When Can We Trust Real-World Data To Evaluate New Medical Treatments?

Gregory E. Simon[1,*], Andrew B. Bindman[2], Nancy A. Dreyer[3], Richard Platt[4], Jonathan H. Watanabe[5], Michael Horberg[6], Adrian Hernandez[7] and Robert M. Califf[8]

Concerns regarding both the limited generalizability and the slow pace of traditional randomized trials have led to calls for greater use of real-world evidence (RWE) in the evaluation of new treatments or products. RWE studies often rely on real-world data (RWD), including data extracted from healthcare records or data captured by mobile phones or other consumer devices. Global assessments of RWD sources are not helpful in assessing whether any specific RWD element is fit for any specific purpose. Instead, evidence generators and evidence consumers should clearly identify the specific health state or clinical phenomenon of interest and then consider each step between that clinical phenomenon and its representation in a research database. We propose specific questions regarding potential error or bias affecting each of those steps: Would a person experiencing this clinical phenomenon present for care in this setting or interact with this recording device? Would this clinical phenomenon be accurately recognized or assessed? How might the recording environment or tools affect accurate and consistent recording of this clinical phenomenon? Can data elements from different sources be harmonized, both technically (same format) and semantically (same meaning)? Can the original data elements be consistently reduced to a useful clinical phenotype? Addressing these questions requires a range of clinical, organizational, and technical expertise. Transparency regarding each step in the creation of RWD is essential if evidence consumers are to rely on RWE studies.

Traditional randomized clinical trials often fail to produce the evidence needed to inform practical decisions of patients, clinicians, health systems, and regulators.[1,2] By design, traditional clinical trials typically include highly selected patients receiving tightly controlled treatments in specialized research settings.[1–3] Those characteristics often reduce both the generalizability of trial results to real-world practice and the efficiency of evidence generation.[3–5] Recognizing the need for more relevant or generalizable evidence and a more efficient evidence-generating process, the National Academies of Science, Engineering, and Medicine organized a series of workshops, sponsored by the US Food and Drug Administration (FDA), focused on "Examining the Impact of Real-World Evidence on Medical Product Development."[6] Those workshops identified and explored specific issues in the design and interpretation of real-world studies. This paper addresses one common aspect of real-world evidence (RWE) studies: use of real-world data (RWD) or data not created primarily by and for research.[7,8] Two companion papers address two other common aspects of RWE studies: delivery of treatments in real-world practice and valid inference from observational or nonrandomized comparisons.

Although the RWE and RWD labels have been variably defined, this discussion will follow definitions recently suggested by the FDA. In that scheme, RWD include data regarding patient health status or health care delivery routinely from a variety of sources, including electronic health records (EHRs), insurance claims, and patient-generated data created outside of healthcare settings. Regardless of the specific source, a defining characteristic of RWD is use of routinely collected data not generated in traditional research encounters. RWE refers to evidence derived from analysis of RWD, generated by a variety of research designs, including both randomized and nonrandomized comparisons.

THE RELATIONSHIP BETWEEN REAL-WORLD EVIDENCE AND REAL WORLD DATA

Although RWE and RWD often overlap in practice, they are conceptually distinct.[1] RWD may support a range of RWE studies, including randomized trials. For example, the Salford Lung Study[9] used a traditional randomized design and protocolized treatment but relied on RWD (extracted from community health system and pharmacy records) to assess both treatment exposure and clinical outcomes as a primary mode of data collection. The IMPACT-Afib trial[10] also used a patient-level randomized design and relies entirely on insurance claims data to assess eligibility, receipt of treatment, and study outcomes.

Following the definition above, RWD can include a wide range of data types and data sources. This discussion focuses on two common categories of RWD:

[1]Kaiser Permanente Washington Health Research Institute, Seattle, Washington, USA; [2]Kaiser Foundation Health Plan and Hospitals, Redwood City, California, USA; [3]IQVIA Real World Solutions, Durham, North Carolina, USA; [4]Harvard Pilgrim Health Care Institute and Harvard Medical School, Hartford, Connecticut, USA; [5]University of California Irvine School of Pharmacy and Pharmaceutical Sciences, Irvine, California, USA; [6]Kaiser Permanente Mid-Atlantic Permanente Research Institute and Mid-Atlantic Permanente Medical Group, Rockville, Maryland, USA; [7]Duke Clinical Research Institute, Durham, North Carolina, USA; [8]Verily Life Sciences and Google Health, South San Francisco, California, USA. *Correspondence: Gregory E. Simon (gregory.e.simon@kp.org)

Received February 9, 2021; accepted March 24, 2021. doi:10.1002/cpt.2252

- Data generated by routine healthcare operations, including data from extracted from EHRs, insurance claims for medical or pharmacy services, and other clinical or administrative data systems used to support healthcare delivery or payment.
- Data created outside of healthcare settings, including data captured by mobile phones (including data actively recorded by users or data automatically collected by passive sensing), data captured by other consumer devices, such as fitness trackers or glucometers, and data generated by commercial entities for "non-medical" purposes (consumer purchases, internet searches, and social media interactions).

The former are sometimes called the "data exhaust" of healthcare operations, whereas the latter are the data exhaust of daily life.

In any study, use of RWD does not dictate other aspects of research design, neither requiring nor precluding other departures from traditional clinical trial methods. RWD may be also combined with more traditional sources of research data. For example, pragmatic randomized trials involving treatment in community settings by community providers may rely on RWD extracted from EHRs or pharmacy records,[9] or may rely on data collected specifically for research. Nonrandomized research designs may rely on traditional research-specific data collection[11,12] or on data collected in routine healthcare operations.[13,14]

WHEN ARE RWD FIT FOR PURPOSE?

Not all RWD are fit or suitable for generating credible evidence. The potential advantages of using RWD (relevance and efficiency) must be weighed against concerns regarding quality and consistency. Miksad and Abernethy[7] have recently described the attributes required for RWD to be "fit for purpose" for generating valid evidence, including high quality, completeness, transparency, generalizability, timeliness, and scalability. Those attributes are relevant to efficiency, generalizability, and validity. We focus here on the attributes most essential for valid inference or trustworthiness: quality, completeness, and transparency. We specify key steps in the creation and curation of RWD and specify question that users of RWD should consider regarding each of those steps.

We cannot rely on global judgments regarding the credibility or utility of any data extracted from health records or recorded on mobile phones. Instead, we must assess the credibility of a specific data source for a specific measurement task. For any real-world investigation, including both observational or randomized comparisons, RWD may be used to evaluate range of health states or healthcare events, including eligibility or inclusion criteria, covariates or potential confounders, treatment exposures, treatment outcomes, and adverse events. Any RWD data source could be fit for one of those purposes and unfit for another. In other words, we should not ask "Can real-world data support valid inference regarding the effectiveness or safety of medical treatments?" Instead we should ask "*When* can a specific data element from a specific real-world data source support valid inference regarding the effectiveness or safety of a specific medical treatment?" For example, EHR data from a referral center might be appropriate for assessing baseline characteristics and initial exposure to a new medical product but might fail to accurately capture subsequent outcomes or adverse events presenting in community practice. Specific procedure codes from insurance claims data could accurately identify a specific health service in one care setting and not in another. A mobile phone sensor might accurately record sedentary time in one clinical or demographic group and not in another.

We must remember that the measures or definitions used in traditional clinical trials (such as expert clinician review of medical records) are also subject to error. In many cases, those so-called gold standards are also proxies for the actual health states or clinical phenomenon we hope to measure. When we attempt to validate an RWD-based measure against a traditional clinical trial measure, we are usually not assessing criterion validity (agreement between an imperfect measure and a true gold standard). Instead, we are assessing convergent validity (agreement between two imperfect measures of a true state that is not directly observed). For example, insurance claims from real-world practice settings likely provide much more accurate assessment of treatment effects on healthcare utilization or cost than do interview-based data regarding treatment in research settings. Health records data may be more accurate than recall for assessment of past health states.[15]

The pathway from a health state or healthcare event of interest to its representation in a research dataset includes a series of specific steps or transitions. For data derived from EHRs or insurance claims, for example, that pathway typically travels from patient to provider to EHR to health system data warehouse to multisite research dataset. For data captured by a mobile or consumer device, that pathway typically travels from patient to device to device vendor server to harmonized research database to research dataset. Each of those transitions is liable to error, and a robust assessment of data quality must identify and evaluate each of these steps. We propose below a specific framework and process for identifying points of transition and assessing sources of error at each of those transition points.

DEFINING THE CLINICAL PHENOMENON OF INTEREST

To thoroughly investigate the pathway from a clinical phenomenon to a research dataset, we must first clearly specify where that pathway begins. Efforts to take advantage of routinely collected data often begin with the question "What data do we have?" or "What data could we easily collect?" rather than "What is the thing we are trying to measure?" We can only assess the credibility of a variable in an analytic dataset after we precisely define the clinical phenomenon that we hope it accurately represents. A clinical phenomenon or healthcare event of interest could be defined by a biological event (e.g., thromboembolic stroke), a subjective experience (e.g., symptoms of depression), or a measurable performance (e.g., timed walk). A treatment exposure could be defined in terms of a one-time procedure (e.g., prosthetic valve implantation), an ongoing exposure (e.g., regular use of an antihypertensive drug over a specific period), or a self-care behavior (e.g., regular use of a digital therapeutic or mobile phone app). For any of these events, we often hope to identify a computable phenotype, an accurate indication, or representation of that clinical phenomenon computed from some RWD source. Our strategies for assessing any computable phenotype would likely vary depending on the nature or location of that clinical phenomenon (biological event vs. subjective experience vs. measurable performance) or

treatment exposure (provider-delivered vs. patient administered). To assess the pathway for a biological event, we would likely trace back to a measured physiologic parameter (e.g., blood pressure) or laboratory or imaging data (e.g., troponin level). In contrast, the pathway for a patient experience would likely pass through patient-reported outcomes or clinician-reported symptoms. The pathway for a treatment exposure might be traced back to a either a provider-delivered intervention or a patient behavior.

QUESTIONS TO ASSESS WHETHER A REAL-WORLD DATA SOURCE IS FIT FOR A SPECIFIC PURPOSE

After clearly specifying each clinical phenomenon or health state of interest, the investigator hoping to assess the credibility of RWD should identify each specific step or transition in the pathway from that clinical phenomenon to its ultimate representation as a clinical phenotype in a research dataset. Distinguishing each step permits formulating specific questions regarding distinct sources of error or bias at each of those transitions. Those questions are shown in **Table 1** and discussed below.

Question 1 (presentation): Would a person experiencing this health state or clinical phenomenon present for assessment?

For data derived from health system records, we must consider whether a person experiencing this phenomenon would present for care at all—or present in setting(s) from which records data are available. Regarding EHR data, we would ask whether a person in the study population experiencing this clinical phenomenon or receiving this health service would present to a facility using this EHR system or instead receive care in some other setting. Regarding insurance claims data, we would ask whether care for a person experiencing this clinical phenomenon would generate an insurance claim or if care might instead be paid out of pocket or paid by alternative insurance. We should also consider how likelihood of presentation (and completeness of eventual data capture) vary across different stages of illness or phases of care and how characteristics of the individual or health system could make presentation for initial or follow-up care more or less likely. When any individual receives care across multiple healthcare systems, records from any single health system may be more likely to capture stable or chronic health states (such as diabetes) than episodic health states (such as transient ischemic attacks).

For data captured by mobile or consumer devices, we must consider whether a person experiencing this phenomenon or health state would interact with the sensing or recording device or process. Regarding passive sensing by mobile devices, we would ask how often the device would be worn or carried (or not) when the event of interest occurs. Regarding data actively recorded by patients or participants, we would ask whether patients or participants would respond to questions or prompts—and how response might vary by personal characteristics and situation. When individuals interact intermittently with a mobile device or sensor, completeness of data from that device or sensor may be greater for stable phenomena (such as weight) than episodic phenomena (such as falls). We should also consider how likelihood of capture or ascertainment could vary both between individuals and within individuals over time.

Example. The ADAPTABLE trial[16] compared alternative doses of aspirin for prevention of all-cause mortality, hospitalization for myocardial infarction, or hospitalization for stroke. Hospitalization outcomes were ascertained using records data from participating health systems, but investigators recognized that this data source might not capture hospitalizations in other facilities. Consequently, ascertainment by records data was supplemented with patient-reported data regarding hospitalization for potential outcome events—with cross-validation of patient-reported data and records data when feasible.

Question 2 (recognition/assessment): Would the clinical phenomenon or health state of interest be accurately recognized or assessed?

For data extracted from health system records, we should ask whether real-world providers in study settings would be able to accurately recognize and/or assess the phenomenon of interest.

Table 1 Questions regarding use of RWD to accurately represent a specific clinical phenomenon or health state

	Data extracted from EHRs or insurance claims	Data recorded by mobile sensors or other connected consumer devices
Presentation	Would a person experiencing this phenomenon present for care in this setting?	Would a person experiencing this phenomenon interact with the sensor or device?
Recognition/assessment	Would clinicians in this setting accurately recognize or diagnose this phenomenon?	Can people experiencing this phenomenon accurately report it? Or can passive sensors accurately detect it?
Recording	How might the technical/social/economic environment affect recording of this phenomenon?	How might characteristics of specific recording systems or devices affect accuracy of detection or assessment?
Harmonization	Can primary data elements be combined—both technical and semantically?	Can data elements from different sensing devices or recording systems be combined?
Reduction	Will processes to reduce primary data to clinical phenotypes perform similarly across settings?	Will processes to reduce primary data to clinical phenotypes perform similar across devices or systems?

EHRs, electronic health records; RWD, real-world data.

We should also ask what characteristics of patients, providers, or health systems might affect accuracy of recognition or assessment.

For data captured by mobile or consumer devices, we should ask whether people experiencing the health state of interest accurately report that experience through a consumer device or whether available sensing technologies could accurately detect this phenomenon—considering both false positive and false negative error rates. We should also ask how accuracy of reporting or accuracy of sensing could vary according to any individual's demographic or clinical characteristics.

Example. The Suicide Prevention Outreach Trial[17] evaluated two population-based outreach programs to prevent self-harm or suicide attempt among high-risk outpatients. Self-harm events or suicide attempts were ascertained from diagnostic codes for intentional self-harm, extracted from health system EHRs and insurance claims data. Previous research[18] suggested that some self-harm events or suicide attempts might be misdiagnosed as having accidental or "undetermined" intent. Identifying these additional self-harm events required review of clinical text from selected injury and poisoning encounters to identify those "missed" self-harm events.[19]

Question 3 (recording): How might the recording environment or tools affect accurate and consistent recording of the phenomenon of interest?

For data captured by health system records, we should ask how the environment (EHR characteristics, financial incentives, and social influences) in any healthcare system might affect a provider's accurate recording of their diagnosis or assessment. Effects of financial incentives or social influences (such as stigma attached to specific diagnoses or treatments) could vary across patient groups.

For data captured by mobile or consumer devices, we should ask how characteristics of any specific recording system (a specific sensor, mobile app, or device vendor) might affect accurate or consistent recording of the phenomenon or health state of interest. Availability and quality of mobile device sensorss or internet-connected consumer products may vary significantly across demographic and clinical populations.

Example. The LASSY observational study[20] used an interrupted time series method to examine the impact of safety advisories regarding antidepressant medications on antidepressant prescribing and risk of suicidal behavior. In the International Classification of Disease, 9th revision-Clinical Modification (ICD-9-CM) classification system, self-harm or suicide attempt was indicated by an primary injury or poisoning diagnosis accompanied by a supplemental cause-of-injury code (or E-code) indicating self-harm. Review of data from participating health systems indicated that use of supplementary E-codes varied widely between health systems and within health systems over time.[21] Sudden changes in E-code use corresponded with introduction of or changes to EHRs systems. Given the interrupted time series design, changes in recorded suicide attempt rates in any health system due to safety advisories could not be distinguished from changes due to changes in health records systems. Consequently, analyses of suicide attempt rates used an alternative specification not dependent on use of E-codes.

Question 4 (harmonization): Can data from different sources be harmonized – both technically (recorded in the same format) and semantically (have the same actual meaning)?

For data captured by health system records, we should consider both the technical barriers to combining data from multiple sources as well as the potential for different health care environments (EHR characteristics, financial incentives, and social influences) to influence the meaning of seemingly identical data from different sources. Technical barriers may preclude harmonization or require sacrificing important detail or precision.

For data captured by mobile or consumer devices, we should ask how data elements from different sensing devices or patient interfaces can be reliably combined without loss or distortion of meaning. This question involves both compatibility of data types (technical interoperability) and variation in clinical performance characteristics across recording systems (semantic interoperability). Differences in financial incentives or social influences may, for example, affect recording of weight, diet, or physical activity.

Example. The ADAPTABLE trial[16] comparing aspirin doses identified potential participants with known atherosclerotic cardiovascular disease using health system records. Eligibility criteria were initially based on diagnosis codes for myocardial infarction, procedure codes for cardiac revascularization procedures, and angiography data indicating 75% or greater stenosis of one or more coronary arteries. These criteria could be consistently assessed using health systems' research databases organized according to a common data model (i.e., technically interoperable). However, incomplete capture of past procedures and differences in classification of data regarding assessment of coronary artery disease led to inconsistent performance of this computable phenotype across health systems (i.e., not semantically interoperable). Consequently, the computable phenotype for eligibility was modified to include additional diagnostic codes and accommodate site-specific data systems.

Question 5 (reduction): Can primary data elements be consistently reduced to a useful clinical phenotype?

For data captured by health system records, we should ask whether processes for reducing primary data elements (such as an assortment of diagnosis codes) to summary clinical phenotypes would have similar effects across different patient populations, health systems, or healthcare settings. Summarization or reduction may obscure potential problems with source data quality or procedures for harmonization. More complex computable phenotypes (such as combining diagnosis and drug dispensing data across time to define a new episode of treatment) may have differential performance across patient populations or healthcare settings (such as specialty clinics vs. primary care clinics).

For data captured by mobile or consumer devices, we should ask whether process for reducing original data elements (such as 24-hour electrocardiogram (ECG) sensing) to summary clinical phenotypes (such as an episode of atrial fibrillation) would have

similar effects across different recording devices, different patient/consumer interfaces, or different subgroups of patients or research participants.

Example. The PCORnet Bariatric Surgery Study[22] compared percent total weight loss following alternative bariatric procedures using data extracted from health system records. Reducing primary data (weight measurements extracted from EHRs) to a measure of weight loss required definition of time windows for weight measurements in routine clinical care, definition and rejection of implausible weight values, and exclusion of weight values during periods when records data indicated pregnancy loss or delivery during the preceding 3 months or following 9 months.

DISCUSSION

We aim to focus discussions regarding RWD from global judgments to more structured and specific assessments of data quality. We argue that such an assessment should begin with a clear specification of the clinical phenomena of interest, then describe the environment and process that generate the primary data, then map specific steps in the pathway from the clinical phenomenon to a research dataset. After describing that pathway, evidence producers and evidence consumers can then identify and evaluate specific sources of error bias at each step. This mapping and evaluation should occur separately for each clinical phenomenon (eligibility criterion, treatment exposure, covariate or potential confounder, outcome, etc.) rather than for a study or trial as a whole. Any specific data source might accurately reflect one phenomenon of interest while performing poorly for others.

As illustrated by the questions above, identifying error or bias at each step typically requires local understanding of data generation, recording, and processing. For data derived from health records, that would include understanding of health system practice patterns, record-keeping processes, and data systems. For data derived from mobile or consumer devices, that would include understanding of device characteristics, user experience, and technical aspects of data storage. In some cases, empirical data may be available to assess potential sources of error or bias. For example, we may be able to use full-text records data to re-evaluate accuracy of diagnosis[19] or use original data elements to examine consistency of data across different care settings.[21] In many cases, however, data to independently assess error or bias will not be available. At a minimum, that assessment should identify possible sources of error or bias, consider steps taken to address and mitigate those specific sources, and evaluate sensitivity of findings to both random and nonrandom errors.

Explicit assessment of each step in the chain requires transparency from data generators, data aggregators, and data stewards regarding processes and sources of error. Assessing whether a person experiencing a health state of interest would present for care in a specific care setting or health system requires transparency regarding patterns of insurance coverage, market share, and member/patient demographics. Assessing accuracy of recognition or diagnosis requires transparency regarding variation in diagnosis and treatment rates. Assessing data harmonization, aggregation, and reduction requires transparency regarding details of those processes—down to disclosure of each line of code used for data extraction, aggregation, and transformation.

The potential impact of specific errors depends on the framing of the study question or hypothesis. If we aim to detect differences between alternative treatments or practices, systematic error or bias is usually more concerning than random error. For example, we would be much more concerned if a provider's likelihood of recognizing or diagnosing an outcome of interest differed systematically between treatments than if that error was randomly distributed between treatments. Even completely random measurement error, however, is a significant concern in the case of a noninferiority comparison between a new treatment and an established one. Although random error typically creates a bias toward a null finding in either a randomized or observational comparison, a null finding in a noninferiority comparison could falsely imply clinical benefit equal to that of an established treatment.

Some sources of error may be amenable to repair or remediation. In general, later steps in the data pathway are more often under the influence or control of evidence creators. Error or bias at the stage of data reduction (e.g., incorrect specification of a computable phenotype) can often be addressed after data have been collected, extracted, and aggregated. In contrast, error at the point of clinical assessment or initial recording can only be addressed prospectively. For example, if we suspect that a significant proportion of people experiencing a clinical phenomenon of interest never present to a specific care setting, effective remediation is not possible. Evidence creators would instead need to limit research to care settings with fewer barriers to care-seeking or acknowledge that some groups (e.g., the uninsured) will not be represented.

Each step in the data pathway is liable to change over time, so users of RWD should periodically re-assess potential sources of error or bias. For a prospective study or evaluation, such as a pragmatic trial, that re-assessment would occur in real time. For a retrospective study or evaluation, that re-assessment would examine temporal discontinuities in key measures. Any research using RWD generated during 2020 must consider the dramatic effect of the coronavirus disease 2019 (COVID-19) pandemic on use of health services.[23] For data extracted from health records, re-assessment should pay special attention to changes in health system organization, payment models, or information systems that might affect presentation for care, diagnosis, or recording. For data captured by mobile or consumer devices, re-assessment should pay special attention to hardware or software upgrades or retirements and to factors affecting users' access or connectivity.

The range of specific questions we propose illustrates the range of expertise necessary to generate or evaluate evidence dependent on RWD. Addressing questions regarding presentation for care requires expertise regarding financing and organization of health services as well as individual-level determinants of help-seeking behavior, such as race and ethnicity. Addressing questions regarding data capture by mobile or consumer electronic devices requires expertise in assessment and design of user interfaces and user experience. Addressing questions regarding accuracy of diagnosis or assessment requires expertise in clinical epidemiology as well as domain-specific clinical knowledge. Addressing questions regarding accuracy of EHR recording requires expertise in clinical informatics

and healthcare operations. Addressing questions regarding accuracy of mobile device or consumer device recording may require expertise ranging from psychometrics to sensor engineering. Addressing questions regarding data harmonization, data aggregation, and data reduction requires expertise in data science and biostatistics.

CONCLUSIONS

Global judgments regarding the completeness or validity of any RWD source are not helpful in assessing whether RWD are fit for any purpose. Instead, evidence generators and evidence consumers should ask when a specific data element derived from a specific data source can completely and accurately assess a specific health state or healthcare event of interest. We propose a series of questions to assess when specific RWD are fit for a specific purpose, applying those questions both to data derived from healthcare records and data derived from consumer devices or other mobile sensors. Assessing fitness for purpose will often require both ground-level knowledge and specific methodologic expertise.

FUNDING

This study was supported by Simon—NIMH Cooperative Agreement U19MH092201.

CONFLICT OF INTEREST

G.E.S. is an employee of Kaiser Permanente Washington. A.B.B. is an employee of Kaiser Foundation Health Plan and Hospitals. R.P. has no relevant financial interests. J.H.W. has no relevant financial interests. N.A.D. us an employee of IQVIA Real World Solutions. M.H. is an employee of Kaiser Permanente Mid-Atlantic. A.H. has no relevant financial interests. R.M.C. is an employee of Verily and Google Health, Board member for Cytokinetics.

1. Corrigan-Curay, J., Sacks, L. & Woodcock, J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *JAMA* **320**, 867–868 (2018).
2. Sherman, R.E. *et al*. Real-world evidence - what is it and what can it tell us? *N. Engl. J. Med.* **375**, 2293–2297 (2016).
3. Eapen, Z.J., Lauer, M.S. & Temple, R.J. The imperative of overcoming barriers to the conduct of large, simple trials. *JAMA* **311**(14), 1397–1398 (2014).
4. Lauer, M.S., Gordon, D., Wei, G. & Pearson, G. Efficient design of clinical trials and epidemiological research: is it possible? *Nat. Rev. Cardiol.* **14**(8), 493–501 (2017).
5. Martin, L., Hutchens, M., Hawkins, C. & Radnov, A. How much do clinical trials cost? *Nat. Rev. Drug Discov.* **16**(6), 381–382 (2017).
6. National Academies of Sciences Engineering and Medicine, editor *Examining the Impact of Real-World Evidence on Medical Product development: Proceedings of a Workshop Series 2019* (National Academies Press, Washington, DC, 2019).
7. Miksad, R.A. & Abernethy, A.P. Harnessing the power of real-world evidence (RWE): a checklist to ensure regulatory-grade data quality. *Clin. Pharmacol. Ther.* **103**, 202–205 (2018).
8. Makady, A., de Boer, A., Hillege, H., Klungel, O. & Goettsch, W. What is real-world data? A review of definitions based on literature and stakeholder interviews. *Value Health* **20**(7), 858–865 (2017).
9. Woodcock, A. *et al*. Effectiveness of fluticasone furoate plus vilanterol on asthma control in clinical practice: an open-label, parallel group, randomised controlled trial. *Lancet* **390**, 2247–2255 (2017).
10. Garcia, C.J. *et al*. Practical challenges in the conduct of pragmatic trials embedded in health plans: Lessons of IMPACT-AFib, an FDA-Catalyst trial. *Clin. Trials* **17**, 360–367 (2020).
11. Lemery, S.J. *et al*. U.S. Food and Drug Administration approval: ofatumumab for the treatment of patients with chronic lymphocytic leukemia refractory to fludarabine and alemtuzumab. *Clin. Cancer Res.* **16**, 4331–4338 (2010).
12. Gokbuget, N. *et al*. Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. *Blood Cancer J.* **6**, e473 (2016).
13. Brown, J.S., Holmes, J.H., Shah, K., Hall, K., Lazarus, R. & Platt, R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med. Care* **48**(6 Suppl), S45–S51 (2010).
14. Brown, J.S., Kahn, M. & Toh, S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med. Care* **51**(8 Suppl 3), S22–S29 (2013).
15. Simon, G.E., Rutter, C.M., Stewart, C., Pabiniak, C. & Wehnes, L. Response to past depression treatments is not accurately recalled: comparison of structured recall and patient health questionnaire scores in medical records. *J. Clin. Psychiatry* **73**, 1503–1508 (2012).
16. Marquis-Gravel, G. *et al*. Rationale and design of the aspirin dosing-A patient-centric trial assessing benefits and long-term effectiveness (ADAPTABLE) TRIAL. *JAMA Cardiol.* **5**, 598–607 (2020).
17. Simon, G.E. *et al*. Population-based outreach versus care as usual to prevent suicide attempt: study protocol for a randomized controlled trial. *Trials* **17**(1), 452 (2016).
18. Walkup, J.T., Townsend, L., Crystal, S. & Olfson, M. A systematic review of validated methods for identifying suicide or suicidal ideation using administrative or claims data. *Pharmacoepidemiol. Drug Saf.* **21**(Suppl 1), 174–182 (2012).
19. Simon, G.E. *et al*. Estimating the number of self-harm events not identified by encounter diagnoses in health system records. *medRxiv*. https://doi.org/10.1101/2020.09.24.20200998.
20. Lu, C.Y. *et al*. Changes in antidepressant use by young people and suicidal behavior after FDA warnings and media coverage: quasi-experimental study. *BMJ* **348**, g3596 (2014).
21. Lu, C.Y. *et al*. How complete are E-codes in commercal plan claims databases? *Pharmacoepidemiol. Drug Saf.* **23**, 218–220 (2014).
22. Arterburn, D. *et al*. Comparative effectiveness and safety of bariatric procedures for weight loss: A PCORnet Cohort Study. *Ann. Intern. Med.* **169**, 741–750 (2018).
23. Jeffery, M.M. *et al*. Trends in emergency department visits and hospital admissions in Health care systems in 5 states in the first months of the COVID-19 pandemic in the US. *JAMA Intern. Med.* **180**, 1328–1333 (2020).