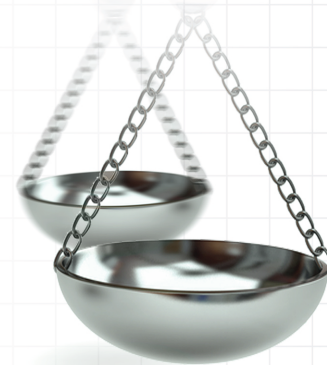




# A review of stakeholder recommendations for defining fit-for-purpose real-world evidence algorithms

Journal of **Comparative Effectiveness Research**

Julie Beyrer<sup>\*1</sup> , Hamed Abedtash<sup>1</sup> , Kenneth Hornbuckle<sup>2</sup>  & James F Murray<sup>1</sup>

<sup>1</sup>Department of Value, Evidence, and Outcomes, Eli Lilly & Company, Indianapolis, IN 46285, USA

<sup>2</sup>Department of Global Patient Safety, Eli Lilly & Company, Indianapolis, IN 46285, USA

\*Author for correspondence: Tel.: +1 317 651 8236; [beyrerj@lilly.com](mailto:beyrerj@lilly.com)

**Aim:** The credibility and value of real-world evidence (RWE) are either supported or undermined by the algorithms (i.e., operational definitions) used. **Methods:** We conducted a targeted evidence review of key RWE decision makers' published recommendations on RWE algorithms through April 2021. Stakeholders were regulatory bodies, other governmental agencies and payer organizations. **Results:** Our review identified recommended criteria: relevance, validity, reliability, responsiveness, transparency and replicability, safety, feasibility and quality process. Stakeholders routinely recommended accuracy measures, subgroups evaluation and specific considerations for assessing exposures and covariates and the underlying real-world data (RWD) quality. **Conclusion:** The importance of stakeholder guidance on fit-for-purpose RWE algorithms is growing. We highlight gaps that future guidance and stakeholder recommendations could address.

First draft submitted: 19 January 2022; Accepted for publication: 25 February 2022; Published online: 17 March 2022

**Keywords:** decision making • fit for purpose • operational definitions • payers • real-world data • real-world evidence • real-world evidence algorithms • regulators • stakeholders

As real-world evidence (RWE) is increasingly used to inform decision making, the importance of stakeholder guidance on the quality of RWE is growing. While, to date, much of the focus has been on RWE data quality, transparency and replicability [1–7], less attention has been paid to the importance of RWE algorithms. This paper provides a deeper dive into RWE algorithms, a fundamental part of creating and interpreting RWE. An algorithm is a set of rules to be followed in calculations or problem-solving operations. RWE algorithms provide the rules for defining or deriving variables for observational research and are comprised of real-world data (RWD), coding schemes, metadata and logic.

The criteria for selecting RWE algorithms are important because the credibility and value of RWE are either supported or undermined by the algorithms used. An algorithm with undesirable properties may have a detrimental effect on the study (e.g., leading to uninterpretable results or a wrong decision based on faulty evidence). Studies addressing the same research question using the same data and analysis methods may arrive at different results and conclusions simply due to differences in the algorithms used [8,9]. Differences in algorithms used or variability in the practice of selecting and reporting the algorithms, may introduce or mask bias and can also make it impossible to validate or replicate results [9,10].

Transparent algorithms are necessary but not sufficient. An evaluation of the relevance and quality of algorithms to support the intended use is also necessary. However, while many publications describe the algorithms used, it is often unclear what, if any, appraisal of the algorithm's operating characteristics and suitability for the study (or fit for purpose) was performed. Evaluation of the algorithm's fit-for-purpose properties is essential for producing high quality and relevant RWE.

The aim of our review was to synthesize and describe the existing stakeholder-recommended criteria and operating characteristics to judge the suitability of an algorithm and identify gaps still to be addressed. Our basic premise is the credibility and value of RWE are directly affected by the fit-for-purpose properties of the algorithms used.

**Box 1. Definitions**

These are the working definitions of the four algorithm types addressed:

**Cohorts**

A cohort is a group of people who share a defining characteristic (e.g., period of birth) or a group followed or traced over time [15]. It is often defined by people who experienced a common diagnosis, procedure or treatment in a selected time period. The cohort is defined by inclusion and exclusion criteria and may require confirming the presence or absence of a key factor such as a treatment or a diagnosis. The cohort definition also includes the duration of time between an individual's entrance and exit of the study (i.e., entry date and end of the observation period) [60].

**Exposures**

An exposure is a known or potential variable of interest assumed to affect an outcome of interest. In a clinical outcome model, exposure is the primary independent variable of interest. Often this variable has a known or hypothesized causal relationship with the outcome of interest. An exposure is usually an exogenous agent (e.g., medication or procedures) but it may also be environmental factors (e.g., socioeconomic factors), endogenous factors (e.g., individual demographics) or other factors (e.g., the intensity, frequency or temporality of medication exposure) [12,15,61,62].

**Covariates**

A covariate is a known or potentially influential variable related to the outcome under study. In a clinical outcome model, covariates are other independent variables including effective modifiers or confounders that may be assumed to affect (influence) an outcome and/or exposure. A covariate may be exogenous, endogenous or other factors.

**Outcomes**

An outcome defines the result (e.g., the health status) of exposure (e.g., a treatment) and covariates. An outcome can be a benefit (i.e., increase in a desired outcome) or a harm (e.g., an adverse event) [15,63,64]. There are diverse types of outcomes based on their characteristics:

- Clinical outcome. Examples include disease occurrence, death, symptoms and/or health status.
- Humanistic outcome. Examples include the quality of life and functional status.
- Economic outcome. Examples include resource utilization, medical cost, cost of treatment, cost of reduced morbidity and cost of years of life saved.

Outcomes can be either a surrogate end point or a final end point. In some cases, related outcomes are integrated into a single 'composite outcome'.

**Methods**

A targeted evidence review was conducted to identify the stakeholder-recommended criteria for judging the suitability of RWE algorithms. We identified regulatory and payer organizations representing key decision makers: US FDA, European Medicines Agency (EMA), International Conference on Harmonisation (ICH), National Quality Forum (NQF), National Committee for Quality Assurance, Agency for Healthcare Research and Quality (AHRQ), and Patient-Centered Outcomes Research Institute. Stakeholder web sites were searched through April 2021. Documents providing recommendations or criteria for evaluating RWE algorithms were included. Given the paucity of recommendations specific to RWE algorithms, we included documents describing recommendations for evaluating RWD or clinical outcomes assessments (COA) measurement; we assumed many of the same measurement recommendations that apply to RWD and other drug development tools like COA would apply to RWE algorithms. The source documents identified and reviewed are shown in [Appendix 1](#). Specifically, we address four types of algorithms: cohorts (including subgroups), exposures, covariates and outcomes. Our working definitions are given in [Box 1](#). There are situations when an algorithm may be used for multiple types; for example, a cohort algorithm may be used to identify an exposure, a covariate, and/or an outcome (e.g., algorithm identifying cardiovascular disease in a diabetic patient). For each of the four algorithm types (cohorts, exposures, covariates and outcomes) we reviewed the source documentation to extract and collate the criteria.

**Results****General criteria for real-world algorithms**

[Table 1](#) describes the general criteria that apply to all four algorithm types identified from the review. The criteria include relevance, validity, reliability, responsiveness, transparency and replicability, safety, feasibility and quality process. We noted that a definition of the concept is not always provided in the stakeholders' publications even though the concept is included in their recommendations. Therefore, we have provided a working definition derived from a synthesis of the definitions and criteria we found in our review ([Table 1](#)).

**Table 1. The General Criteria for Assessing the Suitability of Real-world Evidence.**

| Criteria                                  | Definitions                                                                                                                                                                                                                                                                                                                                                                        | Ref.                   |
|-------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------|
| Relevance                                 | Degree to which the algorithm provides important and necessary information to the conclusions drawn from the algorithm in either a decision or its application. For example, are the needed data for defining the concept present, and do the codes adequately represent the underlying medical concepts they are intended to represent?                                           | [11,13,14,16–23,28]    |
| – Clinical relevance                      | Consistency with current clinical guidelines or clinical expert judgment. There should be evidence documenting a link between clinical processes and outcomes. For example, establishing a clinically appropriate outcome definition is critical for selecting a fit-for-purpose outcomes algorithm                                                                                | [11,13,16,19–22]       |
| – Generalizability and representativeness | Ability of the algorithm to produce findings that address the question posed for a specific decision or application, including the target population and target data source                                                                                                                                                                                                        | [11–14,17–23,28]       |
| – Potential for improvement               | Ability for healthcare systems or patients to improve their performance or outcomes. Examples of improvement include demonstration of quality problems and improvement, data variation across systems, performance across organizations and disparities in care across population groups                                                                                           | [11,19,20]             |
| Validity                                  | Evidence that algorithms measure what they are intended to measure. The relevant measures of validity depend on the specific characteristics of the algorithm (e.g., content validity, construct validity and criterion validity)                                                                                                                                                  | [11–23,28]             |
| Reliability                               | Extent to which the results produced by the algorithm and/or the algorithm's performance are consistent and reproducible over time, in different datasets or in multiple uses/scenarios. The relevant measures of reliability depend on specific characteristics of the algorithm (e.g., test–retest or intra-rater reliability, inter-rater reliability and internal consistency) | [11–14,16,18–22,28]    |
| Responsiveness                            | The ability to detect change and identify differences in results over time in individuals or groups who have changed with respect to the algorithm concept                                                                                                                                                                                                                         | [11,12,14,16,21,22]    |
| Transparency and replicability            | Degree to which the algorithm is communicated clearly so others can understand and replicate it. Transparency has been defined as openness and honesty about study variables, end points and other aspects of real-world data study development and conduct. This includes the concept of clear specifications of the algorithm                                                    | [11–14,18,20–23,28,58] |
| Safety considerations                     | Ensuring any measure or algorithm has no potential for individual harm during its collection or its application to a decision                                                                                                                                                                                                                                                      | [16]                   |
| Feasibility                               | Specifically, the constructs below define the concept of feasibility                                                                                                                                                                                                                                                                                                               | [11–14,19–22,28]       |
| – Logistical feasibility                  | Are the required data available? Are the data reasonably accessible?                                                                                                                                                                                                                                                                                                               | [11,12,14,17,19–22,28] |
| – Reasonable cost                         | Does the measure impose an undue burden on scarce resources (i.e., time or money)? A measure should not impose an inappropriate burden on healthcare systems (e.g., expensive primary data collection) or patients (e.g., time)                                                                                                                                                    | [11–14,17,19–22]       |
| – Privacy and confidentiality             | Does data collection meet accepted standards of privacy and confidentiality?                                                                                                                                                                                                                                                                                                       | [11–14,17,19–22,28]    |
| Quality process                           | Applying processes or procedures designed to support quality assurance and confidence that the algorithm is suitable. For example, early consultation with decision makers, pre-specification, expert involvement, sensitivity analyses and compliance with privacy, ethical, regulatory requirements                                                                              | [11–14,16–23,28]       |

### Specific criteria for each type of real-world algorithm

The following criteria identified from the literature review are more directly applicable to each of the four algorithm types.

#### *Specific criteria for cohort algorithms*

The key consideration for cohorts is how well the population identified by the algorithm reflects the actual population of interest. A relevant case or cohort definition (e.g., incident vs prevalent cases) is a prerequisite for obtaining meaningful measurement of accuracy. Some accuracy measures used to help understand this are: sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Note that although these measures are foundational for describing cohorts, the stakeholders' recommendations include these accuracy measures for all four algorithm types [11–14]:

- Sensitivity measures the ability of a test or algorithm to correctly identify individuals with the target condition. Sensitivity is the probability a person with the target condition (case) in the population will be identified as such by the algorithm [15];
- Specificity measures the ability to correctly identify individuals who do not have the target condition. Specificity is the probability a person without the condition (non-case) will be correctly identified as such by the algorithm [15];
- Positive predictive value (PPV), also known as precision, measures the proportion of the cohort (after applying the algorithm) with the target condition. PPV is the probability a person identified as a case by the algorithm is truly a case (e.g., does have the target condition) [15];

| Table 2. Calculation of sensitivity, specificity and positive and negative predictive values. |                                         |                                        |               |
|-----------------------------------------------------------------------------------------------|-----------------------------------------|----------------------------------------|---------------|
| Test                                                                                          | Actual positive (e.g., disease present) | Actual negative (e.g., disease absent) |               |
| Test positive                                                                                 | True positive (a)                       | False positive (b)                     | PPV = a/a + b |
| Test negative                                                                                 | False negative (c)                      | True negative (d)                      | NPV = d/c + d |
|                                                                                               | Sensitivity = a/a + c                   | Specificity = d/b + d                  |               |
| NPV: Negative predictive value; PPV: Positive predictive value.                               |                                         |                                        |               |

- Negative predictive value (NPV) measures the proportion of the patients not identified in the cohort who do not have the target condition. NPV is the probability a person not identified as a case by the algorithms is truly a non-case (e.g., does not have the target condition) [15].

The accuracy measures are represented in [Table 2](#). Finally, an assessment of algorithm performance by subgroups, or patient characteristics likely to affect algorithm performance, is generally recommended [11–14,16–21]. These include but are not limited to age, gender, race, disease severity and comorbidities.

*Specific criteria for exposure & covariate algorithms*

While exposures and covariates are fundamentally different variables, they share criteria for assessing their suitability, with exposures having additional considerations. The AHRQ user guide for developing an observational comparative effectiveness research protocol [12] provides extensive recommendations or considerations for exposures and covariates measures, although many of these concepts are covered in other stakeholder recommendations as well [11,13,14,17–23]. The guide recommendations are largely focused on medication exposures, but these considerations could also be applied to other types of exposures. We have summarized the AHRQ considerations as questions and categorized them as shown in [Box 2](#).

*Specific criteria for outcome algorithms*

Of the four algorithm types, outcomes are the predominant focus of the criteria we identified. Among stakeholders, regulators are generally viewed as setting the highest bar for measuring outcomes. For example, in the USA, outcome assessment methods to support FDA drug approvals must be well defined and reliable by law [24]. The regulatory guidance reviewed on RWE algorithms for outcome measures was limited to a single use case: pharmacoepidemiologic safety studies in electronic healthcare datasets [13]. We expect regulatory guidance will continue to evolve on the use of RWD and RWE, including outcome algorithms (i.e., acceptable real-world end points).

At the time of this review, two guidance documents from the FDA provide important insights: Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data [13] and Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims [11]. However, there is no comprehensive or single source of regulatory guidance on RWE algorithms at the time of our review. What we present is a compilation of criteria drawn from several existing sources (See [Appendix 3](#)).

The FDA pharmacoepidemiologic safety studies in electronic healthcare data guidance [13] describes the evidentiary support needed for outcome measures in safety surveillance in pharmacoepidemiologic safety studies (e.g., accuracy/internal validity as well as generalizability/external validity). Beyond safety outcomes, there is limited regulatory guidance on RWE algorithms.

In the absence of more specific regulatory recommendations, we focus on other regulatory precedents on outcomes and endpoints. Outcome assessment includes these three types: biomarkers, all-cause mortality and COAs [25]. The FDA guidance on COAs is particularly relevant for RWE algorithms, though thoughtful adaptation to various types of outcomes is required. We focus on COAs as described in the 21st Century Cures Act (Federal Register 2016) [26]. COAs directly measure a patient’s symptoms, overall mental state, or the effects of a disease or condition on how the patient functions [26]. [Figure 1](#) illustrates and defines the types of COAs: clinician-reported, observer-reported, patient-reported and performance outcome [29].

COA outcomes in RWE analyses submitted to the FDA and other regulatory agencies need to document measurement capability following the guidelines detailed in the FDA’s 2009: guidance for Industry Patient-Reported

**Box 2. Considerations for identifying fit-for-purpose algorithms for exposures and covariates.****Properties of medications**

- What is the relationship between the dosage and dose-response?
- What are the pharmacokinetic and pharmacodynamic properties of the medication (e.g., half-life of the medication)?
- What is the recommended administration (e.g., oral, infused, etc.) of the medication?
- What were the actual versus recommended administration and frequency?

**Relationship with outcome**

- Is the outcome dependent on single or cumulative exposures?
- What intensity of exposure (e.g., dose, frequency and duration) is needed to observe the outcome?
- Does the outcome vary with different intensity of exposure?
- Is there a relationship (and what is the relationship) between the mode or context of delivery and the outcome?
- Is the relationship between exposure and outcome linear or some other type of relationship?
- Are there other indications or potential reasons for use of the medication other than the one of interest?
- What is the potential for bias between exposures/covariates and outcomes? For example, consider whether important covariates are present in the data source and the effect on conclusions regarding the exposure–outcome relationship.

**Relevant time window of measurement**

- What time window of measurement is relevant (calendar time and etiologic relevant time window)?
- Is a complete picture of the exposure available in the dataset?
- How immediate is the outcome?
- Is there an indication and latent period (e.g., a time period during which additional exposure will have no effect on outcome)?
- Are there any changes in exposure status? And if so, are 'spillover' effects (e.g., effects that persist after medication is discontinued) possible?
- How rapidly is the exposure's effect lost?
- Are the exposure data available with the relevant frequency during the relevant time window of measurement?

**Measurement error**

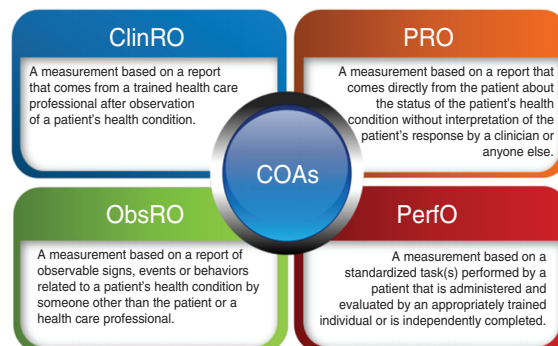
- Are there any known differences in measurement methods between reporters (e.g., data abstractors, labs, etc.)?
- Are there any changes in the measurement method(s) over time?
- What were the quality control procedures for the measurement method (e.g., inter-observer or intra-observer reliability)?
- Are there any other aspects of the data (e.g., data provenance, health plan coverage of drug, Healthcare Common Procedure Coding System code effective date delays, etc.) that might lead to gaps or missing exposure or covariate data?

**Measurement scale**

- What is the method of measurement (e.g., script fills, patient report, recall, etc.)?
- What is the quantitative representation of the exposure (e.g., continuous, categorical, dichotomous)?
- What was the frequency of measurement?

Above information taken from [12].

**Figure 1. Overview of clinical outcome assessment types.** This illustrates and defines the types of COAs: clinician-reported, observer-reported, patient-reported and performance outcome [29]. There are certain types of COAs derived from mobile health technologies (e.g., activity monitors, sleep monitors) that do not fall into one of the other types of COAs. ClinRO: Clinician-reported outcome; COA: Clinical outcome assessment; ObsRO: Observer-reported outcome; PRO: Patient-reported outcome; PerFO: Performance outcome.



Outcome Measures: Use in Medical Product Development to Support Labeling Claims [11]; this Guidance was extended in 2014 to cover not only PROs but all COAs [27].

The properties considered in a COA's measurement review are content validity, construct validity, criterion validity, reliability (or reproducibility), responsiveness (also referred to as the ability to detect change), and score

Table 3. Summary of evidentiary requirements and measurement properties considered in the review of COA outcome instruments.

| Measurement property                       | Type                                   | What is assessed?                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | FDA review considerations                                                                                                                                                                                                                                                                                                                       |
|--------------------------------------------|----------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Reliability                                | Test–retest or intra-rater reliability | <ul style="list-style-type: none"> <li>Stability of scores over time when no change is expected in the concept of interest</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                       | <ul style="list-style-type: none"> <li>ICC coefficient</li> <li>Time period of assessment</li> </ul>                                                                                                                                                                                                                                            |
|                                            | Internal consistency                   | <ul style="list-style-type: none"> <li>Extent to which items comprising a scale measure the same concept</li> <li>Inter-correlation of items that contribute to a scale score</li> <li>Internal consistency of the items comprising a scale score</li> </ul>                                                                                                                                                                                                                                                                                                                | <ul style="list-style-type: none"> <li>Cronbach's <math>\alpha</math> for summary scores</li> <li>Item–total correlations</li> </ul>                                                                                                                                                                                                            |
|                                            | Inter-rater reliability                | <ul style="list-style-type: none"> <li>Agreement among responses when the COA is administered by <math>\geq 2</math> different raters</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                            | <ul style="list-style-type: none"> <li>ICC coefficient</li> </ul>                                                                                                                                                                                                                                                                               |
| Validity                                   | Content validity                       | <ul style="list-style-type: none"> <li>Evidence that the instrument measures the concept of interest including evidence from qualitative studies that the items and domains of an instrument are appropriate and comprehensive relative to its intended measurement concept, population and use. Testing other measurement properties will not replace or rectify problems with content validity.</li> </ul>                                                                                                                                                                | <ul style="list-style-type: none"> <li>Derivation of all items</li> <li>Qualitative interview schedule</li> <li>Interview or focus group transcripts</li> <li>Items derived from the transcripts</li> <li>Composition of patients used to develop content</li> <li>Cognitive interview transcripts to evaluate patient understanding</li> </ul> |
|                                            | Construct validity                     | <ul style="list-style-type: none"> <li>Evidence that relationships among items, domains and concepts conform to <i>a priori</i> hypotheses concerning logical relationships that should exist with measures of related concepts or scores produced in similar or diverse patient groups</li> </ul>                                                                                                                                                                                                                                                                          | <ul style="list-style-type: none"> <li>Strength of correlation testing <i>a priori</i> hypotheses (discriminant and convergent validity)</li> <li>Degree to which the COA instrument can distinguish among groups hypothesized <i>a priori</i> to be different (known groups validity)</li> </ul>                                               |
| Responsiveness or ability to detect change | N/A                                    | <ul style="list-style-type: none"> <li>Evidence that a COA instrument can identify differences in scores over time in individuals or groups (similar to those in the clinical trials) who have changed with respect to the measurement concept</li> </ul>                                                                                                                                                                                                                                                                                                                   | <ul style="list-style-type: none"> <li>Within person change over time</li> <li>Effect size statistic</li> </ul>                                                                                                                                                                                                                                 |
| Interpretation of scores                   | N/A                                    | <ul style="list-style-type: none"> <li>Regardless of whether the primary end point for the clinical trial is based on individual responses to treatment or the group response, it is usually useful to display individual responses, often using an <i>a priori</i> responder definition (i.e., the individual patient COA score change over a predetermined time period that should be interpreted as a treatment benefit). The responder definition is determined empirically and may vary by target population or other clinical trial design characteristics</li> </ul> | <ul style="list-style-type: none"> <li>Summary of the logic and methods used to interpret the clinical meaningfulness of clinical trial results at the individual patient level</li> <li>Responder definition (i.e., definition of meaningful within-person changes specific to the clinical trial population)</li> </ul>                       |

COA: Clinical outcome assessment; ICC: Inter-class correlation; N/A: Not applicable.

interpretation (not considered a measurement property) [11]. Table 3 provides the FDA's definition of each of these measurement topics as well as example methods or measurement properties for demonstrating a well-defined and reliable COA [11].

### Criteria for real-world data

The suitability of RWE algorithms is directly affected by the quality of the underlying RWD. The generation and selection of an RWD source affects and informs the selection of RWE algorithms. Although an evaluation of RWD was not within scope of this review, the critical importance of understanding the nature and source of RWD and ensuring the data are sufficient to generate RWE for the decision or intended purpose is emphasized in the stakeholder recommendations [11–14,16–23,28]. The main criteria identified in these sources are as follows:

- **Relevance.** The extent to which the existing RWD source is adequate for evaluating the question, including the availability of the necessary data elements [12–14,16,21–23,28];
- **Reliability.** The extent to which the data results are the same across different measurements over time [11–14,16,18,19,21,22,28]. Data accrual (or data collection) is the process of overseeing the collection and storage of the data to include sufficient documentation to understand the process, including data provenance [12–14,16,17,21–23,28]. Data quality assurance and control is the process of assuring that data errors are minimized, quality standards are met, and data are reproducible [11,13,14,16,17,19,21–23,28];
- **Validity.** Whether the data accurately and faithfully represent what they are intended to represent [11–14,16,17,20–22,28].



## Discussion

This review attempted to synthesize the published stakeholder recommendations on RWE algorithms through April 2021. While stakeholder recommendations largely reflect the ‘relevant and reliable’ RWE framework [23], they also include related concepts like transparency and replicability, safety considerations, feasibility and quality process that should continue to be addressed in future regulator guidance. Stakeholder recommendations for evaluating RWE algorithms include both quantitative assessments (such as measures of accuracy) and qualitative assessments (such as clinical relevance).

We identified several important gaps in the current stakeholder recommendations on RWE algorithms. One important gap is the lack of comprehensive guidance on RWE algorithms, which represents a risk with growing impact as the stakes for RWE for a decision-making increase. For example, regulatory recommendations are absent regarding the relevant validation approach or types of evidence for different contexts of use. Context of use is increasingly recognized as a key factor for determining the adequacy of measurement tools [12,29]. The existing RWE algorithm recommendations from stakeholders focus on only a few specific contexts of use, including pharmacoepidemiologic safety study outcomes [13,17] and quality of care measures [19,20]. For other algorithm use cases (e.g., real-world outcome measures contributing evidence about the effectiveness of a medical product), future recommendations should describe suitable validation approaches and types of evidence. Guidance on other types of validation approaches besides criterion validity would be valuable. For some routinely collected measures of clinical effectiveness (e.g., tumor response or progression), there is no clear reference standard, and it may not be apparent what approaches would be acceptable to validate these types of real-world outcome measures. These outcome measures may need to be assessed in other ways to demonstrate clinical validity, such as considering face validity with experts, measuring reliability between abstractors, benchmarking with external standards or evaluating performance in terms of prediction or correlation with other related events (i.e., clinical validation [30]). Example approaches may include evaluating the correlation of results between related outcome measures [31,32]; correlation or comparison of results between published trials and corresponding real-world patient cohorts [33,34]; and correlation of results for the same population, intervention, and outcome across different RWD sources [35–37]. The type of validation or evidence necessary for evaluating an algorithm’s suitability for a particular context of use may not be obvious today, as there are examples of different expert interpretations about the relevant validation approach for different contexts of use, including real-world external control arm scenarios [38,39]. Comprehensive guidance on the type of validation approach relevant to various contexts of use is needed. Analysis of past guidance along with the growing number of RWE submissions could provide valuable insights on relevant validation approaches for different contexts of use.

Another related gap is the absence of recommendations on thresholds (quantitative or qualitative) for determining whether an algorithm is fit for purpose. For example, the FDA Mini-Sentinel health outcomes of interest initiative recommended a quantitative threshold of PPV of 0.7 or greater for acceptable algorithm performance [40]. While maximizing PPV is important, 0.7 is not an immutable threshold. Algorithm accuracy improves as operating characteristics (e.g., sensitivity and PPV) increase, but there is a trade-off between algorithm sensitivity and PPV. The decision regarding whether to use an algorithm with higher sensitivity or PPV can be made only in the context of use. In the context of rare diseases [41], designs where inclusive study populations are needed [42], or for observing variation in care in quality-of-care assessments [43], the sensitivity of cohort algorithms will be an important consideration. In the context of outcome measures, reduced algorithm sensitivity by itself (i.e., provided that misclassification is non-differential with respect to exposure status and specificity is perfect) will not bias the relative risk of an outcome but will have a detrimental impact on odds ratios [42]. This is not an exhaustive list of contexts of use but offers a few examples of considerations for making trade-offs between measures of algorithm accuracy. In short, the acceptable thresholds are informed by the algorithm’s context of use. The absence of thresholds in stakeholder recommendations may be attributable to the difficulty of defining relevant thresholds for all contexts of use. A potentially helpful framework for conceptualizing thresholds is the certainty framework, which reflects two dimensions: factors that influence the level of certainty about the algorithm’s fit for the use case versus factors that influence the level of certainty needed to make a decision [44]. The recent FDA draft guidance on RWD for assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products [45] includes these concepts; however, it does not describe the level of certainty needed for different contexts of use nor does it provide criteria for evaluating whether the validation evidence is sufficient

to meet the level of certainty needed. A framework that addresses the two dimensions of certainty, and including both quantitative and qualitative aspects, is needed.

With respect to algorithm accuracy and generalizability, a gap was noted in stakeholder recommendations regarding adjustment of predictive values (PPV and NPV) for case prevalence. Sensitivity and specificity are characteristics of the test, while PPV and NPV values are dependent on the prevalence of the condition in the population; thus, a low prevalence condition may give rise to a low PPV despite high algorithm sensitivity and specificity. Researchers evaluating and selecting algorithms for their studies need to understand the prevalence of the concept of interest in the validation study as well as in the data source in which the algorithm will be applied. If the data in the gold standard data source do not reflect the true prevalence and absence of the cases in the real-world target population (e.g., the validation dataset is enriched for cases), then the PPV and NPV cannot be accurately estimated directly from the data. These points are addressed in the new FDA draft guidance [45], although future guidance could offer potential solutions for handling this limitation. For example, when the true case prevalence is known, the PPV and NPV can be calibrated using the following formulas based on Bayes' theorem [46,47]:

$$PPV = \frac{sensitivity \times prevalence}{sensitivity \times prevalence + (1 - specificity) \times (1 - prevalence)}$$

And

$$NPV = \frac{specificity \times (1 - prevalence)}{(1 - sensitivity) \times prevalence + specificity \times (1 - prevalence)}$$

Additionally, emerging COA guidance (e.g., patient-focused drug development) [29], on validation of real-world patient- or physician-generated data could also be an important future component of stakeholder recommendations on RWE algorithms; for example, the concept of patient-relevant algorithms could be incorporated. Finally, a health equity lens could be applied to future algorithm recommendations. Subgroup analyses by race, ethnicity and language or social determinants of health factors, where available, may support or refute the relevance and validity of an algorithm from a health equity perspective. For example, subgroup analyses may be conducted to reveal whether algorithmic bias is present due to disparities in care and could introduce, perpetuate or exacerbate disparities in health outcomes [48–50].

A strength of this review is its unique, current synthesis of stakeholder recommendations for RWE algorithms. A synthesis of existing stakeholder recommendations is timely given the many and significant multi-stakeholder efforts underway to evaluate 'fit for purpose' of RWD in different contexts of use (e.g., RCT DUPLICATE [51], OPERAND [52], Friends of Cancer Research pilot programs [53], Duke-Margolis [3–5], other RWE demonstration projects [54,55] and related frameworks [6,56]) and to increase rigor in transparency and replicability of algorithms (e.g., STaRT-RWE [57], ISPOR RWE transparency initiative [58]). A limitation of this review is that there are additional stakeholders we did not consider; for example, the Japan Pharmaceuticals and Medical Devices Agency has released guidance on RWE algorithms for safety surveillance studies [59] that was not included in our review. At the time this manuscript was being developed, the FDA published draft guidance on RWD, which included recommendations on validation of RWE algorithms (referred to as 'operational definitions' in the guidance) [45]. We took this new guidance into consideration as we identified gaps for future stakeholder recommendations to address.

## Conclusion

Fit-for-purpose algorithms are essential for generating credible RWE and trustworthy decisions, and the stakes are growing. This paper collates the recommendations regarding RWE algorithms by diverse stakeholders, including regulators, payers and other governmental healthcare decision makers. Comprehensive guidance is needed on the various types of validation approaches relevant for different contexts of use. As noted, there are many criteria to consider, and trade-offs will be made to find an acceptable balance of all the evidence. Future recommendations can help clarify criteria and thresholds for deeming an algorithm as fit for purpose (e.g., a framework that describes the level of certainty needed for different contexts of use versus the level of certainty the available evidence provides). Future recommendations could also include suggestions for overcoming limitations in algorithm generalizability (e.g., adjusting PPV for different case prevalence across data sources), cite COA validation standards in the emerging



regulatory guidance on patient-focused drug development, and apply a health equity lens to algorithm evaluation. This is not an exhaustive list of gaps to be addressed but highlights the key topics and needs we identified in our review of the available stakeholder recommendations.

### Future perspective

The importance of stakeholder guidance on fit-for-purpose RWE algorithms continues to grow. As regulatory, payer, and other governmental stakeholders increasingly rely on RWE in their decision making, we anticipate the evolution of guidance on fit-for-purpose RWE algorithms and related topics (e.g., acceptable real-world end points) to address known and emerging scientific and policy gaps.

#### Executive summary

- The importance of and risks associated with real-world evidence (RWE) algorithms are growing as RWE is increasingly used to inform decision making by regulators, payers and others. A review of stakeholder criteria for RWE algorithms has not been previously published. We synthesized RWE algorithm recommendations from the US FDA, European Medicines Agency (EMA), International Conference on Harmonisation (ICH) regulators; payers: National Quality Forum and National Committee for Quality Assurance (NQF and NCQA); and other governmental organizations: the Agency for Healthcare Research and Quality and Patient-Centered Outcomes Research Institute (AHRQ, PCORI). We identified gaps that could be addressed in future stakeholder guidance.
- Key stakeholder considerations comprised qualitative and quantitative measures of relevance, validity, reliability, responsiveness, transparency and replicability, safety, feasibility and quality process. Stakeholders recommended accuracy measures (sensitivity, specificity and positive and negative predictive values), subgroups evaluation, specific considerations when assessing exposures and covariates, and assessment of the underlying RWD quality.
- Gaps include comprehensive guidance on relevant and acceptable validation approaches and the level of certainty needed for different contexts of use, along with criteria for assessing the level of certainty the validation evidence provides (i.e., conceptual framework for thresholds). Future guidance could also address adjustment of positive and negative predictive values (PPV and NPV, respectively) for case prevalence in different data sources and incorporate clinical outcome assessment validation principles (e.g., patient relevance) and a health equity lens to algorithm evaluation.

### Acknowledgments

The authors acknowledge and thank the following people for their input, review and comments during the development of this targeted stakeholder review: A Ali, C Vehling, D Haldane, J Mount, N Kellier-Steele, X Zhang, BL Thompson, K Marie Schroeder and K Kinchen. The authors thank D Schamberger for editorial review and D Nelson for quality review of the manuscript.

### Financial & competing interests disclosure

This study was funded by Eli Lilly and Company. J Beyrer, H Abedtash, K Hornbuckle and JF Murray are employees and shareholders of Eli Lilly and Company. JF Murray has served in leadership positions within the International Society for Pharmacoconomics and Outcomes Research (ISPOR). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

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

### Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

### References

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

1. Berger ML, Sox H, Willke RJ *et al.* Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Value Health* 20(8), 1003–1008 (2017).
2. Wang SV, Schneeweiss S, Berger ML *et al.* Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1.0. *Value Health* 20(8), 1009–1022 (2017).

3. Daniel G, Silcox C, Bryan J, McClellan M, Romine M, Frank K. Characterizing RWD quality and relevancy for regulatory purposes. Duke-Margolis Center for Health Policy, Washington, DC, USA (2018) (Accessed 8 May 2021). [https://healthpolicy.duke.edu/sites/default/files/2020-03/characterizing\\_rwd.pdf](https://healthpolicy.duke.edu/sites/default/files/2020-03/characterizing_rwd.pdf)
4. Mahendraratnam N, Eckert J, Mercon K *et al.* Understanding the need for non-interventional studies using secondary data to generate real-world evidence for regulatory decision making and demonstrating their credibility (2019) (Accessed 8 May 2021). <https://healthpolicy.duke.edu/publications/understanding-need-non-interventional-studies-using-secondary-data-generate-real-world>
5. Mahendraratnam N, Silcox C, Mercon K *et al.* Determining real-world data's fitness for use and the role of reliability (2019) (Accessed 8 May 2021). <https://healthpolicy.duke.edu/publications/determining-real-world-datas-fitness-use-and-role-reliability>
6. Gatto NM, Reynolds RF, Campbell UB. A structured preapproval and postapproval comparative study design framework to generate valid and transparent real-world evidence for regulatory decisions. *Clin. Pharmacol. Ther.* 106(1), 103–115 (2019).
7. Orsini LS, Berger M, Crown W *et al.* Improving transparency to build trust in real-world secondary data studies for hypothesis testing—why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Value Health* 23(9), 1128–1136 (2020).
8. Curtis MD, Griffith SD, Tucker M *et al.* Development and validation of a high-quality composite real-world mortality endpoint. *Health Serv. Res.* 53(6), 4460–4476 (2018).
9. Hall GC, Lanes S, Bollaerts K, Zhou X, Ferreira G, Gini R. Outcome misclassification: impact, usual practice in pharmacoepidemiology database studies and an online aid to correct biased estimates of risk ratio or cumulative incidence. *Pharmacoepidemiol. Drug Saf.* 29(11), 1450–1455 (2020).
10. Lanes SF, de Luise C. Bias due to false-positive diagnoses in an automated health insurance claims database. *Drug Saf.* 29(11), 1069–1075 (2006).
11. United States, Food and Drug Administration (FDA). Patient-reported outcome measures: use in medical product development to support labeling claims. Guidance for Industry (2009) (Accessed 30 November 2018). [www.fda.gov/downloads/drugs/guidances/ucm193282.pdf](http://www.fda.gov/downloads/drugs/guidances/ucm193282.pdf)
- **Describes fit-for-purpose assessments of clinical outcomes assessment (COA) tools, an analogous and earlier regulatory precedent of fit-for-purpose measurement.**
12. Developing a protocol for observational comparative effectiveness research: a user's guide. Agency for Healthcare Research and Quality (AHRQ). Effective Health Care Program on developing a protocol for observational comparative effectiveness research: a user's guide. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM (Eds). Agency for Healthcare Research and Quality, MD USA (Accessed 30 November 2018). [www.ncbi.nlm.nih.gov/books/NBK126190/pdf/Bookshelf\\_NBK126190.pdf](http://www.ncbi.nlm.nih.gov/books/NBK126190/pdf/Bookshelf_NBK126190.pdf)
- **Provides scientific considerations for assessing real world evidence (RWE) algorithm fit-for-purpose, including unique considerations for exposures and covariates in Chapter 4.**
13. United States, Food and Drug Administration (FDA). Best practices for conducting and reporting pharmacoepidemiologic safety studies using electronic healthcare data. Guidance for industry and FDA staff. Updated April 29, 2020 (Accessed 30 November 2018). [www.fda.gov/downloads/drugs/guidances/ucm243537.pdf](http://www.fda.gov/downloads/drugs/guidances/ucm243537.pdf)
- **Describes the US FDA regulatory perspective on validated RWE algorithms for use in pharmacoepidemiology safety studies using electronic healthcare data.**
14. European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. The ENCePP guide on methodological standards in pharmacoepidemiology (Revision 5). Updated 5 July, 2016 (Accessed 22 October 2021). [www.encepp.eu/standards\\_and\\_guidances/documents/ENCEPPGuideofMethStandardsinPE\\_Rev5.pdf](http://www.encepp.eu/standards_and_guidances/documents/ENCEPPGuideofMethStandardsinPE_Rev5.pdf)
- **Provides scientific considerations for assessing RWE algorithm fit-for-purpose.**
15. Porta M (Ed.). *A Dictionary of Epidemiology. (6th Edition)*. Oxford University Press, NY, USA (2014).
16. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH harmonised tripartite guideline: statistical principles for clinical trials—E9 (1998). (Accessed 14 December 2021). [https://database.ich.org/sites/default/files/E9\\_Guideline.pdf](https://database.ich.org/sites/default/files/E9_Guideline.pdf)
17. European Medicines Agency. Head of Medicines Agencies. Guideline on good pharmacovigilance practices (GVP) Module VIII - Post-authorization safety studies (Rev 3) (2017) (Accessed 23 April 2018). [www.ema.europa.eu/docs/en\\_GB/document\\_library/Scientific\\_guideline/2012/06/WC500129137.pdf](http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129137.pdf)
- **Describes the European regulatory perspective on validated RWE algorithms in post-authorization safety studies.**
18. United States, Food and Drug Administration (FDA). Drug and device manufacturer communications with payors, formulary committees, and similar entities — questions and answers. Guidance for industry and review staff (Accessed 30 November 2018). [www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm537347.pdf](http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm537347.pdf)
19. National Committee for Quality Assurance. *The healthcare effectiveness data and information set 2018 – volume 1 (Epub)*. National Committee for Quality Assurance (September 2017).
20. National Quality Forum. Review and update of guidance of evaluating evidence and measure testing (2013). (Accessed 30 November 2018). [www.qualityforum.org/Publications/2013/10/Review\\_and\\_Update\\_of\\_Guidance\\_for\\_Evaluating\\_Evidence\\_and\\_Measure\\_Testing\\_-\\_Technical\\_Report.aspx](http://www.qualityforum.org/Publications/2013/10/Review_and_Update_of_Guidance_for_Evaluating_Evidence_and_Measure_Testing_-_Technical_Report.aspx)

21. Patient-Centered Outcomes Research Institute (PCORI). PCORI methodology standards (2019) (Accessed 30 November 2018). [www.pcori.org/research-results/about-our-research/research-methodology/pcori-methodology-standards](http://www.pcori.org/research-results/about-our-research/research-methodology/pcori-methodology-standards)
22. European Medicines Agency. Scientific guidance on post-authorisation efficacy studies (2015). (Accessed 22 November 2018). [www.ema.europa.eu/docs/en\\_GB/document\\_library/Scientific\\_guideline/2015/11/WC500196379.pdf](http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/11/WC500196379.pdf)
- **Describes the European regulatory perspective on validated RWE algorithms in post-authorization efficacy studies.**
23. United States, Food and Drug Administration (FDA). Framework for FDA's real-world evidence program (2018). (Accessed 22 November 2018). [www.fda.gov/media/120060/download](http://www.fda.gov/media/120060/download)
- **Describes the US FDA framework for assessing fit-for-purpose RWE.**
24. National Archives and Records Administration, U.S. Department of Health and Human Services. Adequate and well-controlled studies. 21 CFR 314.126. (Accessed 17 December 2018). [www.ecfr.gov/cgi-bin/text-idx?SID=9196841351bca689cfdfc2b81e8f585f&mc=true&node=pt21.5.314&rgn=div5#se21.5.314\\_1126](http://www.ecfr.gov/cgi-bin/text-idx?SID=9196841351bca689cfdfc2b81e8f585f&mc=true&node=pt21.5.314&rgn=div5#se21.5.314_1126)
25. Powers JH 3rd, Patrick DL, Walton MK *et al*. Clinician-reported outcome assessments of treatment benefit: report of the ISPOR clinical outcome assessment emerging Good Practices Task Force. *Value Health* 20(1), 2–14 (2017).
26. United States Congress. *Public Law*. 114–255 (2016) (Accessed 30 November 2018). [www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf](http://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf)
27. Critical Path Institute. FDA, US Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER) guidance for industry and FDA staff: qualification process for drug development tools (2014) (Accessed 17 November 2021). [www.c-path.org/wp-content/uploads/2014/01/FDA-releases-guidance-for-drug-development-tool-qualification.pdf](http://www.c-path.org/wp-content/uploads/2014/01/FDA-releases-guidance-for-drug-development-tool-qualification.pdf)
28. United States, Food and Drug Administration (FDA). Guidance for industry and Food and Drug Administration staff: use of real-world evidence to support regulatory decision-making for medical devices (2017). [www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices](http://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices) (Accessed 8 May 2021).
29. United States, Food and Drug Administration (FDA). Presented at the Patient-focused Drug Development Guidance Public Workshop; October 15–16, 2018. Methods to identify what is important to patients & select, develop or modify fit-for-purpose clinical outcomes assessments. (Accessed 22 November 2021). [www.fda.gov/media/116281/download](http://www.fda.gov/media/116281/download)
30. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet]. Food and Drug Administration, US (2016). Silver Spring, MD Validation. 14 Nov 2017 (Updated 16 Nov 2020). [www.ncbi.nlm.nih.gov/books/NBK464453/](http://www.ncbi.nlm.nih.gov/books/NBK464453/) Co-published by National Institutes of Health (US), Bethesda (MD).
31. Griffith SD, Miksad RA, Calkins G *et al*. Characterizing the feasibility and performance of real-world tumor progression end points and their association with overall survival in a large advanced non-small-cell lung cancer data set. *JCO Clin. Cancer Inform.* 3, 1–13 (2019).
32. Griffith SD, Tucker M, Bowser B *et al*. Generating real-world tumor burden endpoints from electronic health record data: comparison of RECIST, radiology-anchored, and clinician-anchored approaches for abstracting real-world progression in non-small cell lung cancer. *Adv. Ther.* 36(8), 2122–2136 (2019).
33. Ma X, Bellomo L, Magee K, Bennette CS *et al*. Characterization of a real-world response variable and comparison with RECIST-based response rates from clinical trials in advanced NSCLC. *Adv. Ther.* 38(4), 1843–1859 (2021).
34. Huang Bartlett C, Mardekian J, Cotter MJ *et al*. Concordance of real-world versus conventional progression-free survival from a phase III trial of endocrine therapy as first-line treatment for metastatic breast cancer. *PLoS ONE* 15(4), e0227256 (2020).
35. Friends of Cancer Research (FOCR). FOCR on considerations for use of real-world evidence in oncology: lessons learned from Friends of Cancer Research collaborations. (Accessed 8 May 2021). [www.focr.org/sites/default/files/Use\\_of\\_Real-World\\_Evidence\\_in\\_Oncology-2020.pdf](http://www.focr.org/sites/default/files/Use_of_Real-World_Evidence_in_Oncology-2020.pdf)
36. Lasiter L, Tymejczyk O, Garrett-Mayer E *et al*. Real-world overall survival using oncology electronic health record data: friends of cancer research plot. *Clin. Pharmacol. Ther.* 111(2), 444–454 (2022).
37. Rivera DR, Henk HJ, Garrett-Mayer E *et al*. The Friends of Cancer Research real-world data collaboration pilot 2.0: methodological recommendations from oncology case studies. *Clin. Pharmacol. Ther.* 111(1), 283–292 (2022).
38. United States, Food and Drug Administration (FDA). Center for Drug Evaluation and Research; Office of Surveillance and Epidemiology; Office of Pharmacovigilance and Epidemiology; Division of Epidemiology I. Application number: 761163orig1s000 (other review[s]). Division of Epidemiology I (DEPI-I) review of the final report of a real-world evidence for efficacy study (MOR208C206). July 20, 2020. [www.accessdata.fda.gov/drugsatfda\\_docs/nda/2020/761163Orig1s000OtherR.pdf](http://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/761163Orig1s000OtherR.pdf) (Accessed 8 May 2021).
39. United States, Food and Drug Administration (FDA). Center for Drug Evaluation and Research application number: 761163orig1s000: multi-discipline review. BLA multi-disciplinary review and evaluation BLA 761163 Monjuvi (tafasitamab-cxix). [www.accessdata.fda.gov/drugsatfda\\_docs/nda/2020/761163Orig1s000MultidisciplineR.pdf](http://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/761163Orig1s000MultidisciplineR.pdf) (Accessed 8 May 2021).
40. Schmouk GT, Lee TA, Pickard AS *et al*. United States Food and Drug Administration (FDA) Mini-Sentinel Project. FDA White Paper. Mini-Sentinel methods — alternative methods for health outcomes of interest validation (2013). [www.sentinelinitiative.org/sites/default/files/surveillance-tools/validations-literature/Mini-Sentinel-Alternative-Methods-for-Health-Outcomes-of-Interest-Validation\\_0.pdf](http://www.sentinelinitiative.org/sites/default/files/surveillance-tools/validations-literature/Mini-Sentinel-Alternative-Methods-for-Health-Outcomes-of-Interest-Validation_0.pdf) (Accessed 30 November 2018).

41. Setoguchi S, Solomon DH, Glynn RJ, Cook EF, Levin R, Schneeweiss S. Agreement of diagnosis and its date for hematologic malignancies and solid tumors between medicare claims and cancer registry data. *Cancer Causes Control* 18(5), 561–569 (2007).
42. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. *J. Clin. Epidemiol.* 65(3), 343–349.e2 (2012).
43. Gold HT, Do HT. Evaluation of three algorithms to identify incident breast cancer in medicare claims data. *Health Serv. Res.* 42(5), 2056–2069 (2007).
44. Cocoros NM, Arlett P, Dreyer NA *et al.* The certainty framework for assessing real-world data in studies of medical product safety and effectiveness. *Clin. Pharmacol. Ther.* 109(5), 1189–1196 (2021).
45. United States, Food and Drug Administration (FDA). Draft guidance for industry. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. [www.fda.gov/regulatory-information/s](http://www.fda.gov/regulatory-information/s)earch-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory (Accessed 22 November 2021).
- **Describes the most recent US FDA draft guidance on assessing fit-for-purpose RWE algorithms (operational definitions) in studies using electronic medical records or claims data.**
46. Nattinger AB, Laud PW, Bajorunaite R, Sparapani RA, Freeman JL. An algorithm for the use of Medicare claims data to identify women with incident breast cancer. *Health Serv. Res.* 39(6 Pt 1), 1733–1749 (2004).
47. Fischer F. Using Bayes theorem to estimate positive and negative predictive values for continuously and ordinally scaled diagnostic tests. *Int. J. Methods Psychiatr. Res.* 30(2), e1868 (2021).
48. Reed A. Medicare AI will infer race to close health equity gap. Bloomberg Law. <https://news.bloomberglaw.com/privacy-and-data-security/medicare-ai-will-infer-race-to-close-health-equity-gap> (Accessed 22 November 2021).
49. Williams DR, Wyatt R. Racial bias in health care and health: challenges and opportunities. *JAMA* 314(6), 555–556 (2015).
50. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA* 322(24), 2377–2378 (2019).
51. RCT Duplicate. RCT duplicate: randomized controlled trials duplicated using prospective longitudinal insurance claims: applying techniques of epidemiology (2020). [www.rctduplicate.org](http://www.rctduplicate.org)
52. OptumLabs. Using RWD in regulatory decision-making (2020). [www.optumlabs.com/work/data-regulatory-decision.html](http://www.optumlabs.com/work/data-regulatory-decision.html)
53. Friends of Cancer Research. Considerations for use of real-world evidence in oncology (2020). [https://friendsofcancerresearch.org/sites/default/files/2020-10/Use\\_of\\_Real-World\\_Evidence\\_in\\_Oncology\\_0.pdf](https://friendsofcancerresearch.org/sites/default/files/2020-10/Use_of_Real-World_Evidence_in_Oncology_0.pdf)
54. United States, Food and Drug Administration (FDA). FDA Announces 4 Grant Awards for Projects Exploring the Use of Real-World Data to Generate Real-World Evidence in Regulatory Decision-Making (2020). [www.fda.gov/drugs/science-and-research-drugs/fda-announces-4-grant-awards-projects-exploring-use-real-world-data-generate-real-world-evidence](http://www.fda.gov/drugs/science-and-research-drugs/fda-announces-4-grant-awards-projects-exploring-use-real-world-data-generate-real-world-evidence)
55. Jaksa A, Mahendraratnam N. Learning from the past to advance tomorrow's real-world evidence: what demonstration projects have to teach us. *J. Comp. Eff. Res.* 10(16), 1169–1173 (2021).
56. Gatto NM, Campbell UB, Rubinstein E *et al.* The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework. *Clin. Pharmacol. Ther.* 111(1), 122–134 (2022).
57. Wang SV, Pinheiro S, Hua W *et al.* STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ* 372, m4856 (2021).
58. Orsini LS, Berger M, Crown W *et al.* Improving transparency to build trust in real-world secondary data studies for hypothesis testing-why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Value Health* 23(9), 1128–1136 (2020).
59. [PMDA] Japan Pharmaceuticals and Medical Devices Agency. Formulation of “basic concepts for validating outcome definitions used in post-marketing database surveillance” (2020) (translation from Japanese). [www.pmda.go.jp/files/000235927.pdf](http://www.pmda.go.jp/files/000235927.pdf)
60. Ryan Patrick. Observational Health Data Sciences and Informatics (OHDSI). Forum; Building a validated OHDSI outcome library; Response in comment section (2018). <https://forums.ohdsi.org/t/building-a-validated-ohdsi-outcome-library/4422/9> (Accessed 22 November 2021).
61. Silva dS. *Cancer epidemiology: Principles and methods*. International Agency for Research on Cancer (IARC) (1999) (Accessed 21 November 2018). <http://publications.iarc.fr/Non-Series-Publications/Other-Non-Series-Publications/Cancer-Epidemiology-Principles-And-Methods-1999>
62. White E, Hunt JR, Casso D. Exposure measurement in cohort studies: the challenges of prospective data collection. *Epidemiol. Rev.* 20(1), 43–56 (1998).
63. Donabedian A. The Milbank Memorial Fund Quarterly, Vol. 44, No. 3, Part 2: Health Services Research I. A Series of Papers Commissioned by the Health Services Research Study Section of the United States Public Health Service. Discussed at a Conference Held in Chicago, pp. 166 (Jul. 1966) (15–16 October 1965).
64. Kozma CM, Reeder CE, Schulz RM. Economic, clinical, and humanistic outcomes: a planning model for pharmacoeconomic research. *Clin. Ther.* 15(6), 1121–1132 (1993).

65. National Quality Forum (NQF). About us. [www.qualityforum.org/About\\_NQF/](http://www.qualityforum.org/About_NQF/) (Accessed 14 December 2021).
66. Agency for Healthcare Research and Quality (AHRQ). Agency for Healthcare Research and Quality: a profile. <https://www.ahrq.gov/cpi/about/profile/index.html> (Accessed 14 December 2021).
67. Patient-Centered Outcomes Research Institute (PCORI). About PCORI. <https://www.pcori.org/about/about-pcori> (Accessed 14 December 2021).