## Title

Examining the Use of Real-World Evidence in the Regulatory Process



## Journal

Clinical Pharmacology & Therapeutics, 107(4)

## ISSN

0009-9236

## Authors

Beaulieu-Jones, Brett K
Finlayson, Samuel G
Yuan, William
et al.

## Publication Date

2020-04-01

## DOI

10.1002/cpt.1658

# Examining the Use of Real-World Evidence in the Regulatory Process

Brett K. Beaulieu-Jones[1,*], Samuel G. Finlayson[1], William Yuan[1], Russ B. Altman[2], Isaac S. Kohane[1], Vinay Prasad[3] and Kun-Hsing Yu[1,*]

**The 21st Century Cures Act passed by the United States Congress mandates the US Food and Drug Administration to develop guidance to evaluate the use of real-world evidence (RWE) to support the regulatory process. RWE has generated important medical discoveries, especially in areas where traditional clinical trials would be unethical or infeasible. However, RWE suffers from several issues that hinder its ability to provide proof of treatment efficacy at a level comparable to randomized controlled trials. In this review article, we summarized the advantages and limitations of RWE, identified the key opportunities for RWE, and pointed the way forward to maximize the potential of RWE for regulatory purposes.**

Real-world data (RWD) and real-world evidence (RWE) have received substantial attention from medical researchers and regulators in recent years.[1,2] The US Food and Drug Administration (FDA) defines data relating to patient health status and the delivery of healthcare (such as electronic health records (EHRs), claims and billing activities, product and disease registries, and patient-generated data) as real-world data (RWD), and the analysis of these data regarding usage and effectiveness are termed real-world evidence (RWE).[3] The European Medicines Agency (EMA) similarly defines RWD as defined as "routinely collected data relating to a patient's health status or the delivery of health care from a variety of sources other than traditional clinical trials" and expressed interest in using RWD for regulatory decision making.[4] RWE presents great potential to accelerate therapy development and to monitor the successes and failures or both newly approved and existing therapies.[5] It is critical that stakeholders, including researchers, both academic and industry, providers, regulators, administrators, and patients understand the limitations of RWE. RWE is not generated with a particular study question in mind but are generated primarily for clinical care and billing purposes. As such, appropriate use of RWE must be driven by well-designed guidelines and regulations to ensure accurate, unbiased findings. If used correctly, RWE could supplement traditional clinical research to aid therapeutic development, clinical decision making efficiency gains in healthcare, and improved access to therapeutics for underserved populations. Examples of promise in each of these areas can be seen in the results of the EMA's adaptive pathways pilot.[6] If used incorrectly, RWE could lead to spurious approvals, financial waste, and most importantly cause harm to patients.

To date, RWE has been used primarily to perform postmarketing surveillance to monitor drug safety and detect adverse events. RWE has also been particularly effective when the outcome of interest is rare, in cases where a very long follow-up period is required to assess the health outcomes, or when it is difficult to perform randomized controlled trials (RCTs), such as in pediatric or pregnant populations. An early example was the discovery of a link between the ingestion of diethylstilbestrol during pregnancy and vaginal adenocarcinoma of the offsprings using observational data.[7] More recent studies linked the use of angiotensin-converting enzyme inhibitors while pregnant to congenital malformations[8] and the exposure of selective serotonin reuptake inhibitors to persistent pulmonary hypertension in newborns.[9,10] Most recently, RWE has shown postmarketing evidence that it may have an important role to play in understanding drug effectiveness and adverse events based on differences of metabolism in various racial and genetic groups.[11–14]

There is a growing interest in the usage of RWE by regulatory agencies to evaluate the safety and efficacy of medical treatments.[2,5,15–17] In particular, Congress has mandated that the FDA increase focus on RWE for regulatory decision making both for new approvals and evaluating additional indications for approved therapies[18] and the FDA has testified on progress toward implementing this focus.[19] The EMA recently accepted an RWE-based control arm during their analysis of Alecensa effectiveness compared with the standard of care.[20–22] In addition, the FDA has established partnerships with private companies whose goal is to use RWD in regulatory decision making, including using synthetic control arms.[20]

Although some have been enthusiastic about the ability for observational RWD to substitute for RCTs, others have expressed caution. Booth *et al.*[23] contend that RWD should not be used as a replacement for clinical trials due to the inability to compare outcomes of nonrandomized groups. A recent comprehensive empirical analysis of treatments in oncology confirms this finding. The results on replicating clinical trials in observational data are highly mixed, Concato *et al.*[24] concluded that well-designed observational trials closely estimated the effects of treatment when compared with RCTs on the same subject. On the other hand, Soni *et al.*[25]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA; [2]Departments of Bioengineering, Genetics, Medicine, and Biomedical Data Science, Stanford University, Stanford, California, USA; [3]Division of Hematology Oncology, Department of Public Health and Preventive Medicine, Center for Health Care Ethics, Knight Cancer Institute, Oregon Health & Science University, Portland, Oregon, USA. *Correspondence: Brett Beaulieu-Jones (brett_beaulieu-jones@hms.harvard.edu) and Kun-Hsing Yu (Kun-Hsing_Yu@hms.harvard.edu)

**Received August 20, 2019; accepted September 17, 2019. doi:10.1002/cpt.1658**

found a poor correlation between the hazard ratio seen in observational studies vs. randomized trials on the same topic. There is limited evidence that some RCTs may be difficult to replicate.[26–28] This could be due to population and effect sizes, population demographics, or other factors.

RWE may provide valuable insight into the effectiveness and generalizability of interventions in practice, even as RCTs are unlikely to be supplanted as the gold standard for measuring intervention efficacy. It is important to consider the higher level of evidence RCTs provide than RWE when making a regulatory decision.[29] We should aim to use the highest standard possible while acknowledging it is not feasible for RCTs to answer all clinical questions related to drug effectiveness. The reasons where RCT might not be feasible include (i) prohibitive cost,[30,31] (ii) when the standard of care is effective and/or administering a placebo is unethical,[32] and (iii) in rare diseases where patient recruitment is challenging.[33–35] In addition, RCTs are typically performed in a relatively homogenous cohort that is less diverse than the real-world population in terms of age, race, socioeconomic status, geography, clinical setting, disease severity, patient history, and patient willingness to seek treatment.[36–40] Finally, over time, indications for therapeutics often expand to indications and population groups they were not originally tested in. Pragmatic and other modern trial designs may mitigate some of these challenges, for example, by improving generalizability, but fundamental issues of cost, time, and difficult recruitment remain.[41] Pharmaceuticals have been approved without RCTs but should be limited to cases where the potential burden of an incorrect treatment estimation is outweighed by the burden of conducting an RCT.[42] It is critical to utilize alternatives to supplement but not supplant RCTs both in the form of pragmatic trials and RWE-based analyses.

The FDA has taken action in an attempt to reduce the burden in both time and cost of bringing a new therapy to market, through the accelerated regulatory decision regulations initially put in place in 1992 and expanded in 2012.[43] These regulations allow for surrogate and intermediate end points when a therapy addressed a serious condition without existing options. In the 5 years following final guidance from the FDA in May 2014, 71 therapies have been approved through the accelerated pathway.[44] This is in comparison to 25 in the 5 years prior to this guidance. In addition, in 2012, the FDA established the "breakthrough therapy designation" for therapies intended to treat serious or life-threatening conditions where the therapy may demonstrate substantial improvement.[45] From April 2015 to March 2019, 24 drugs have received breakthrough therapy designation. Due to the fact that clinical end points (i.e., outcomes that show direct clinical benefits, such as increased overall survival) may take a long time to develop, many drug trials with breakthrough therapy designation use surrogate end points, which are measurements or signs predictive of clinical outcomes but do not directly measure clinical benefits.[46,47] Examples of surrogate end points include blood pressure for hypertension drugs and serum low-density lipoprotein cholesterol for hypercholesterolemia treatments.[47] Shorter trials and the use of surrogate end points present a strong need for postapproval surveillance for both safety and effectiveness, especially in the context of traditional clinical end points. The use of shorter trials and surrogate end points to accelerate regulatory decisions may suggest that the traditional process is unnecessarily slow and wasteful. However, it is also possible that the extensive use of these shortcuts will lead to suboptimal decisions because there is no guarantee improvement as measured by surrogate end points will translate to traditional end points.

In this light, we ask the question: "What is the role of RWE in the regulatory process?" In this review, we first lay out some of the primary reasons RWE is not suited to replace RCTs. We then examine some areas RWE is well-suited to supplement and enhance the regulatory process as well as providing postapproval guidance.

## LIMITATIONS FOR RWE FOR REGULATORY DECISION MAKING

Here, we examine the limitations of commonly used methods for applying RWD to inform regulatory decision making. We focus on the limitations of RWE to be used in the comparative effectiveness analyses conducted in phase II and phase III RCTs. We, therefore, narrowly examine three types of studies that have been proposed as ways to perform these comparative effectiveness analyses of therapies from RWD: (i) Virtual Comparative Effectiveness Studies ascertain outcomes in both an intervention and control group from an RWD source.[48,49] (ii) Studies using Historical Control Arms compare retrospective RWD-derived controls against an uncontrolled treatment arm.[50] (iii) Studies using Synthetic (Real-World) Control Arms pair an uncontrolled treatment arm with concurrent RWD controls.[22]

Each of the three study designs of RWE has significant issues that prevent the ability to achieve a level of evidence on par with RCTs. Because of challenges in drawing causal conclusions of treatment efficacy from RWE they are not suited to replace RCTs. **Table 1** shows how the specific limitations discussed apply to each form of study design. Most of these mechanisms to use RWE are affected by multiple limitations. Below, we describe these concerns and then link them to each of these three forms of evidence.

### A. Unobserved confounders

Of the sources of RWD, the EHR is generally considered to provide the most granular view of patient care; although insurance claims may provide a higher level of completeness of care. Neither of these data sources is designed for secondary analysis: EHR is primarily intended for patient care, whereas claims data is designed for financial billing and reimbursement. As a result, there are potentially unobserved factors influencing a physician's decision to pursue a particular course of treatment in a systematic manner, preventing the direct comparison of outcomes between treatment arms or direct comparisons to RCT findings. Traditionally, these factors are addressed using randomization because random treatment assignment would not allow for systematic differences between exposed/nonexposed arms. This is not possible using observational data. Several examples of these factors include:

1. Physician opinion. A physician may have just read a paper or attended a seminar recommending a particular treatment, may have seen the treatment work well for a patient that they

**Table 1 Examination of the issues pertinent to each purpose of RWE**

| | Virtual comparative effectiveness studies | Historical control arms | Synthetic control arms |
|---|---|---|---|
| A. Unobserved confounders | X | X | X |
| B. Medicine changes over time | | X | |
| C. Trials may change participant and provider behavior | X | X | X |
| D. Closer monitoring of adverse effects in trials | X | X | X |
| E. Lack of pretrial registration and the potential for multiple testing errors | X | | |
| F. Weaknesses of propensity score matching | X | X | X |
| G. Inability to compare RWD preapproval | X | | |
| H. Opportunities for conflicts of interest to affect results | X | X | X |
| I. Patterns of completeness of data and loss of follow-up differ | X | X | X |
| J. Measurement error in identifying patient status from RWD | X | X | X |

RWD, real-world data; RWE, real-world evidence.

deem similar, or they may simply have a gut feeling that a treatment is right for a patient. Pessimistically, pharmaceutical companies may exert influence regarding the choice of treatment.[51,52] This influence can often be associated with patient characteristics.

2. Patient request. A patient, through his or her own research or through advertisement, may request a specific course of treatment.[53,54] This bias is particularly evident in cases when the patient is attempting to optimize a different outcome from the trial (e.g., shared decision making).[55]
3. Knowledge of a trial. Patients who choose to enroll in a trial are likely to be distributed differently than the general disease population. Similarly, clinicians participating in a trial or prescribing an off-label drug are likely to make different choices regarding patient care. In particular, there is some evidence that physicians are more likely to attempt experimental treatment[37,40] in those who appear healthier. For this reason, unobserved confounders can affect all three considered uses of RWE.
4. Differential access to treatment. The availability or ease of offering a particular treatment may be dependent on administrative, logistical, or insurance coverage-based barriers.[56] These differences are magnified by the integration of multisite and geographically diverse data into observational studies. Consistent annotation or quantification of these factors is not a central theme within RWD datasets, nor are the magnitudes and directions of these effects on physician behavior completely understood.

### B. Medicine changes over time

Historical control arms are by nature historical and as new treatments and technologies, new guidelines, and environmental or socioeconomic changes are introduced medicine changes. This was demonstrated by Sacks *et al.*[50] in 1982 when they showed that 80% (44 of 56) historical control trials found the treatment of interest better than the control, but only 20% of RCTs agreed. For six different clinical areas, they found the results of the trials were more dependent on the method of control groups than on the therapy being considered. In a similar vein, Zia *et al.*[57] identified 43 phase III clinical trials that used identical therapeutic regimens to their corresponding phase II study. Only 28% of the phase III studies were "positive" and 81% had lower effect sizes than their corresponding phase II study. The effect of time trends in medicine is evident even over the course of a single outcome-adaptive trial.[58,59] Outcome-adaptive trials work by adjusting treatment assignment probabilities based on which treatment arm is doing better in order to subject as many participants as possible to the most promising treatment. When a treatment arm seems promising at the beginning of a trial, patients are disproportionately enrolled in the promising arm. This means that the average date of enrollment can be much later in some arms than others.

### C. Trials may change participant and provider behavior—the "Hawthorne Effect"

Clinical trial protocols may result in different behavior than typical clinical practice. Although difficult to measure, there is weak evidence of a protocol or "Hawthorne effect" leading to participants of clinical trials having better outcomes than typical clinical practice.[60] McCarney *et al.*[61] performed a placebo-based randomized trial in dementia, which found that participants receiving more frequent follow-up visits achieved better cognitive and carer-rated quality-of-life outcomes. Similarly, behavioral changes, such as the more frequent follow-up of clinical trial participants, may result in better medication adherence than real-world settings.[62] In traditional RCTs, both arms of the trial may experience the Hawthorne effect. When using historic or synthetic controls only the traditional intervention arm would experience the Hawthorne effect.

### D. Closer monitoring of adverse effects in trials

The combination of more frequent follow-up and specific attention paid to adverse drug events may lead to lower rates of adverse drug events in routine clinical practice compared with clinical trials. In addition, relatively minor adverse effects may not be billed against or recorded in the context of more serious diagnoses (e.g., nausea on a chemotherapy protocol). Several studies have shown that harm in oncology RCTs is underreported and may not even follow protocols.[63] RWE may underestimate important patient safety concerns.

### E. Lack of pretrial registration and multiple testing

It is critical that clinical trials be registered prior to any analysis of results that occur to prevent unreported multiple testing.[64,65] Many statistical methods have been proposed to minimize false discovery in studies involving testing multiple hypotheses simultaneously, and the need for correction is well recognized in biomedical research. However, some forms of multiple testing are more difficult to identify. As an illustration, there are several widely distributed datasets (e.g., Truven MarketScan, Optum Claims data, etc.), and it is likely that multiple investigators will ask similar questions using these datasets. The multiple testing nature of large and uncoordinated efforts may not be apparent to individual investigators involved nor to the scientific community, and the likelihood of false positives would remain uncorrected.[66] To highlight this issue, Silberzahn *et al*.[67] distributed the same dataset to 29 teams with a total of 61 data analysts and asked the question, "Are soccer referees more likely to give red cards to dark-skin-toned players than light skin-toned-players?" Twenty of the 29 teams found a positive effect, the other 9 found no significant relationship and the estimated effect sizes ranged from 0.89–2.93. This study makes explicit a phenomenon that is largely hidden from view, multiple chances and analytical approaches to explore an observational hypothesis may result in different point estimates. Given the financial stakes in regulatory outcomes, there are strong incentives for reporting of positive results. Because explorational analyses may be performed prior to registration or in the absence of registration, multiple hypothesis testing may plague RWE efforts. Preregistration of methods in RCT studies prevents this type of manipulation, both deliberate and inadvertent, and provides a measure of methodological transparency.

### F. Weaknesses of propensity score methods

Propensity score matching and adjustment are popular approaches to account for high-dimensional confounders in observational studies.[68] In propensity score matching, researchers construct a matched population of treated and nontreated individuals based on their probabilities of receiving the treatment. By pairing every treated patient with one or more nontreated patients that were roughly equally likely to have received the treatment, propensity score matching seeks to balance the underlying factors associated with treatment assignment. In propensity score adjustment, the high-dimensional confounders are summarized by a propensity score, which can then be adjusted in downstream analyses.[68,69]

In practice, propensity score methods are difficult to properly execute and evaluate.[70] For instance, it is very difficult to determine if a given propensity score model has been correctly specified. The traditional method for evaluating predictive performance fails to properly evaluate the quality of a propensity score model because unmeasured confounding and the randomness in treatment assignment both contribute to the deviation of the model from the observed data. As such, it can be difficult for a reader to determine whether the model has sufficiently adjusted for confounding.[70]

The deployment of the propensity score is also not straightforward. Because propensity scores from two patients are rarely exactly equal, defining "close enough" propensities for two patients to count as a match involves a delicate balance between excessively loose cutoffs (which risks undercutting the notion of matching itself) and highly stringent cutoffs (which may exclude too many patients from the analysis). The process of propensity matching almost inevitably changes the population being studied in ways difficult to interpret by systematically excluding some subset of patients that fail to achieve a proper match in the other "arm" of the analysis. Ultimately, this means that propensity score matching may provide causal estimates about the effect of the intervention on a different population than the study originally sought to investigate.

### G. Inability to compare RWD preapproval of the experimental treatment

The efficacy of a treatment cannot be evaluated from RWD until it is used widely enough for there to be a substantial body of data. In general, this means that treatment needs to already be approved for the indication of interest. In rare cases, there may be significant off-label usage, but the reasoning behind this off-label usage should be carefully considered in terms of its impacts on patient selection.[71,72] Significant differences between the patients receiving the off-label treatment vs. the existing standard of care could exist, resulting in issues of generalizability. The inability to compare observations preapproval to observations in the general population is especially pertinent to the experimental arm, but there may also be subtle changes in the standard of care that can be difficult to identify from RWD (e.g., dosing, timing, and adherence).[73,74] Finally, off-label prescription with the hope of generating RWD is an inefficient mechanism of hypothesis testing, which is often optimized by formal trials.

### H. Opportunity for conflict of interest

Therapy approval decisions are binary outcomes with the potential for profound impact on patients, employees, shareholders, and other stakeholders. Because of this, there represents an outsized incentive for a variety of parties to cheat the system, regardless of the study design. It was recently revealed that there was "data manipulation" in the data provided to the FDA during the approval process for Zolgensma.[75,76] In RWE, this poses a potentially bigger issue given the retrospective nature of data. The bar for exploitation is lower than prospective studies and exploitation becomes less black and white. It is not possible to ensure that trial organizers have not already analyzed the data to ensure that the control arm is penalized. This could be done through manipulation of the inclusion and exclusion criteria, the method of propensity matching between the trial subjects and

the subjects derived from RWD or other intentional selections. For example, Sacks *et al*.[50] found that historical control groups generally did significantly worse than RCT control groups across 50 reported clinical trials. The retrospective population selection task exhibits the opportunity for exploitation that is difficult to uncover, especially when considering potential financial implications. Although there is the opportunity for bad actors to cheat the system regardless of study type, RWE-based studies can be performed by a smaller set of investigators where there are less exposure and transparency to the protocol.

#### I. Completeness of data and loss of follow-up

Of the three most commonly used sources of RWD, EHR and hospital-based administration both suffer from the fact that they do not record any care received outside of a particular hospital system and oftentimes contain incomplete regard with respect to a particular EHR (e.g., inpatient vs. ambulatory). This leads to challenges in ensuring completeness of care in RWD-based studies using EHR data. In addition, over 50% of the United States receives healthcare insurance through their employment.[77] The average tenure of employment is just 4.3 years for men and 4 years for women.[78] This short tenure means even insurance claims datasets present challenges when considering the completeness of care and follow-up coverage. Record incompleteness, defined as instances when fewer than 50% of enrollees have at least one claim in a given year, can be caused by administrative phenomena, such as company or record mergers, as well as subcontracting of service.[79]

#### J. Measurement error in identifying patient status from RWD

EHRs and administrative data were not initially intended for specific studies, thus, these records may not be sufficiently granular to ascertain the phenotypic status of the patients. Researchers need to make additional assumptions or resort to proxy measures to infer the disease status of the patients under study. Additionally, previous studies showed that different hospitals have dissimilar approaches of disease and procedure coding, even when using the same standard lexicon of diagnostic and procedure codes. Data harmonization efforts and calibration studies are necessary to enhance the accuracy of inferring patient statuses using the limited, and sometimes inconsistent, descriptors in the RWD.

## THE WAY FORWARD: HOW CAN WE CAPTURE VALUE FROM RWE TO EFFECTIVELY IMPROVE PATIENT TREATMENT?

Although RWD analyses are susceptible to the biases and issues summarized above, they hold promise in complementing trial analyses and enable the evaluation of numerous biological hypotheses at a minimal cost. Below, we discuss the approaches that could maximize the value and potential of RWE with particular attention to the approval and postapproval surveillance processes (**Table 2**).

**Table 2 Key needs and opportunities for RWE**

| Key areas | Examples and approaches |
|---|---|
| 1. Measuring post-approval safety and effectiveness | Integrate multiple RWD study designs and leverage modern statistical approaches to better estimate the effects of treatments |
| | Compare differences between effectiveness and efficacy |
| | Establish best practices, guidelines, and reporting standards |
| | Follow accelerated approval and surrogate end-point trials to understand the long-term effects on traditional end points |
| 2. Development of future therapies | Identify populations underserved by current therapies and clinical trials |
| | Discover disease subtypes or potential patient subpopulation that might benefit from novel treatment modalities |
| | Facilitate trial recruitment at diverse clinical sites and the inclusion of diverse populations in future studies |
| 3. Measuring healthcare value and quality | Determine the value-based reimbursement of drugs |
| | Evaluate how closely clinical guidelines are followed and whether guidelines lead to positive outcomes |

RWD, real-world data; RWE, real-world evidence.

#### Integrate results from multiple designs of observational studies to triangulate effect estimates

As discussed in the previous section, RWE generated from different study designs suffers from different sources of biases, and causal inference based on RWD relies on strong assumptions rarely met in practice. Nonetheless, by combining the results generated by different study designs, we can better estimate the risks and benefits of treatment strategies.[80] For example, cohort studies using RWD suffer from unmeasured confounding, whereas the use of instrumental variables relies on instrumental assumptions.[81] Because these two study designs require different sets of assumptions, we can estimate the extent of assumption violation and its impact on the risk estimates. Similarly, we can further incorporate the results from natural experiments[82] and negative control analyses[83] to gauge the effects of treatments and unmeasured confounders. By comparing the results from different approaches, we can determine the possible range of the true estimates with a greater level of confidence. Nonetheless, given the fact that different study designs and analytical methods possess distinct pros and cons, researchers need to be vigilant about the interpretation of their combined results.[80] Effect triangulation approaches have successfully estimated causal effects in settings with strong confounding, such as the effects of lowering systolic blood pressure on the risk of coronary heart disease.[80,84]

#### Leverage new statistical approaches for causal analyses

Philosophers have attempted to understand causality since the age of enlightenment,[85] and epidemiologists proposed several criteria to evaluate the linkage between causes and effects.[86] Recent developments of statistical methods allow causal inference from observational data while minimizing biases insurmountable by conventional approaches.[70] As an illustration, in the presence of time-varying confounders, traditional variable

adjustment approaches will inevitably result in biases, due to the fact that subsequent measurements after the baseline are likely affected by the treatments. The g-methods, a group of statistical approaches, can account for the time-varying confounders and treatment-confounder feedback that commonly reside in observational data.[87–89] In addition, recently developed multiple robust statistical approaches can reduce the risk of model misspecification by relaxing the assumptions needed to achieve unbiased estimates.[90,91] For example, doubly robust methods can consistently estimate the effects of treatments if either the confounder-treatment relation or the confounder/treatment-outcome relation is correctly modeled,[90] which would be helpful in settings where model misspecification raises significant concern. It is worth noting that the causal identification conditions (i.e., exchangeability, positivity, and consistency) still need to hold in order to get accurate effect estimates. In high-dimensional settings, machine-learning approaches can model the high-level interactions among the variables and facilitate dimension reduction.[92] These methods can accommodate the large number of variables extracted from EHRs and other RWDs.

### Establish a robust infrastructure for randomized registry trials

Integration of EHRs into a large-scale data registry would allow for real-time matching against clinical trials and for physicians to be immediately notified if a patient was a potential fit.[93] After patients were enrolled in the trial, the treatment could be randomized and follow-up could occur at their normal point of care.[94] This would have the potential to massively reduce enrollment and follow-up costs while increasing the diversity of populations included in clinical trials. This cost reduction could enable randomized trials to answer a wider scope of questions. In particular, it may allow randomized registry trials to be performed postapproval for comparative effectiveness analysis and additional trials sponsored by government and nonprofit organizations.

### Perform postapproval surveillance and validate that efficacy translates to effectiveness

Although clinical trials measure drug efficacy, it is important to perform postapproval surveillance to determine whether efficacy is generalizable to a broader population. Postapproval surveillance would allow drug pricing to take into account the real-world value delivered. In addition, due to size restrictions, clinical trials may not capture rare adverse effects[95] or drug–drug interactions[96] that could be discovered through RWD analysis.

### Use RWD to measure how closely trial populations resemble real-world populations

This would enable trial organizers to plan representative trials and regulators to determine where additional trials may be necessary. In addition, if there is truly a heterogeneity of effects, it is unlikely to be detected in a homogenous trial population. It is, therefore, important to monitor both effectiveness and safety in the larger heterogeneous population that the therapy is given to.

### Preregistration of RWD analyses

Establishing a preregistration requirement for RWD analyses is an effective approach to reduce the risk of multiple hypotheses testing and p-hacking.[97,98] However, many observational datasets, such as Medicare,[99] Medicaid,[100] and Marketscan,[101] are available and fully accessible to the researchers before the conception of specific RWD studies, which makes adequate preregistration a significant challenge. Combining a preregistration mechanism with a requirement to validate the identified effects on prospective patient cohorts could mitigate the risk of false discovery due to p-hacking.[97]

### Complement RCTs and pragmatic trials with RWD

Although RCTs and pragmatic trials generate high-quality evidence for establishing causality[24,102] and inform real-world practice,[41,103] respectively, it is infeasible to answer all clinically important research questions by setting up a series of trials. In addition to using multiple sources of RWE to arrive at better risk estimates ("Integrate Results From Multiple Designs of Observational Studies to Triangulate Effect Estimates" section), researchers could further leverage the hypotheses generated by RWD to inform trial design, or when the subgroup analyses of trials are underpowered, conduct RWD studies to further identify the participants who would likely benefit from the treatments under study.[104]

### Establish reporting guidelines of RWE

Many reporting guidelines have been established for observational studies, but the specific requirements for using RWE for regulatory approval remains unclear. Widely accepted guidelines for academic publication include the REporting of studies Conducted using Observational Routinely collected health Data (RECORD)[105] and STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) guidelines.[106] However, researchers have called for more stringent and specific approaches for subfields of epidemiology, such as the RECORD-PE statement for pharmacoepidemiological research.[107] Similar efforts are needed to enhance the validity and reproducibility of RWE for regulatory purposes. In particular, regulatory agencies need to set specific requirements for the study participants, variables, measurement methods, data curation procedures, analytical approaches, and accessibility of the data and codes used in demonstrating the effectiveness of treatments using RWD.

### Establish a structure for the postapproval evaluation of clinical practice guidelines and predictive algorithms

Historically, postapproval marketing of drugs involved a wide network of pharmaceutical sales representatives attempting to visit physicians to provide samples and detailing.[51,108] In the 21st century, there has been an increased interest in establishing standardized clinical practice guidelines, for example, Choosing Wisely,[109] and in providing predictive drug recommendations as a part of personalized,[110] and precision medicine,[111] Standardizing care is a core component of improving the process with which care is delivered in the structure, process, and outcome framework for evaluating healthcare quality.[112] Personalized drug recommendations offer the promise to provide patients with the drugs they are most

likely to benefit from. This transition from individualized decision making presents a great potential to improve care and to deliver evidence-based medicine. However, it also presents the potential for postapproval marketing to shift from one-on-one encounters to influence at the system level through guidelines and algorithms. It is critical for RWE-based guidelines and recommendation systems to be thoroughly evaluated by independent third parties in the form of regulatory agencies and physician societies.

### Determine value-based reimbursement of drugs

Value-based drug pricing has been discussed for over a decade.[113–115] RWD analyses can reveal the actual effectiveness of the drugs in real-world use cases, and, hence, inform the value created for the patients.[116] Tracking the longitudinal health outcomes of patients receiving the treatments under the usual circumstances of healthcare practice is crucial for determining the true value of therapies. If it is to succeed, how value is attributed must be carefully regulated and monitored.

## CONCLUSIONS

RWE presents a unique opportunity to accelerate the development of new therapies and to evaluate both the efficacy and the effectiveness of these treatments. However, researchers and regulators should take heed of the limitations of RWE and the potential biases lurking in RWD. Although there is significant value in utilizing RWE preregulatory approval (e.g., identifying subpopulations of need) and postregulatory approval (e.g., safety and surveillance), there are significant barriers to reliably using observational data as a key component of the regulatory process. Conflicting studies on attempts to replicate clinical trials using RWE show the potential risks and brittleness of RWE-based comparative effectiveness.[24,25] This view is further supported by the discrepancies between trial results and those from RWD,[117–120] including a recent failed attempt at Facebook[121] to replicate RCTs with large-scale RWD analyses in a nonmedical context.

The appropriate use of RWE must be driven by forward-thinking best practices, guidelines, and regulations to avoid spurious or biased findings. Increased data availability presents many opportunities but also brings with it the potential for biases in a system with outsized financial incentives. Traditional approaches, including preregistration, may not be sufficient when it is not possible to know which analyses have already been performed. It is critical to acknowledge the history and strengths of traditional RCTs especially in regard to initial approvals. RCTs should not be replaced for approval but can be supplemented to better understand treatment effectiveness in the real world. Prospective follow-up studies driven by unconflicted parties will be critical to decision making around postapproval therapy surveillance and reimbursement. The appropriate use of RWE offers promise to accelerate the development of therapies while making their delivery safer, more targeted, and more efficient in real-world settings.

### FUNDING

This study was funded by the National Library of Medicine (NLM) T15LM007092 (B.K.B.-J.), Harvard Data Science Fellowship (K.-H.Y.), and National Institute of General Medical Sciences (NIGMS) T32GM007753 (S.G.F.).

### CONFLICT OF INTEREST

All other authors declared no competing interests for this work.

### DISCLOSURES

Dr. Prasad reports receiving royalties from his book Ending Medical Reversal; that his work is funded by the Laura and John Arnold Foundation; that he has received honoraria for grand rounds/lectures from several universities, medical centers, and professional societies, and payments for contributions to Medscape; and that he is not compensated for his work at the Veterans Affairs Medical Center in Portland, Oregon, or the Health Technology Assessment Subcommittee of the Oregon Health Authority. Dr. Beaulieu-Jones reports owning equity in Progknowse Inc. outside of the submitted work. Progknowse is a company working with academic and community-based health systems to integrate clinical data and enhance data science capabilities.



1. Jarow, J.P., LaVange, L. & Woodcock, J. Multidimensional evidence generation and FDA regulatory decision making: defining and using 'real-world' data. *JAMA* **318**, 703–704 (2017).
2. Corrigan-Curay, J., Sacks, L. & Woodcock, J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *JAMA* **320**, 867 (2018).
3. US Food and Drug Administration. Real-world evidence <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence> (2019).
4. Cave, A., Kurz, X. & Arlett, P. Real-world data for regulatory decision making: challenges and possible solutions for Europe. *Clin. Pharmacol. Ther.* **106**, 36–39 (2019).
5. 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).
6. Final report on the adaptive pathways pilot <https://www.ema.europa.eu/en/documents/report/final-report-adaptive-pathways-pilot_en.pdf>
7. Herbst, A.L., Ulfelder, H. & Poskanzer, D.C. Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. *N. Engl. J. Med.* **284**, 878–881 (1971).
8. Cooper, W.O. *et al*. Major congenital malformations after first-trimester exposure to ACE inhibitors. *N. Engl. J. Med.* **354**, 2443–2451 (2006).
9. Chambers, C.D. *et al*. Selective serotonin-reuptake inhibitors and risk of persistent pulmonary hypertension of the newborn. *N. Engl. J. Med.* **354**, 579–587 (2006).
10. Kieler, H. *et al*. Selective serotonin reuptake inhibitors during pregnancy and risk of persistent pulmonary hypertension in the newborn: population based cohort study from the five Nordic countries. *BMJ* **344**, d8012 (2012).
11. Lakoski, S.G., Lagace, T.A., Cohen, J.C., Horton, J.D. & Hobbs, H.H. Genetic and metabolic determinants of plasma PCSK9 levels. *J. Clin. Endocrinol. Metab.* **94**, 2537–2543 (2009).
12. Wang, D., Guo, Y., Wrighton, S.A., Cooke, G.E. & Sadee, W. Intronic polymorphism in CYP3A4 affects hepatic expression and response to statin drugs. *Pharmacogenomics J.* **11**, 274–286 (2011).
13. Kivistö, K.T. *et al*. Lipid-lowering response to statins is affected by CYP3A5 polymorphism. *Pharmacogenetics* **14**, 523–525 (2004).
14. Pinto, N. & Dolan, M.E. Clinically relevant genetic variations in drug metabolizing enzymes. *Curr. Drug Metab.* **12**, 487–497 (2011).
15. The FDA and Flatiron Health Expand Real-World Data Cancer Research Collaboration. Flatiron Health. Flatiron health

<https://flatiron.com/press/press-release/the-fda-and-flatiron-health-expand-real-world-data-cancer-research-collaboration/> (2019).
16. Khozin, S. *et al*. Real-world outcomes of patients with metastatic non-small cell lung cancer treated with programmed cell death protein 1 inhibitors in the year following US Regulatory Approval. *Oncologist* **24**, 648–656 (2018).
17. Khozin, S. *et al*. Characteristics of real-world metastatic non-small cell lung cancer patients treated with nivolumab and pembrolizumab during the year following approval. *Oncologist* **23**, 328–336 (2018).
18. Ver Date Sep 11 2014 12:03, M. 130 STAT. 1034 PUBLIC LAW 114–255—DEC. 13, 2016 <https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf>
19. Testimony: Scott Gottlieb on 21st Century Cures Act, 7/25/18. U.S. Food and Drug Administration <https://www.fda.gov/news-events/congressional-testimony/implementing-21st-century-cures-act-2018-update-fda-and-nih-07242018> (2019).
20. Goldsack, J.*et al*.Synthetic control arms are a good option for some clinical trials – STAT. *STAT* <https://www.statnews.com/2019/02/05/synthetic-control-arms-clinical-trials/> (2019).
21. Petrone, J. Roche pays $1.9 billion for Flatiron's army of electronic health record curators. *Nat. Biotechnol.* **36**, 289–290 (2018).
22. Davies, J. *et al*. Comparative effectiveness from a single-arm trial and real-world data: alectinib versus ceritinib. *J. Comp. Eff. Res.* **7**, 855–865 (2018).
23. Booth, C.M., Karim, S. & Mackillop, W.J. Real-world data: towards achieving the achievable in cancer care. *Nat. Rev. Clin. Oncol.* **16**(5), 312–325 (2019).
24. Concato, J., Shah, N. & Horwitz, R.I. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med.* **342**, 1887–1892 (2000).
25. Soni, P.D. *et al*. Comparison of population-based observational studies with randomized trials in oncology. *J. Clin. Oncol*. **37**, 1209–1216 (2019).
26. Whitelaw, A., Placzek, M., Dubowitz, L., Lary, S. & Levene, M. Phenobarbitone for prevention of periventricular haemorrhage in very low birth-weight infants. A randomised double-blind trial. *Lancet* **2**, 1168–1170 (1983).
27. Kuban, K.C. *et al*. Neonatal intracranial hemorrhage and phenobarbital. *Pediatrics* **77**, 443–450 (1986).
28. Silverman, W. Replication of clinical trials. *Lancet* **328**, 348 (1986).
29. Shekelle, P.G., Woolf, S.H., Eccles, M. & Grimshaw, J. Clinical guidelines: developing guidelines. *BMJ* **318**, 593–596 (1999).
30. Martin, L., Hutchens, M., Hawkins, C. & Radnov, A. How much do clinical trials cost? *Nat. Rev. Drug Discov.* **16**, 381–382 (2017).
31. Sertkaya, A., Wong, H.-H., Jessup, A. & Beleche, T. Key cost drivers of pharmaceutical clinical trials in the United States. *Clin. Trials* **13**, 117–126 (2016).
32. Omuro, A. & DeAngelis, L.M. Glioblastoma and other malignant gliomas: a clinical review. *JAMA* **310**, 1842–1850 (2013).
33. Schork, N.J. Personalized medicine: time for one-person trials. *Nature* **520**, 609–611 (2015).
34. Bogaerts, J. *et al*. Clinical trial designs for rare diseases: studies developed and discussed by the International Rare Cancers Initiative. *Eur. J. Cancer* **51**, 271–281 (2015).
35. Yu, K.-H. *et al*. Data-driven analyses revealed the comorbidity landscape of tuberous sclerosis complex. *Neurology* **91**, 974–976 (2018).
36. Knepper, T.C. & McLeod, H.L. When will clinical trials finally reflect diversity? *Nature* **557**, 157–159 (2018).
37. Mishkin, G., Arnaldez, F. & Percy Ivy, S. Drivers of clinical trial participation—demographics, disparities, and eligibility criteria. *JAMA Oncol.* **5**, 305–306 (2019).
38. Murthy, V.H., Krumholz, H.M. & Gross, C.P. Participation in cancer clinical trials: race-, sex-, and age-based disparities. *JAMA* **291**, 2720–2726 (2004).
39. Oh, K., Hu, F.B., Manson, J.E., Stampfer, M.J. & Willett, W.C. Dietary fat intake and risk of coronary heart disease in women: 20 years of follow-up of the nurses' health study. *Am. J. Epidemiol.* **161**, 672–679 (2005).
40. Unger, J.M., Hershman, D.L., Fleury, M.E. & Vaidya, R. Association of patient comorbid conditions with cancer clinical trial participation. *JAMA Oncol*. **5**, 326 (2019).
41. Ford, I. & Norrie, J. Pragmatic trials. *N. Engl. J. Med.* **375**, 454–463 (2016).
42. Hatswell, A.J., Baio, G., Berlin, J.A., Irs, A. & Freemantle, N. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999–2014. *BMJ Open* **6**, e011666 (2016).
43. *US Food and Drug Administration Safety and Innovation Act (FDASIA)* (US FDA, Silver Spring, MD, 2012).
44. US Food and Drug Administration (FDA). CDER Drug and Biologic Accelerated Approvals Based on a Surrogate Endpoint. fda.gov <https://www.fda.gov/media/88907/download> (2019)
45. US Food and Drug Administration (FDA). Number of breakthrough therapy designation requests received. FDA Data Access <https://www.accessdata.fda.gov/scripts/fdatrack/view/track.cfm?program=cber&xml:id=CBER-All-Number-of-Breakthrough-Therapy-Requests-Received-and-Approvals> (2019).
46. Prentice, R.L. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat. Med.* **8**, 431–440 (1989).
47. Center for Drug Evaluation & Research Table of Surrogate Endpoints. US Food and Drug Administration <https://www.fda.gov/drugs/development-resources/table-surrogate-endpoints-were-basis-drug-approval-or-licensure> (2019).
48. Velentgas, P. , Dreyer, N.A. , Nourjah, P. & Smith, S.R. *Protocol for Observational Comparative Effectiveness Research: A User's Guide* (Agency for Healthcare Research and Quality (US), Rockville, MD, 2013).
49. Black, N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* **312**, 1215–1218 (1996).
50. Sacks, H., Chalmers, T.C. & Smith, H. Jr. Randomized versus historical controls for clinical trials. *Am. J. Med.* **72**, 233–240 (1982).
51. Wazana, A. Physicians and the pharmaceutical industry: is a gift ever just a gift? *JAMA* **283**, 373–380 (2000).
52. Spurling, G.K. *et al*. Information from pharmaceutical companies and the quality, quantity, and cost of physicians' prescribing: a systematic review. *PLoS Med.* **7**, e1000352 (2010).
53. Wilkes, M.S., Bell, R.A. & Kravitz, R.L. Direct-to-consumer prescription drug advertising: trends, impact, and implications. *Health Aff.* **19**, 110–128 (2000).
54. Donohue, J.M., Cevasco, M. & Rosenthal, M.B. A decade of direct-to-consumer advertising of prescription drugs. *N. Engl. J. Med.* **357**, 673–681 (2007).
55. Stacey, D. *et al*. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst. Rev.* **4**, CD001431 (2017).
56. Niedzwiecki, M.J., Hsia, R.Y. & Shen, Y.-C. Not all insurance is equal: differential treatment and health outcomes by insurance coverage among nonelderly adult patients with heart attack. *J. Am. Heart Assoc.* **7**, pii: e008152 (2018).
57. Zia, M.I., Siu, L.L., Pond, G.R. & Chen, E.X. Comparison of outcomes of phase II studies and subsequent randomized control studies using identical chemotherapeutic regimens. *J. Clin. Oncol.* **23**, 6982–6991 (2005).
58. Korn, E.L. & Freidlin, B. Outcome–adaptive randomization: is it useful? *J. Clin. Oncol.* **29**, 771–776 (2011).
59. Korn, E.L. & Freidlin, B. Adaptive clinical trials: advantages and disadvantages of various adaptive design elements. *J. Natl. Cancer Inst.* **109**, (2017). https://doi.org/10.1093/jnci/djx013.
60. Braunholtz, D.A., Edwards, S.J.L. & Lilford, R.J. Are randomized clinical trials good for us (in the short term)? Evidence for a 'trial effect'. *J. Clin. Epidemiol.* **54**, 217–224 (2001).
61. McCarney, R. *et al*. The Hawthorne effect: a randomised, controlled trial. *BMC Med. Res. Methodol.* **7**, 30 (2007).
62. Osterberg, L. & Blaschke, T. Adherence to medication. *N. Engl. J. Med.* **353**, 487–497 (2005).
63. Seruga, B. *et al*. Under-reporting of harm in clinical trials. *Lancet Oncol.* **17**, e209–e219 (2016).
64. Green, S.K. & Krautkramer, C.J. The need for a centralized clinical trials registry. *AMA J. Ethics* **6**, 505–508 (2004).

65. Zarin, D.A., Tse, T., Williams, R.J. & Rajakannan, T. Update on trial registration 11 years after the ICMJE policy was established. *N. Engl. J. Med.* **376**, 383–391 (2017).
66. Mandl, K.D. & Manrai, A.K. Potential excessive testing at scale: biomarkers, genomics, and machine learning. *JAMA* **321**, 739–740 (2019).
67. Silberzahn, R. *et al*. Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv. Methods Pract. Psychol. Sci.* **1**, 337–356 (2018).
68. Stuart, E.A. & Rubin, D.B. Best practices in quasi-experimental designs. *Best Pract. Quant. Method* 155–176 (2008).
69. Rosenbaum, P.R. & Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
70. Hernan, M.A. & Robins, J.M. *Causal Inference* (CRC, Boca Raton, FL, 2010).
71. Radley, D.C., Finkelstein, S.N. & Stafford, R.S. Off-label prescribing among office-based physicians. *Arch. Intern. Med.* **166**, 1021–1026 (2006).
72. Conroy, S. *et al*. Survey of unlicensed and off label drug use in paediatric wards in European countries. *BMJ* **320**, 79–82 (2000).
73. Sokol, M.C., McGuigan, K.A., Verbrugge, R.R. & Epstein, R.S. Impact of medication adherence on hospitalization risk and healthcare cost. *Med. Care* **43**, 521–530 (2005).
74. Ho, P.M., Bryson, C.L. & Rumsfeld, J.S. Medication adherence: its importance in cardiovascular outcomes. *Circulation* **119**, 3028–3035 (2009).
75. Bryan, W.W.AveXis Post Approval Office Director Memo STN125694.0. *FDA.gov* <https://www.fda.gov/media/128116/download> (2019).
76. Feuerstein, A., Sheridan, K., Garde, D. & Cooney, E.FDA: Novartis knew of manipulation of data supporting Zolgensma approval. STAT <https://www.statnews.com/2019/08/06/novartis-was-aware-of-manipulation-of-data-supporting-gene-therapy-approval-fda-says/> (2019).
77. Berchick, E.R., Hood, E. & Barnett, J.C.Health Insurance Coverage in the United States: 2017. Current Population Reports, P60-264, US Government Printing Office, Washington, DC <http://www.census.gov/content/dam/Census/library/publications/2018/demo/p60-264.pdf> (2018).
78. Employee Tenure in 2018. Bureau of Labor Statistics <https://www.bls.gov/news.release/tenure.nr0.htm> (2018).
79. Tyree, P.T., Lind, B.K. & Lafferty, W.E. Challenges of using medical insurance claims data for utilization analysis. *Am. J. Med. Qual.* **21**, 269–275 (2006).
80. Lawlor, D.A., Tilling, K. & Davey Smith, G. Triangulation in aetiological epidemiology. *Int. J. Epidemiol.* **45**, 1866–1886 (2016).
81. Greenland, S. An introduction to instrumental variables for epidemiologists. *Int. J. Epidemiol.* **29**, 1102 (2000).
82. Shadish, W.R., Cook, T.D. & Campbell, D.T.Experimental and quasi-experimental designs for generalized causal inference <https://www.alnap.org/system/files/content/resource/files/main/147.pdf> (2002).
83. Lipsitch, M., Tchetgen Tchetgen, E. & Cohen, T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* **21**, 383–388 (2010).
84. Ference, B.A. *et al*. Clinical effect of naturally random allocation to lower systolic blood pressure beginning before the development of hypertension. *Hypertension* **63**, 1182–1188 (2014).
85. Hume, D. *An Enquiry Concerning Human Understanding* (A. Millar, London, 1748).
86. Hill, A.B. The environment and disease: association or causation? *Proc. R. Soc. Med.* **58**, 295–300 (1965).
87. Robins, J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* **7**, 1393–1512 (1986).
88. Robins, J.M., Blevins, D., Ritter, G. & Wulfsohn, M. G-estimation of the effect of prophylaxis therapy for *Pneumocystis carinii* pneumonia on the survival of AIDS patients. *Epidemiology* **3**, 319–336 (1992).
89. Robins, J.M., Hernán, M.A. & Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–560 (2000).
90. Bang, H. & Robins, J.M. Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973 (2005).
91. Vansteelandt, S., Vanderweele, T.J. & Robins, J.M. Multiply robust inference for statistical interactions. *J. Am. Stat. Assoc.* **103**, 1693–1704 (2008).
92. Yu, K.-H., Beam, A.L. & Kohane, I.S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2**, 719–731 (2018).
93. Roberts, K. *et al*. Biomedical informatics advancing the national health agenda: the AMIA 2015 year-in-review in clinical and consumer informatics. *J. Am. Med. Inform. Assoc.* **24**, e185–e190 (2017).
94. Miotto, R. & Weng, C. Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. *J. Am. Med. Inform. Assoc.* **22**, e141–e150 (2015).
95. Yang, S. *et al*. Autoimmune effects of lung cancer immunotherapy revealed by data-driven analysis on a nationwide cohort. *Clin. Pharmacol. Ther.* **107**, 388–396 (2020).
96. Tatonetti, N.P., Ye, P.P., Daneshjou, R. & Altman, R.B. Data-driven prediction of drug effects and interactions. *Sci. Transl. Med.* **4**, 125ra31 (2012).
97. Munafò, M.R. *et al*. A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 21 (2017).
98. Wicherts, J.M. *et al*. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. *Front. Psychol.* **7**, 1832 (2016).
99. Ambs, A. *et al*. Overview of the SEER–Medicare Health Outcomes Survey linked dataset. *Health Care Financ. Rev.* **29**, 5–21 (2008).
100. Bright, R.A., Avorn, J. & Everitt, D.E. Medicaid data as a resource for epidemiologic studies: strengths and limitations. *J. Clin. Epidemiol.* **42**, 937–945 (1989).
101. Adamson, D.M., Chang, S. & Hansen, L.G. Health Research Data for the Real World: The MarketScan Databases (Thompson Healthcare, New York, 2008).
102. Bothwell, L.E. & Podolsky, S.H. The emergence of the randomized, controlled trial. *N. Engl. J. Med.* **375**, 501–504 (2016).
103. Roland, M. & Torgerson, D.J. Understanding controlled trials: what are pragmatic trials? *BMJ* **316**, 285 (1998).
104. Agarwala, V. *et al*. Real-world evidence in support of precision medicine: clinico-genomic cancer data as a case study. *Health Aff.* **37**, 765–772 (2018).
105. Benchimol, E.I. *et al*. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med.* **12**, e1001885 (2015).
106. von Elm, E., et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* **370**, 1453–1457 (2007).
107. Langan, S.M. *et al*. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ* **363**, k3532 (2018).
108. Gagnon, M.-A. & Lexchin, J. The cost of pushing pills: a new estimate of pharmaceutical promotion expenditures in the United States. *PLoS Med.* **5**, e1 (2008).
109. Cassel, C.K. & Guest, J.A. Choosing wisely: helping physicians and patients make smart decisions about their care. *JAMA* **307**, 1801–1802 (2012).
110. Hamburg, M.A. & Collins, F.S. The path to personalized medicine. *N. Engl. J. Med.* **363**, 301–304 (2010).
111. Collins, F.S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **363**, 1–3 (2010).
112. Donabedian, A. Evaluating the quality of medical care. *Milbank Q.* **83**, 691–729 (2005).
113. Claxton, K. *et al*. Value based pricing for NHS drugs: an opportunity not to be missed? *BMJ* **336**, 251–254 (2008).
114. Danzon, P.M. & Taylor, E. Drug pricing and value in oncology. *Oncologist* **15** (suppl. 1), 24–31 (2010).
115. Hwang, T.J., Kesselheim, A.S. & Sarpatwari, A. Value-based pricing and state reform of prescription drug costs. *JAMA* **318**, 609–610 (2017).
116. Haynes, B. Can it work? Does it work? Is it worth it? The testing of healthcare interventions is evolving. *BMJ* **319**, 652–653 (1999).

117. Grodstein, F., Manson, J.E. & Stampfer, M.J. Hormone therapy and coronary heart disease: the role of time since menopause and age at hormone initiation. *J. Womens Health* **15**, 35–44 (2006).
118. Manson, J.E. *et al*. Estrogen plus progestin and the risk of coronary heart disease. *N. Engl. J. Med.* **349**, 523–534 (2003).
119. Hernán, M.A., Robins, J.M. & García Rodríguez, L.A. Discussion on 'Statistical issues arising in the Women's Health Initiative'. *Biometrics* **61**, 922–930 (2005).
120. Hernán, M.A. *et al*. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* **19**, 766–779 (2008).
121. Gordon, B.R., Zettelmeyer, F., Bhargava, N. & Chapsky, D.A comparison of approaches to advertising measurement: evidence from big field experiments at Facebook <https://papers.ssrn.com/sol3/papers.cfm?abstract_xml:id=3162023> (2018).