, 3, 4]	Optum Labs [2, 4]
Ottawa Health Research Institute [1, 4]	PatientsLikeMe [0, 1, 5]
Pfizer, Inc. [0, 1, 3, 4]	Premera BlueCross [4]
Priority Health [0, 5]	Quintiles, Inc. [0, 1, 4]
Sanofi, US [0, 3, 4]	Strategic Communications & Planning (SCP) [1, 2, 4]
Tufts Medical Center [0]	UCB Biopharma [0, 3, 4]
UnitedHealth Group [4]	University of British Columbia [1]
University of California, San Francisco (UCSF) [1]	University of Maryland, School of Pharmacy [0, 4]
University of Toronto [1]	University of Utah [4]
Virginia Tech, College of Science [1]	

KEY	
0 – advisory committee	**1** – methods workgroup
2 – dissemination workgroup	**3** – sponsors
4 – stakeholder meeting (virtual or in person)	**5** – other (topic vetting or outreach calls, key informant interviews)

ADVISORY COMMITTEE

Joseph Chin, MD *(ex officio)* Acting Deputy Director Coverage and Analysis Group Centers for Medicare and Medicaid Services	**Gregory Daniel, PhD, RPh** Deputy Director Duke-Robert J. Margolis, MD, Center for Health Policy Duke University
Nancy Dreyer, PhD Global Chief of Scientific Affairs & Sr. Vice President Real-world & Late Phase Research Quintiles	**Scott Flanders, PhD** Senior Director Health Economics and Clinical Outcomes Research- Oncology Astellas
John Fox, MD Senior Medical Director and Associate Vice President of Medical Affairs Priority Health	**Jonathan Jarow, MD** Senior Medical Advisor to the Center Director CDER Food and Drug Administration (FDA)
Sachin Kamal-Bahl, PhD Senior Director Global Health and Value, Innovation Center Pfizer	**Julie Locklear, PharmD** Vice President Health Economics and Outcomes Research EMD Serono
Joan McClure Senior Vice President Clinical Information and Publications National Comprehensive Cancer Network	**Elizabeth McGlynn, PhD** Director Center for Effectiveness and Safety Research Kaiser Permanente
Peter Neumann, ScD Director, Center for the Evaluation of Value and Risk in Health at the Institute for Clinical Research and Health Policy Studies Tufts Medical Center	**Sally Okun, RN, MMHS** Vice President Advocacy, Policy, and Patient Safety PatientsLikeMe
Tom Oliver Director, Guidelines Quality and Guidelines American Society of Clinical Oncology	**Eleanor Perfetto, PhD** Professor, Pharmaceutical Health Services Research, School of Pharmacy, University of Maryland Senior Vice President, Strategic Initiatives, National Health Council

METHODS WORKGROUP

Kristen Bibeau, PhD Director of Epidemiology and Real World Evidence, Center for Quantitative Methods and Information Sciences, Teva Pharmaceuticals	**George Browman, MD** Professor, Clinical Epidemiology & Biostatistics McMaster University Clinical Professor, School of Population and Public Health, University of British Columbia
Scott Flanders, PhD Senior Director Health Economics and Clinical Outcomes Research Astellas Medical Affairs – Americas	**Jennifer Graff, PharmD** Vice President Comparative Effectiveness Research National Pharmaceutical Council
Craig Henderson, MD Adjunct Professor Department of Medicine (Hematology/Oncology) University of California, San Francisco (UCSF)	**David Henry, MB, ChB** Professor, Institute of Health Policy, Management and Evaluation University of Toronto
Sachin Kamal-Bahl, PhD Senior Director Global Health and Value, Innovation Center Pfizer	**Mark Levenson, PhD** Deputy Director, Center for Drug Evaluation and Research (CDER) Office of Biostatistics US Food and Drug Administration (FDA)
Gary Lyman, MD, MPH Co-Director Hutchinson Institute for Cancer Outcomes Research (HICOR) Fred Hutchinson Cancer Research Center	**Sally Morton, PhD** Dean College of Science Virginia Tech
Jim Murray, PhD Research Fellow Global Health Outcomes Eli Lilly and Company	**Josée Poirier, PhD** Director Program Design & Research MeYou Health
Beverley Shea, PhD Senior Methodologist Ottawa Health Research Institute	**Michael Stoto, PhD *(Chair)*** Professor of Health Systems Administration and Population Health Georgetown University
Ming Zhang, PhD Senior Director, Group Leader RWE Design & Analytics Janssen Pharmaceuticals	

DISSEMINATION WORKGROUP

Aylin Altan, PhD Vice President Research Optum Labs	**John Beilenson, AB, MA (Co-Chair)** President Strategic Communications & Planning (SCP)
Rabia Kahveci, MD HTA Consultant Agency for Healthcare Research and Quality (AHRQ)	**Megan Klopchin** Associate Global Patient Outcomes & Real World Evidence Eli Lilly
Karen Lencoski, MBA Director Therapeutic Area Government Strategy Astellas	**Julie C. Locklear, PharmD, MBA** Vice President & Head Health Economics & Outcomes Research EMD Serono
Joan McClure, MD Senior Vice President Clinical Information & Publications National Comprehensive Cancer Network	**Troy Sarich, PhD, BSc** Vice President Real World Evidence Janssen Pharmaceuticals
Julie Simmons, CMP (Co-Chair) Manager Marketing and Communications Center for Medical Technology Policy	**Marcus Wilson, PharmD** President HealthCore

STAKEHOLDER MEETING PARTICIPANTS

Regulatory Authorities	
Jonathan Jarow, MD Senior Medical Advisor to the Center Director Center for Drug Evaluation & Research Food & Drug Administration (FDA)	**Mark Levenson, PhD** Deputy Director Division of Biometrics VII Food & Drug Administration (FDA)
Accountable Care Organizations and Integrated Health Systems	
Michele Gilliam, RDLD, CPHQ* Director of Performance Improvement Maine Health ACO	**Lindsay Jubelt, MD*** Senior Medical Director, Mount Sinai Health Partners Assistant Professor, General Internal Medicine Mount Sinai Hospital
Pamela Pelizzari, ScB Healthcare Consultant Milliman	**Megan Priolo, MHS** Chief Operation Officer Greater Baltimore Health Alliance Greater Baltimore Medical Center
Lucy Savitz, PhD Assistant Vice President Delivery System Science Intermountain Healthcare	
Life Sciences	
Charlie Barr, MD, MPH Group Medical Director and Head, Evidence Science and Innovation Genentech	**Scott Flanders, PhD** Director, Health Economics and Clinical Outcomes Research-Oncology Astellas Pharma US
Kenneth (Ken) Iwata, PhD Senior Medical Affairs Director, Oncology Astellas Pharma US	**Bryan Johnstone, PhD** Vice President Sanofi US
Sachin Kamal-Bahl, PhD Vice President/ Head of Global Innovation Center Pfizer, Inc.	**Karen Lencoski, JD** Director Therapeutic Area Government Strategy Astellas Pharma US
James (Jim) Murray, PhD Research Fellow Eli Lilly and Company	**Hemant Phatak, PhD** Senior Director, US Health Economics & Outcomes Research – Oncology EMD Serono
Catherine Piech, MBA Vice President HECOR Janssen Pharmaceuticals	**Jonathan Plumb, MSc, BSc** Head, Global Payer Evidence UCB Pharma

Brande Yaist, MS Senior Director, Center of Expertise, Global Patient Outcomes and Real World Evidence Eli Lilly & Company	**Ming Zhang, PhD** Senior Director, Group Leader RWE Design and Analytics Janssen Pharmaceuticals
Payer Organizations	
Robin Cisneros* National Director, Medical Technology Assessment and Products The Permanente Foundation	**Jo Carol Hiatt, MD, MBA*** Chair, National Product Council Assistant Medical Director, SCPMG Business Administration The Permanente Foundation
Marguerite Koster, MA, MFT* Senior Manager, Evidence-Based Medicine (EBM) Services Kaiser Permanente	**Robert (Bob) McDonough, MD, JD, MPP** Senior Director of Clinical Policy Research & Development Aetna
Lewis (Lew) Sandy, MD Senior Vice President, Clinical Advancement UnitedHealth Group	**John Watkins, PharmD, MPH, BCPS*** Formulary Manager Premera Blue Cross
Professional Societies / Trade Organizations	
Joan McClure, MS Senior Vice President, Clinical Information and Publications National Comprehensive Cancer Network (NCCN)	**Jennifer (Jen) Graff, PharmD** Vice President, Comparative Effectiveness Research National Pharmaceutical Council (NPC)
Health Technology Assessment	
Naomi Aronson, PhD* Executive Director, Clinical Evaluation, Innovation and Policy Blue Cross Blue Shield Association (BCBSA)	**Renee Balliet, PhD** Product Manager, Genetics Products Hayes, Inc.
Stephanie Chang, MD, MPH Director, Evidence-based Practice Center Program Agency for Healthcare Research and Quality (AHRQ)	**Richard Chapman, PhD, MS** Director, Health Economics Institute for Clinical and Economic Review (ICER)
Rabia Kahveci, MD Chair, Antara Numune Health Technology Assessment Unit (ANHTA)	**Jim Reston, PhD, MPH** Associate Director, Health Technology Assessment ECRI Institute
David Wade, MD Chief Medical Officer Hayes, Inc.	

Research and Academia	
Aylin Altan, PhD Vice President, Research Optum Labs	**Diana Brixner, RPh, PhD, FAMCP*** Professor, Department of Pharmacotherapy Executive Director, Pharmacotherapy Outcomes Research Center University of Utah
Nancy Dreyer, PhD Global Chief of Scientific Affairs and Senior Vice President, Real-World & Late Phase Research Quintiles	**Henry (Joe) Henk, PhD** Vice President, Research OptumLabs
Brad Hirsch, MD Senior Medical Director Flatiron	**Jacqueline Milani, MS, CPP** Director University of Maryland School Pharmacy, Pharmaceutical Research Computing
Jodi Segal, MD, MPH Co-Director, Center for Drug Safety and Effectiveness Bloomberg School of Public Health	**Beverly (Bev) Shea*** Senior Methodologist, Ottawa Health Research Institute Clinical Scientist, Bruyère Research Institute Adjunct Professor, Department of Epidemiology and Community Medicine, University of Ottawa
Michael Stoto, PhD* Professor of Health Systems Administration and Population Healthy Georgetown University	**Marcus Wilson, PharmD** President HealthCore
Patient / Consumer Advocates	
Maureen Japha, JD Director, Regulatory Policy FasterCures	**Mark Skinner, JD** President/ CEO Institute for Policy Advancement Ltd.
Other	
John Beilenson President Strategic Communications & Planning (SCP)	**Sam Savitz** PhD Candidate University of North Carolina at Chapel Hill, Department of Health Policy & Management
CMTP Staff	
Jennifer Al Naber Program Manager	**Robert Conley** Research Manager
Janelle King Executive Assistant	**Donna A. Messner** Senior Vice President
Rachael Moloney Research Manager	**Nora Osowski** Research Manager
Sonia Lee Intern	**Sean Tunis** President and CEO

***Participated remotely**

APPENDIX 2: E-SCAN

DOCUMENTS INCLUDED IN REVIEW

Doc #	Document Citation	Audience	Type	Initiative
1	Berger ML, Mamdani M, Atkins D, Johnson ML. Good Research Practices for Comparative Effectiveness Research: Defining, Reporting and Interpreting Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part I. Value in Health 2009;12(8):1044-52.	Researchers	Guidance or Good Research Practices	ISPOR Best Practices Task Force for Comparative Effectiveness Research (Retrospective Database Analysis, Parts I-III)
2	Cox E, Martin BC, Van Staa T, Garbe E, Siebert U, Johnson ML. Good Research Practices for Comparative Effectiveness Research: Approaches to Mitigate Bias and Confounding in the Design of Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report—Part II. Value in Health 2009; 12(8): 1053-61.	Researchers	Guidance or Good Research Practices	
3	Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U.Good Research Practices for Comparative Effectiveness Research: Analytic Methods to Improve Causal Inference from Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part III. Value in Health 2009; 12(8): 1062-73.	Researchers	Guidance or Good Research Practices	
4	Berger ML, Dreyer N, Anderson F, Towse A, Sedrakyan A, Normand S. Prospective Observational Studies to Assess Comparative Effectiveness: The ISPOR Good Research Practices Task Force Report. Value in Health 20120; 15: 217-230.	Researchers, Policy Makers	Guidance or Good Research Practices	ISPOR Best Practices Task Force for Comparative Effectiveness Research (Prospective Observational Studies)
5	Garrison Jr. LP, Neumann PJ, Erickson P, et al. Using real-world data for coverage and payment decisions: The ISPOR real-world data task force report. Value Health 2007;10:326-35.	Payment & coverage decision-makers	Guidance or Good Research Practices	ISPOR Best Practices Task Force for Comparative Effectiveness Research (Real World Data)
6	Berger ML, Martin BC, Husereau D, et al. A Questionnaire to Assess the Relevance and Credibility of Observational Studies to Inform Health Care Decision Making: An ISPOR-AMCP-NPC Good Practice Task Force Report. Value in Health 2014; 17: 143-56.	Decision-makers, evidence assessors	Assessment	Comparative Effectiveness Research (CER) Collaborative Initiative
7	Sterne JAC, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016; 355:i4919.	Decision makers, evidence assessors, researchers	Assessment	Cochrane Collaboration
8	Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomized studies. BMJ 2011; 343: d5928.	Decision makers, evidence assessors, researchers	Assessment	
9	Dreyer NA, Schneeweiss S, McNeil BJ, Berger ML, et al. GRACE Principles: Recognizing high-quality observational studies of comparative effectiveness. Am J Manag Care. 2010;16(6):467-471.	Decision makers, evidence assessors, CER researchers	Assessment	GRACE principles
10	Dreyer NA, Bryant A, Velentgas P. The GRACE Checklist: A validated assessment tool for high quality observational studies of comparative effectiveness research. J Manag Care Spec Pharm 2016;22(10):1107-13.	Decision makers, evidence assessors, CER researchers	Assessment	
11	Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM, eds. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. AHRQ Publication No. 12(13)-EHC099. Rockville, MD: Agency for Healthcare Research and Quality; January 2013. www.effectivehealthcare.ahrq.gov/Methods-OCER.cfm.	Researchers	Guidance or Good Research Practices	Agency for Healthcare Research and Quality
12	PCORI (Patient-Centered Outcomes Research Institute) Methodology Committee. 2013. "The PCORI Methodology Report." pcori.org/research-we-support/research-methodology-standards	Researchers	Guidance or Good Research Practices	Patient Centered Outcomes Research Institute Methodology Committee
13	Real world evidence (RWE) Navigator. Available online at: [https://rwe-navigator.eu/]	Decision makers, industry, researchers	Framework	IMI GetReal
14	The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Guide on Methodological Standards in Pharmacoepidemiology (Revision 4). EMA/95098/2010. Available at http://www.encepp.eu/standards_and_guidances	Researchers	Guidance or Good Research Practices	The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP)
15	The European Networok of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Checklist for Study Protocols (Revision 3). Available at: [http://www.encepp.eu/standards_and_guidances/checkListProtocols.shtml]	Researchers	Guidance or Good Research Practices	
16	Public Policy Committee, International Society of Pharmacoepidemiology. Guidelines for good pharmacoepidemiology practice (GPP). Pharmacoepidemiology and drug safety 2016; 25: 2-10.	Researchers	Guidance or Good Research Practices	ISPE (International Society for Pharmacoepidemiology) Guidelines for Good Pharmacoepidemiology Practices
17	Hall GC, Sauer B, Bourke A, Brown JS, Reynolds MW, Lo Casale R. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemiology and Drug Safety 2011; DOI: 10.1002/pds.2229	Researchers	Guidance or Good Research Practices	
18	U.S. Department of Health and Human Services Food and Drug Administration. Use of real world evidence to support regulatory decision-making for medical devices. Draft guidance for industry and Food and Drug Administration staff. July 27, 2016. Available online at [http://www.fda.gov/ucm/groups/fdagov-public/@fdagov-meddev-gen/documents/document/ucm513027.pdf]	Industry, researchers	Guidance for Industry	U.S. Food and Drug Administration
19*	U.S. Department of Health and Human Services Food and Drug Administration. Good pharmacovigilance practices and pharmacoepidemiologic assessment. Guidance for Industry. March 22, 2005. Available online at: [http://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM126834.pdf]	Industry, researchers	Guidance for Industry	
20†	Guyatt GH, Oxman AD, Vist GE, et al. GRADE guidelines: 4. Rating the quality of evidence-- study limitations (risk of bias). J Clin Epidemiol 2011;64:407-415.	Decision makers, evidence Assessors	Evidence Hierarchy & Assessment	GRADE Working Group
21	Kahn MG, Brown JS, Chun AT, Davidson BN. Transparent reporting of data quality in distributed data networks. eGEMs 2015; 3(1): Art 7.	Evidence Assessors, researchers	Framework	PCOR Data Quality and Transparency Standards Project
22‡	Benchimol EI, Smeeth L, Guttmann A, et al. The Reporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLOS Medicine. 2015; 12(10):e1001885.	Journal Editors, Researchers, Peer Reviewers	Reporting Standards	RECORD statement extention to STROBE Statement)

†Though there are multiple GRADE publications, here we include only the article on assessing limitations of individual studies.

COMPARISON OF RELATED INITIATIVES

RWE Decoder	Related Documents (Doc #)
Module 1. Articulating Research Question (Or: general content regarding elements of a good research question)	**1,4,11,12,13,15**
Decision-making context	4,5,11,12,13,15
Patient-Important Outcomes	11,12,13
PICO	4,11,12,13
(TS)	11,12
Module 2A. Assessing Relevance (Or: general content regarding the relevance, generalizability, or applicability of results)	**6,11,12,13,16,18,22**
Population	2, 4,6,9,11,12,13,14,15,16,17,18,19,22
Intervention	4,6,9,10,11,12,13,14,16,18,19,22
Comparator	4,6,9,10,11,12,13,14,16,18,22
Outcomes	4,6,9,10,11,12,13,14,15,16,18,19,22
Timing	2,6,11,12,18
Setting	6,11,12,13,22
Module 2B. Assessing Rigor	
Quality of the Research Question (Or: general content regarding the statement or quality of research question, or appropriateness of study design)	1,4,6,11,12,13,14,15,16,19,22
PICO(TS) Stated	4,6,9,11,12
Approrpiateness of Study Design	1, 2,3,4,6,9,10,11,12,13,14,15,16,17,18,19,22
Risk of Bias (Or: general content regarding the risk of bias, credibility of findings, or other limitations of study results)	1, 2, 3,4,6,7,8,9,10,11,12,14,15,16,17,18,20,22
Confounding bias	1, 2, 3,4,5,6,7,9,10,11,12,14,15,16,17,21,20,22
Selection bias	3,4,6,7,8,10,11,12,14,15,16,20,22
Bias from classification of intervention	1,2,6,7,9,11,12,14,15,16,17,20,21,22
Bias from deviation from intended intervention	4,7,11,13,14,16,20
Bias from missing data	3,6,7,9,11,12,14,15,16,21,20,22
Bias from measurement outcomes	1,2,6,7,9,10,11,12,14,15,16,17,20,21,22
Bias in results reporting	6,7,8,12,15,20
Performance Bias	8,20
Detection Bias	8,9,10,14,20
Attrition Bias	8,9,12,13,14,20
Other (e.g. contamination, recruitment bias)	8,13,20
Precision (Or: general content regarding precision, including sample size, measures of variance, confidence intervals, etc.)	4,6,11,12,14,15,16,19,22
Confidence Interval	11,22
Data Quality (Or: general content regarding data quality or integrity, data source, choice of dataset, data stewardship, etc.)	5,6,10,11,12,14,15,16,17,18,21,22
Data source & intention	1,4,5,6,9,10,11,12,14,15,16,17,18,19,21,22
Completeness (absence of missing data)	5,9,11,12,14,15,16,17,18,21,22
Fidelity (e.g. a female is coded as a female)	5,11,12,14,15,16,17,18,21,22
Plausibility (e.g. are the data believable)	18,21,22
Cohort construction & linkage	5,11,12,14,15,17,18,21,22
Module 2C. Effect Size, Direction (Or: general content rearding magnitude, direction, or meaningfulness of effect of primary outcome)	**1,6,9,11**

Please note: This scan is not a validated quantitative comparative analysis of content across initiatives. Rather, it provides a holistic overview of existing initiatives as resources for users wishing to explore related or more in-depth material pertinent to the assessment of rigor or relevance for studies. Some of the assignments made in this table of related documents entailed subjective judgement; for example, if specific designs or analytic techniques were relevant to a particular source of bias but were discussed without specific mention of that bias, they were listed in this table as relevant to that source of bias.

APPENDIX 3: USE CASE

INSTRUCTIONS

First, read the use case overview, then read the hypothetical studies described within the use case. An "inner dialogue," from the perspective of a decision maker using the GPC RWE Decoder, follows. Instructions for the RWE Decoder tell the user, for each study, to consider and rate on a scale of 1 – minimally relevant to 4 – very relevant, the **Relevance** of the study, and on a scale of 1 – minimally rigorous to 4 vey rigorous, the methodological **Rigor** of the study. Furthermore, to enter "0" only if the study fails to provide a piece of information that is necessary, thus making the evidence from the study unusable for the current decision-making situation.

MODULE 1: ARTICULATING THE QUESTION	
Q1. What is the nature of your decision? Q2. What do you want to know? Q3. What do patients want to know (e.g. which outcomes are most important to patients?) Q4. What is your research question, rephrased following PICOTS format?	
MODULE 2A: ASSESSING THE RELEVANCE OF EACH RWE STUDY	
PICOTS: Population – **I**ntervention – **C**omparator – **O**utcomes (Primary, Secondary) – **T**iming – **S**etting	
MODULE 2B: ASSESSING THE RIGOR OF EACH RWE STUDY	
QUALITY OF RESEARCH QUESTION	▪ PICOTS Stated ▪ Appropriateness of study design ▪ Scientific argument
POTENTIAL FOR BIAS	▪ Confounding bias ▪ Selection bias ▪ Bias from classification of intervention ▪ Bias from deviation from intended intervention ▪ Bias from missing data ▪ Bias from measurement outcomes ▪ Bias in results reporting Additional sources (randomized studies): ▪ Performance Bias ▪ Detection Bias ▪ Attrition Bias ▪ Other (e.g. contamination, recruitment bias)
PRECISION	▪ Confidence interval
DATA INTEGRITY	▪ Data source & intention ▪ Completeness (absence of missing data) ▪ Fidelity (e.g. a female is coded as a female) ▪ Plausibility (e.g. are the data believable) ▪ Cohort construction & linkage
MODULE 2C: MAGNITUDE AND DIRECTION OF EFFECT*	
Effect Size (primary outcome)	Direction (-, +, or no difference)
MODULE 3: RWE FRAMEWORK VISUAL SUMMARY	
Data point (one per study): (Relevance [X], Rigor [Y], Effect Size [bubble size], Direction [bubble color])	

Use Case Overview

Payers have noticed an increase in off-label prescribing of Drug X, which is only approved for the treatment of MS, among ALS patients in their covered populations. Drug Y is the current standard of care for long-term treatment to slow ALS disease progression, often measured every 6 to 12 months. Drug Y is currently covered by health plans and indicated in clinical practice guidelines for the treatment of ALS.

What is the nature of your decision?

You, a payer and framework user, would like to assess the effectiveness of a newly popular off-label therapy for a degenerative disease, ALS, in health systems in which ALS patients receive routine care, in order to help you decide whether a) restrictions should be put in place to limit, or b) a new policy or program could allow, within certain parameters, the off-label use of this MS treatment for ALS patients in your covered population.

What do you (decision maker) want to know?

You would like to know if Drug X is effective at reducing ALS disease progression, whether as a potential alternate treatment option to Drug Y, or perhaps in combination. Answering this question will then help you as payer decide whether Drug X should be incorporated into new clinical programs, coverage with evidence development programs, or updated clinical practice guidelines for ALS.[14]

What do patients want to know?

Patients want to know if Drug X will impact their physical function, e.g. their independence and ability to perform daily tasks, as well as their overall quality of life.

What is your research question, rephrased and following a PICOTS format?

DOMAIN	DOMAIN DETAILS
Population	Adult ALS patients within our covered population
Intervention	Drug X (off label treatment)
Comparison	Drug Y (current standard of care)
Primary Outcome(s)	Long term disease progression (difference over time in measureable clinical outcomes: limb function, muscle strength)
Secondary Outcome	Patient-reported function, and/or patient-reported quality of life
Secondary Outcome	Surrogate outcomes: motor neuron loss
Timing	6 to 12 months
Setting	Network health systems in which ALS patients receive routine care

Hypothetical Studies available to inform decision-making

Disclaimer: The intent of this use case is not to advocate a one-size fits all interpretation or to prescribe standard thresholds for rigor or relevance. We acknowledge the subjective, context-dependent nature of health care decision-making, and utilize this use case to demonstrate how one might consider their own decision making context, the impact of their decision, and their own organizational requirements for evidence, together with the output from the RWE Decoder.

[14] The primary interest is currently the clinical effectiveness of the drug. Cost may become important in subsequent stages of decision-making, e.g. whether to give favorable formulary status to Drug X, pending CED study results.

Overall ratings of rigor and relevance are informed judgments based on a stepwise assessment of each domain. However, overall scores are NOT required to be average scores of each subdomain.

STUDY 1. (N=160) A manufacturer-sponsored Phase 2 single-blind (patient) RCT measuring safety and adverse effects of Drug X in adults with neurodegenerative disease. A subgroup of forty (n=40) treatment naïve, adult ALS patients with symptom onset within the past year and no other diagnosed comorbid conditions, and who were either treatment naïve or on a stable long term regimen of Drug Y, randomly (1:1) received Drug X or placebo (blinded, and were closely monitored by trained study personnel over the course of trial follow-up). The sample size was too small to generate any precise estimate of effectiveness of Drug X in ALS patients, and did not compare safety outcomes among ALS patients taking Drug X compared to those taking Drug Y.

Findings: Findings in the subgroup were consistent with the broader study sample and original product label. After 6 months, outcomes in the subgroup did not differ from the total study population (i.e., safety outcomes for Drug X in the subgroup were similar to major safety outcomes for Drug X in the total population). Side effects were similar in the subgroup compared to the total study population.

Overall Relevance: 1. Population is somewhat relevant – adult ALS patients, however they do not represent the broad range of people with ALS in the broader covered population, e.g. who may have multiple comorbid conditions, broader socio-demographic and other clinical characteristics. Intervention is more relevant; Drug X is used in the way that would be prescribed to ALS patients in real health delivery settings. However, the use of a placebo comparator instead of Drug Y is not relevant. While the Outcomes provided are helpful information, the study only provides safety information and no information on efficacy or effectiveness regarding disease progression or outcomes of interest to patients. Timing is appropriate and relevant, since a period of several months is sufficient, based on existing clinical knowledge, to observe measurable difference in ALS treatment research. Setting isn't fully described, however the sentence stating patients were "closely monitored by trained study personnel" during follow-up implies a departure from usual care, thus less relevant. The overall Relevance rating of 1 is mainly because the study failed to measure the outcomes you're particularly interested in, in a setting that would reflect the care settings where most of our covered population of ALS patients receive care.

Overall Rigor: 2. Quality of Research Question: 1-2. The study is fairly rigorous, though it is not designed to answer a question of efficacy. Although the objective of the overall Phase 2 study may have been articulated prospectively, and the study design appropriate to observe safety outcomes, it is unclear whether the question for this subgroup was articulated in advance, or whether the analysis was planned or conducted post hoc. You assume there was some planning, since the description implies that within the subgroup, the 40 patients were evenly randomized. The primary aim of the study was not to look at ALS patients, but a broader group of MS patients. Potential for Bias: 2. Though there is randomization (here you assume the investigators provide a table to demonstrate successful randomization, the study is limited by its small sample size. Furthermore, though patients were blinded to treatment choice, thus reducing the potential for bias in the performance of those on Drug X, the outcome assessors were not blinded and thus there is some concern regarding potential bias in the detection of outcomes). Selection bias is also a concern, you would prefer to see information on study dropouts (attrition) and reasons for dropout. The extent of missing data is also not disclosed, and this may be particularly important given the small sample size. Precision: 1. There is no precise estimate for the subpopulation. Data Integrity: 3. You are more confident in the Data Integrity for several reasons: You assume the trial was held to strict standards because of regulatory requirements (e.g. there was a data safety monitoring board and the

entire dataset will be made available to regulators), because the data were collected primarily for research (Intention) and so variables and definitions are clear and do not conflict. However, since there is no discussion of missingness, you still have some reservations regarding completeness.

STUDY 2. (N=1000) The research center of one large private health system conducted a retrospective analysis of combined data from three of their clinics in order to compare disease progression rates among ALS patients before and after they started Drug X. This health system has published numerous studies using their own proprietary data, and have internal research teams familiar with their databases. For this retrospective study, they chose three clinics using the same data definitions, and which previously contributed to a prior research study of ALS patients. In the combined database, they matched patients based on propensity scores calculated using patient age, sex, region of onset, and co-treatment (e.g. treatment naïve, short term Drug Y, long term Drug Y). In order to study whether Drug X slows disease progression, compared to Drug Y, researchers recorded a) muscle strength, and 2) limb function at 6 months' pre-intervention, at baseline (start of Drug X), and 6 months' post-intervention.

Findings: Findings from a paired analysis demonstrated a moderate but clinically meaningful, statistically significant difference in function post-intervention compared to pre-intervention, among those ALS patients who took Drug X. In other words, results indicate that Drug X may be associated with a meaningful slow in disease progression among ALS patients.

Overall Relevance: 4*.* Population, Intervention, and Primary Outcomes are highly relevant, Comparator, Timing and Setting are reasonably relevant. The Population includes a large diverse group of similar patients to the covered population, in similar care settings. The Intervention is the treatment of interest, Drug X, and considers both prior and co-treatment, which is likely how it is being adopted off-label. The Comparator is relevant, being Drug Y, the current standard. The Primary Outcome is the same as stated in the PICOTS table, though neither secondary outcome from the PICOTS table is measured. Timing is relevant (6 though not 12 months). Setting is relevant – the study is conducted in a competing health system however the setting is otherwise one in which ALS patients receive their routine care. Greater discussion among colleagues might help us understand key differences between our care settings and the system in which the study took place.

Overall Rigor: 4. Quality of Research Question: 4. The study and analysis were clearly designed to answer the question of comparing ALS progression among ALS patients before and after they started Drug X. Potential for Bias: 3. Here your first instinct is to give this study a 1 because of the retrospective nature, however, you reconsidered and gave it a 3, considering the investigators took numerous steps to minimize the potential for bias. Precision: 4. This large study generated statistically significant results for a meaningful estimate. Data Integrity: 3. Although the study was conducted using electronic health data, the system has lots of research experience and familiarity with their own data, with a dedicated internal research team. They are therefore likely to understand the data's limitations.

STUDY 3. (N=200) One community clinic in the same system is the site of an investigator-initiated prospective observational study to measure and compare disease progression in ALS patients on Drug X, Drug Y, and a combination. Long term outcomes data are not yet available, however, among the 200 adult ALS patients enrolled in this prospective study, after 3 months of follow-up the clinic has measured biomarkers of upper and lower motor neuron loss and comparing among three treatment arms: Drug X, Drug Y, and Combination. Investigators mention controlling for bias by including several known clinical and demographic confounders in their multi-variate analysis. This clinic does not have a track record of research publications, however, in the published report the investigator will respond to requests for

study data from potential collaborators. There may be future opportunities to replicate the findings from this study.

Findings: Researchers reported a small confidence interval and statistically significant difference close enough to zero to suggest no meaningful difference in motor neuron loss among Drug X patients compared to Drug Y patients. However, less motor neuron loss was observed in those patients taking both Drug X and Drug Y in combination. The investigator cannot yet draw any conclusions as to how the surrogate findings might translate to patient functionality and progression of symptoms.

Overall Relevance: 2. The Population is relevant, adult ALS patients, the Intervention and Comparator are also relevant to practice, being Drug X, a combination of Drug X and Y, and Drug Y. However, the Outcome is a surrogate with no information on its validity or the predictive value of a change; there are no clinical measures of functional improvement. There is no mention of secondary or patient-reported outcomes. The Timing is currently not as relevant, being only 3 months. While the investigators indicate there will be long term outcomes collected, at this time of decision-making, short term surrogate outcomes are all that is available. Finally, the Setting is relevant, being a community clinic that delivers care to part of the covered population.[15]

Overall Rigor: 2. Overall this is a low-rigor study – or more specifically, a low rigor interim analysis. Quality of the Research Question: 2. The objective of the study is to compare clinical measures of ALS disease progression over time across the three treatment arms. However, currently the primary analysis is not available, only an analysis of surrogate outcomes that have not been validated (though there is some supporting evidence). Potential for Bias: 1. While there is some effort to adjust for confounding in a multi-variate analysis, one might assume that the known confounders relate to clinical outcomes of disease progression, but these have not been explored in relation to surrogate outcomes. Non-randomized design with no information on patient selection for use of new drug or drug combination. For instance, those remaining on Drug X may have been more stable (more slowly progressive) than those placed on Drug Y or the Combination (more rapidly progressive). There is no information about deviations from the interventions or about co-interventions. There is no mention of measurement errors or blinding of outcomes measurement, so observers will have known what participants were receiving. There is no information on missing data or selective outcomes reporting (e.g. do the investigators only report the surrogate measures that changed). Precision: 2. Estimates were statistically significant and precision of estimated changes seems reasonable but the meaningfulness, or clinical relevance, of these differences in surrogate outcomes remain unclear. Data Integrity: 2. The clinic does not have a research history or track record of published studies, and there is little description of procedures for data cleaning, missingness, etc. On the other hand, Interpreting the meaning or intent of laboratory data is more straightforward than clinical data. The investigators do indicate that they are willing to share their dataset with future collaborators, so there may be an opportunity (eventually) for replicability studies.

STUDY 4. (N=600) Another clinic, which is actively invested in patient engagement and has a highly utilized patient portal, surveyed patients who stayed on Drug Y and patients who switched from Drug Y to Drug X. Among patients who had been on Drug Y for one and two years, respectively, patients who switched to Drug X reported meaningful slowing of disease progression, (though statistically nonsignificant, with moderately wide confidence intervals), measured as patient-reported function and quality of life compared to those who stayed on Drug X. The study dataset came entirely from patient

[15] Assumptions: It is assumed that in this non-research setting diagnosis is accurate and diagnostic misclassification is not occurring.

surveys collected every six months, though with varying response rates. Limb function and muscle strength was not extracted from health records or used to validate findings among a subset of patients.

Overall Relevance: 2. The population from which the study participants were selected is similar to the ALS patients in our network, suggesting relevance. However, because patients are self-selected, it is possible there are important subgroups of patients that are not involved in the study. Both the intervention (Drug X) and comparator (Drug Y) are relevant. The outcomes are of interest, however, the study only collects self-reported function and quality of life, so clinical measures of disease progression (limb function and muscle strength) are unavailable. Timing is relevant, if patients stay active in survey participation every six months. Setting*:* the use of a patient portal could be considered a representative setting of many health systems, and leveraging a highly used patient portal has the potential to provide highly relevant information directly from patients regarding their firsthand experiences with their diseases and care.

Overall Rigor: 1. The study has very weak rigor. The patients are self-selected and the outcomes are not validated. There is confounding by the treatment. No attempt is made for adjustment for confounding. There is no blinding. On the positive side, the study is surprisingly large. The use of a patient portal opens the study to greater risk of selection bias from self-selection factors and reliance on self-report. High likelihood of selection bias: those who appeared to respond to the new drug being more likely to report their experience. It is unclear what “meaningful improvements” mean in a context where treatment can only be expected to slow the rate of progression. There is no information on treatment switches or co-interventions, missing data or selective reporting of outcomes and analyses. All of these flaws are quite likely in this setting.