Received: 23 December 2019 | Revised: 5 March 2020 | Accepted: 1 April 2020

DOI: 10.1002/pds.5010

COMMENTARY

WILEY

Considerations when evaluating real-world data quality in the context of fitness for purpose

Matthew W. Reynolds[1] | Alison Bourke[2] | Nancy A. Dreyer[3]

[1]Real World Solutions, IQVIA, Maryland

[2]Real World Solutions, IQVIA United Kingdom, UK

[3]Real World Solutions, IQVIA, Massachusetts

Correspondence
Alison Bourke, Real World Solutions, IQVIA United Kingdom, UK.
Email: alison.bourke@iqvia.com

1 | INTRODUCTION

With the growing use of large real-world clinical datasets comes increased scrutiny about the quality of real-world data (RWD). The question of whether the data are good enough for decision-making is often raised.[1] It is important to understand that quality cannot be assessed by looking at data in isolation: Any evaluation of RWD quality underpinning real-world evidence (RWE) must consider whether that data source has the information to answer a given research question.

A way forward in the quality conundrum suggested by Girman et al[2] is to establish a framework for evaluating data appropriateness, also known as fitness for purpose, meaning the degree to which the chosen data source aligns with the ability to accurately and reliably address the research question being posed. Here, we offer a simple framework for characterizing the attributes of any RWD source and key aspects of research questions to facilitate the optimal matching of research needs and data sources to achieve meaningful and reliable results.

2 | CHARACTERISTICS OF RWD SOURCES

Key aspects of any RWD source relate to provenance, access, and curation. In this model, data custodians would provide the agreed-upon information on their data in advance in a clear format and make them accessible to all potential researchers.

1. Provenance

- Is the information collected within a specific health system, clinical setting, and/or geography? What are the data subjects' demographics and reasons for being in the database - this may affect which diseases and treatments are seen. (Note that we refer to "data subjects" rather than "patients" since study subjects may include healthy people in routine care).
- Why is the data being collected? What drives its recording? Might this lead to systematic bias such as coding to more expensive diagnoses?
- Who records the data, for example, doctors, nurses, administrators, patients? What is the process/operational aspects of data collection? Data captured will be influenced by the focus of the data collector and the collection method.
- For how long has data been collected? Have there been changes to the databases over time? For example, were new fields added or coding systems changed? How many patients are in the data source over time?
- What is the lag time between data recording and data availability for research? How often is the database updated?
- Is it possible to collect additional information from data recorders and/or data subjects? For example, information on smoking might not be routinely collected but perhaps could be requested.
- Has the data source been used for research in the past and is there any documentation (eg, published papers)?

2. Access

- What are the processes and costs to access the data? Is the path to access clear in terms of who/how to contact? Are there restrictions on how the data is used and by whom? Does research require approval by an ethics review board?

3. Curation

- Which data fields are collected? A comprehensive data dictionary is essential including which coding systems are used.
- What verification/validation checking is undertaken during and after data collection (eg, range and logic checks or comparison to source documentation)? Has the extent of completeness for key fields been documented, preferably with information on the possible reasons for missing data (eg, weight assessments missing in middle-age men but well completed in pregnant women)?
- Is data transformed from the original recording (eg, patient anonymization efforts or standardized coding of free text)?
- Are all records from individual patients/subjects linked together to avoid double counting?
- If subsets of the database are made available for research, details should be provided regarding standard operating procedures (SOPs) for extracting data? Are quality assurance checks in place to ensure that extraction meets expectations?

3 | KEY PARAMETERS OF ANY RESEARCH QUESTION

1. Cohort: Who are the data subjects of interest and what characteristics are needed to identify them?
   - Is there a restriction on age, gender, and/or geography?
   - In what clinical settings are cohort characteristics recorded?
   - Are the diseases and/or symptoms of interest identified with diagnosis codes, laboratory results, and/or detailed clinical assessments, and how specific must identification be in terms of description and severity, for example, rash or severe allergic contact dermatitis?
   - Does the disease/symptom require some form of validation, and if so, what information is needed for confirmation?
   - Is there a temporal component to identifying the cohort, for example, prior history or conditions/treatment needed to qualify data subjects?
2. Intervention(s): Are there interventions of interest?
   - In what clinical settings are the recording of the interventions likely to be made?
   - How can the intervention be described? If a drug, is the focus on a broad drug class or specific brand/generic and/or dose and route of administration?
   - Is it sufficient to know if the patient was prescribed treatment or that a prescription was filled or taken?
   - Is there a temporal component related to the intervention? For example, interventions that occurred before/after a diagnosis or symptoms?
3. Outcome(s) of interest: Are there one or more outcomes of interest?
   - In what settings is the outcome likely to be recorded? Are some outcomes expected to be challenging to find (eg, financial burden or patient quality of life)?
   - Is the outcome defined via diagnosis codes, treatments, and/or laboratory tests, etc. or other indicators? Is severity important?
   - Does the outcome require some form of validation, and if so, what information is required to confirm?
   - Are there any time specifications/windows for outcome identification?
4. Essential covariates: Are there any important covariates that need to be assessed?
   - Is the covariate information required, present within the RWD? For example, while age and gender are likely to be accessible, socioeconomic status, diet, and exercise may not be.
   - Are there other treatments/interventions/medical history/co-morbidities that need to be incorporated?
5. Timing
   - How much follow-up time is required?
   - What aspects of the health and technology system may change over time that warrant consideration?[3]
   - How quickly are results needed?

Once key components of the research question have been assessed, they must be prioritized. Distinguishing the "must-have" data from the "nice-to-have" is critical since there are many RWD sources available and no one individually is likely to match perfectly to the research needs. Knowing which components of the research study are essential will pair the possible database matches more effectively.

While the list above is not exhaustive in defining aspects of quality, it will help characterize both the data available and the question to be answered, and it should provide a breadth of consideration for assessing a data source's fitness for purpose. Many of the issues raised above may trigger more specific critical queries for any given research question. In some cases, not all the information needed to make a full assessment will be available in easily accessible RWD metrics and so specific questions to data source holders and/or feasibility studies may be needed, especially to assure that there are likely to be enough data subjects of interest.

4 | DISCUSSION

As part of the Sentinel Initiative, the FDA and other organizations are developing detailed recommendations on standards-based approaches to describing data and presenting data quality metrics.[4,5] Here, we offer some straightforward information parameters to guide anyone interested in gauging RWD appropriateness for any research question. The speed of achieving results may also affect the choice of RWD source. If the research question concerns a life or death issue where little evidence is available, then even an imperfect RWD source that can deliver results quickly may be useful to generate intermediate evidence as opposed to waiting potentially years for perfect data to develop.

Ideally, the data assessments described here should be easily accessible; however, some information may be proprietary so that, for

commercial reasons, not all the desired information can be published. It may be helpful to query the data owner, researchers who have used the data, and/or colleagues and professional peers to obtain gain deeper insight.

Finally, it is important to remember that RWD sources and the environment they operate in are dynamic and continually updated, so the assessment of quality metrics will need regulator updating. Moreover, good data quality needs to be accompanied by appropriate study design and analyses. On-going efforts to support the regulatory use of real-world data in the US, Europe, and China will help flesh out the methodological and design issues that will help achieve meaningful and reliable results.[6-8]

5 | CONCLUSION

While access to RWE and its use is increasing,[9] it is natural for scientists to question the "quality" of the "messy" underpinning RWD, especially for those schooled in the use of randomized clinical trials where researchers have active control of data collection, and budgets to support extensive quality control. Understanding quality is particularly important in the regulatory sphere where decisions based on RWE need to be justified judiciously.

Even if all the metrics for evaluating RWD quality were established and agreed-on, such criteria are unlikely to be enough for every scientific research purpose. Most RWD sources will be good fits for some research questions but not others. Data appropriateness needs to be gauged by reviewing the strengths and weaknesses of any dataset under consideration in the context of the research initiative, study design, budget, and time available to assemble relevant information.

Data quality is not an absolute metric that can predict utility in isolation. A quality assessment can only be evaluated with full knowledge of the research question and immediacy of the need of information. Only through the optimal pairing of data source and research question parameters can you have confidence in delivering a reliable conclusion.

CONFLICT OF INTEREST

The author(s) have no conflicts of interest to declare.

ORCID

Alison Bourke https://orcid.org/0000-0002-0005-9016

REFERENCES

1. Dreyer NA. Advancing a framework for regulatory use of real-world evidence: when real is reliable. *Ther Innov Regul Sci*. 2018;52(3):362-368.
2. Girman CJ, Ritchey ME, Zhou W, Dreyer NA. Considerations in characterizing real-world data relevance and quality for regulatory purposes: a commentary. *Pharmacoepidemiol Drug Saf*. 2019;28(4):439-442. https://doi.org/10.1002/pds.4697.
3. Bourke A, Bate A, Sauer BC, Brown JS, Hall GC. Evidence generation from healthcare databases: recommendations for managing change. *Pharmacoepidemiol Drug Saf*. 2016;25:749-754. https://doi.org/10.1002/pds.4004.
4. Sentinel. Standardization and querying of data quality metrics and characteristics for electronic health data. https://www.sentinelinitiative.org/sentinel/methods/standardization-and-querying-data-quality-metrics-and-characteristics-electronic. 2018. Accessed December 23, 2019.
5. Duke-Margolis Center for Health Policy. Determining real-world data's fitness for use and the role of reliability. https://healthpolicy.duke.edu/sites/default/files/u31/rwd_reliability.pdf. 2019. Accessed December 19, 2019.
6. Duke-Margolis Center for Health Policy. Adding real-world evidence to a totality of evidence approach for evaluating marketed product effectiveness. Updated December 19, 2019. https://healthpolicy.duke.edu/publications/adding-real-world-evidence-totality-evidence-approach-evaluating-marketed-product. Accessed January 21, 2020.
7. HMA Heads of Medicines Agencies. *HMA-EMA Joint Big Data Taskforce. Phase II Report: Evolving Data-Driven Regulation*. January 20, 2020. EMA/689902/2019.
8. New Medicinal Products Administration China. Real-World Evidence Supporting Drug Development and Review. Guiding Principles. January 6, 2020.
9. European Parliament. European Parliament resolution on enabling the digital transformation of health and care in the Digital Single Market; empowering citizens and building a healthier society. 2019. https://www.europarl.europa.eu/doceo/document/B-9-2019-0239_EN.html?utm_source=POLITICO.EU&utm_campaign=c7bf50a629-EMAIL_CAMPAIGN_2019_12_18_03_44&utm_medium=email&utm_term=0_10959edeb5-c7bf50a629-190445777. Accessed December 20, 2019.