Use of Real-World Evidence to Drive Drug Development Strategy and Inform Clinical Trial Design

Simon Dagenais[1], Leo Russo[2,*], Ann Madsen[3], Jen Webster[1] and Lauren Becnel[1]

Interest in real-world data (RWD) and real-world evidence (RWE) to expedite and enrich the development of new biopharmaceutical products has proliferated in recent years, spurred by the 21st Century Cures Act in the United States and similar policy efforts in other countries, willingness by regulators to consider RWE in their decisions, demands from third-party payers, and growing concerns about the limitations of traditional clinical trials. Although much of the recent literature on RWE has focused on potential regulatory uses (e.g., product approvals in oncology or rare diseases based on single-arm trials with external control arms), this article reviews how biopharmaceutical companies can leverage RWE to inform internal decisions made throughout the product development process. Specifically, this article will review use of RWD to guide pipeline and portfolio strategy; use of novel sources of RWD to inform product development, use of RWD to inform clinical development, use of advanced analytics to harness "big" RWD, and considerations when using RWD to inform internal decisions. Topics discussed will include the use of molecular, clinicogenomic, medical imaging, radiomic, and patient-derived xenograft data to augment traditional sources of RWE, the use of RWD to inform clinical trial eligibility criteria, enrich trial population based on predicted response, select endpoints, estimate sample size, understand disease progression, and enhance diversity of participants, the growing use of data tokenization and advanced analytical techniques based on artificial intelligence in RWE, as well as the importance of data quality and methodological transparency in RWE.

Real-world data (RWD) are generated by the routine delivery, administration, and reimbursement of healthcare, rather than through controlled experimental settings. Traditional sources of RWD include third-party health insurance claims ("claims"), electronic health records (EHRs), disease or product registries, and patient health surveys.[1] Increasingly, RWD can be linked to more novel types of data, such as genomic (or other "omic") data from biobanks, biopsies and other pathology tests, diagnostic imaging, social determinants of health (SDoH), cancer organoids, as well as patient-derived xenografts (PDXs). Data from qualitative studies, such as surveys, focus groups, and interviews with patients, caregivers, and healthcare providers (HCPs), can also be linked to traditional sources of RWD to provide further insights into preferences, patient-centric endpoints, unmet needs, and help explain why real-world medical practice does not always concord with recommended care. Other types of data, including social media, weather information, and global travel data, can further supplement existing RWD (**Table 1**).

The costs of acquiring these types of RWD can vary substantially, with annual licenses for large, closed network, third-party, private payer claims data in the United States generally costing $100k–300k per therapeutic area (TA), or $400k–$800k for all TAs, structured EHR data costing $1–3 million per TA, and specialty unstructured EHR data requiring extensive curation costing $3–5 million per TA. Some sources of RWD (e.g., Medicare fee for service and specialty disease registries) cannot readily be licensed for direct access by biopharmaceutical companies, requiring that each analysis be completed by an authorized third party under a paid consulting agreement. The costs for such arrangements can range from $50–150k for simple descriptive analyses to $1 million or more for very complex analyses across multiple sources of RWD. Other resources required to analyze RWD include personnel with advanced training in epidemiology, biostatistics, bioinformatics, and related fields, as well data engineers, statistical programmers, data scientists, and information technology resources (e.g., servers, storage, and software) capable of analyzing large amounts of data expeditiously. It is now routine for large biopharmaceutical companies to invest millions of dollars annually to support groups capable of analyzing RWD.

Real-world insights (RWI) generated from any of these data sources can be used to inform internal research or business decision making throughout a product's development lifecycle, including assessing the commercial viability of a program, identifying patient subgroups of interest, understanding temporal trends, and asking better research questions. Researchers in health economics and outcomes research (HEOR) have long used RWD to assess the effectiveness, safety, and value of biopharmaceutical products after they receive regulatory approval. Given that randomized

[1]Real World Evidence, Pfizer Inc, New York, New York, USA; [2]Global Medical Epidemiology, Worldwide Medical and Safety, Pfizer Inc, Collegeville, Pennsylvania, USA; [3]Global Medical Epidemiology, Worldwide Medical and Safety, Pfizer Inc, New York, New York, USA. *Correspondence: Leo Russo (Leo.Russo@pfizer.com)

Received February 26, 2021; accepted October 30, 2021. doi:10.1002/cpt.2480

Table 1 Common sources, types, and examples of real-world data

Source	Type	Subtype	Examples
Administrative	Third-party payer claims	Closed networks	IBM MarketScan, IQVIA PharMetrics, Optum Clinformatics
		Open networks	IQVIA LAAD, DRG RWD, Symphony IDV
		Government	CMS FFS Medicare, Medicaid, VA/DOD
	Hospital chargemaster		Premier, Vizient, IQVIA CDM
	Pharmacy		Surescripts, IQVIA NDTI
Electronic health records	Care setting	Hospitals	Cerner, Epic, Athena
		Clinics	IQVIA AEMR, Optum Panther, IBM Explorys
		Long-term care/Home health	PointClickCare Lighthouse, Optima/Net Health
	Disease	Oncology	Flatiron, Ontada, ConcertAI
		Behavioral health	Kareo, SimplePractice, Valant
		Other	Praxis, TSI Healthcare, Phillips
Patients	Health surveys	Private	Kantar Health NHWS, Gallup National Health
		Public	NHANES, MEPS
	Outcome measures		Kantar Health, Evidation Health
	Multidimensional		PatientsLikeMe, Ciitizen
	Consumer genetic testing		23andMe, Ancestry.com
	Social determinants of health		IQVIA/Experian, MarketScan HPM, Optum SES
	Medical devices		Glooko, Livongo
	Mobile device biometrics	Smartphones	iPhone (HealthKit), Android (Google Fit)
		Smart watches	Apple Watch (HealthKit), Fitbit (Google Fit)
Diagnostics	Laboratory testing	Genetic testing	Invitae, Neogenomics, Ambry Genetics
		Other	Quest, LabCorp
	Clinicogenomics	Oncology	AACR GENIE, Optum Clinicogenomics
	Population genomics		NHGRI 1000 Genomes Project, NIH All of Us
	Diagnostic imaging		Life Image, Ambra Health
Other	Disease registries	Traditional	CorEvitas, Target RWE
		Other	OM1, COTA Healthcare
	Adverse event reports	Regulatory	FDA FAERS, FDA VAERS
		Social media	Twitter, Facebook
	Mortality	Public/Private	CDC WONDER, ObituaryData.com
	Tokenization		HealthVerity, Datavant, Komodo

CDC, Centers for Disease Control and Prevention; FAERS, US Food and Drug Administration Adverse Event Reporting System, FDA, US Food and Drug Administration; NIH, National Institutes of Health; RWE, real-world evidence.

controlled trials (RCTs) have limited generalizability and are insufficiently powered to detect rare adverse events, RWD is also a cornerstone of safety signal detection in pharmacoepidemiology and pharmacovigilance.[2] The US Food and Drug Administration (FDA) uses RWD—including claims and EHR data—extensively in its Sentinel System safety surveillance program.[3]

The 21st Century Cures Act was enacted in December 2016 and required the FDA to propose scenarios under which real-world evidence (RWE)—scientific evidence derived from the rigorous analysis of RWD with appropriate study methodology—can meet regulatory standards for substantial evidence of efficacy and safety and therefore be used to support regulatory decisions. Although key stakeholders, such as biopharmaceutical companies, regulators, payers, health technology assessment agencies, patient advocates, and policy makers, generally agree on the limitations of RCTs (e.g., insufficient follow-up duration, use of surrogate endpoints, and smaller sample sizes), they have yet to reach consensus on the optimal use of RWE.[4–11]

Because much of the recent RWE literature has focused on regulatory aspects (e.g., approval of new products for rare diseases based on single-arm trials with external control arms)[6,11,12], this paper will instead focus on the value of RWE to guide the research and development (R&D) process within biopharmaceutical companies.[13] The following topics will be discussed: (1) use of RWD to guide pipeline and portfolio strategy; (2) use of novel sources of RWD to inform product development; (3) use of RWD to inform clinical development; (4) advanced analytics and tokenization to harness "big" RWD; and (5) considerations when using RWD to inform internal decisions.

USE OF RWD TO GUIDE PIPELINE AND PORTFOLIO STRATEGY

Biopharmaceutical companies must carefully manage their product pipeline, whether through internal research efforts, external agreements such as co-licensing or acquisitions, or partnerships, to invest in the most promising medicines. Historically, RWD was used in the latter part of product development, often once late-phase clinical trials were well under way. Increasingly, companies are leveraging RWD earlier in the development process in ways that are intimately tied to strategy; this use of RWD has only been discussed in a smaller body of literature.

RWD and RWI can be used to develop key portions of Target Product Profiles (TPPs), which help guide internal decisions throughout the product development process. For example, RWD can be analyzed to provide insights into the targeted indication by refining available estimates of disease prevalence and incidence, including temporal trends, to help predict how the total patient population may change over time.[14-16] RWI-enabled approaches are particularly useful for rare diseases, where even small changes in population size can change the viability of a program.

This was demonstrated by researchers interested in an ultrarare oncology indication, neuroendocrine tumors (NETs), who used claims data from the IBM MarketScan and IQVIA PharMetrics databases to update estimates of NET prevalence and incidence.[17] Their analyses confirmed that although NET is very rare, the population size appears to be increasing over time; such findings could be sufficient to tip the balance in favor of developing a therapy for a small but growing market.

RWD can also be used by clinical pharmacology functions to inform the choice of medications for drug-drug interaction (DDI) studies based on frequency of use in the patient population of interest. Information about comorbidities gleaned from RWD can also help assess the potential impact of DDIs based on renal and hepatic impairment and other health conditions. For example, researchers developing therapies for coronavirus disease 2019 (COVID-19) analyzed RWD from 2 EHR databases (University of California, San Francisco Research Data Browser, and Cerner Real World COVID-19 Database) to understand medication use among patients with characteristics similar to the target population of interest.[18] These data were used to identify the most promising candidates to explore further for COVID-19 based on minimizing the clinical impact of potential DDIs predicted from *in vitro* studies.

RWD can also provide insights into potential subgroups of interest within the target indication by helping predict response to therapy or identifying those with the greatest unmet needs according to use of available therapies. Certain sources of RWD (i.e., claims data) can also inform estimates of burden of illness (e.g., healthcare costs, impact on productivity, and mortality), which are key to developing market access, pricing, and HEOR strategies, such as value-based or outcomes-based contracts. At a portfolio level, biopharmaceutical companies can analyze RWD to compare and rank pipeline assets based on forecasted population size, anticipated market share, peak annual revenues, or perceived advantages to existing therapies.

Beyond these insights, the biopharmaceutical industry is also engaging in more streamlined, integrated strategic approaches powered by RWI and RWE. Functions that rely extensively on RWD, such as HEOR and market access, are now being embedded within early clinical development teams to inform the selection of patient-centric endpoints and patient-reported outcomes instruments for pivotal trials based on RWD obtained from registries, EHR data, patient preference studies, or other data sources.[19] By understanding and addressing payer needs early, this interdisciplinary integration can accelerate the development of evidence required to gain postapproval market access and reimbursement.

USE OF NOVEL SOURCES OF RWD TO INFORM PRODUCT DEVELOPMENT

Molecular and clinicogenomic RWD

A variety of factors have contributed to the surging use of "omic" data in recent years, including: (1) decreased costs of whole genome, whole transcriptome, and gene panel tests; (2) increased insurance coverage for these tests in certain populations (e.g., oncology and noninvasive neonatal tests)[20-22]; (3) uptake of these tests in routine clinical practice; (4) availability of "deep" molecular data to identify new druggable targets or reposition existing medicines for new indications[23]; (5) improved understanding of mechanisms of resistance to anti-infectives[24] or anti-cancer therapies[24,25]; (6) advances in identifying synergistic therapies[26]; (7) use of biomarkers to customize care for patients (e.g., so-called "hyperresponders" or patients whose pharmacogenomic profiles correlate with a higher probability of experiencing specific adverse events)[27,28]; and (8) ability to link genomics datasets with EHR data to better understand the frequency and impact of rare co-occurring genomic events (e.g., *NTRK* gene fusions). Together, these advances have led to a variety of discoveries.

For example, these approaches were used early in the COVID-19 pandemic to identify common molecular players in host-coronavirus interactions from recent outbreaks, including COVID-19 and Middle East Respiratory Syndrome-CoV.[29] These findings were then linked to RWD from claims and genetic RWD to identify two current medications that could be repurposed to lessen viral replication and lead to improved patient outcomes. At the 2021 American Society for Clinical Oncology (ASCO) meeting, investigators compared > 325,000 tumor samples with > 28,000 matched plasma samples to detect kinase fusions, with high positive agreement between fusion events in tumors and novel liquid biopsy samples.[30]

Despite these advances, notable limitations exist in the widespread use of "omic" data. First, data from whole genome and transcriptome testing—unlike data from claims or EHR—cannot be de-identified, so extreme care must be taken to understand patient consent to use these data, how they are stored, how access is managed, and how to protect privacy according to regulations such as General Data Protection Regulation (GDPR)[31] while empowering patient autonomy to participate in research.[32]

Second, many sources of RWD require expert curation of unstructured data from EHRs or "omic" data to identify key variables or events of interest. Because many "omic" test results come in proprietary data formats, automated approaches to extraction may not be possible, limiting their utility to researchers and raising concerns about data quality.[33] Initiatives such as the American Association for Cancer Research (AACR) Genomics Evidence Neoplasia Information Exchange (GENIE) in the US and UK Biobank, among others, seek to address this gap by linking clinicogenomic data to other sources of RWD, including curated clinical data.

Third, the costs of acquiring and/or analyzing "omic" data linked to other sources of RWD can be substantial (i.e., tens of millions of dollars per year to license specialty datasets, or billions of dollars to acquire RWD providers, such as Flatiron Health and Foundation Medicine), and the return on investment is difficult to determine. For rare diseases, the costs of RWD can be astronomical on a per patient basis.

Fourth, it may not be possible to obtain direct access to certain datasets, requiring biopharmaceutical companies to submit analytical queries that are executed by the data owners. Such arrangements can be costly and time-consuming, may raise concerns about confidentiality, and can create uncertainty about the permitted uses of analytical reports (e.g., publications, engagements with payers, and submissions to regulators).

Medical imaging RWD, radiomics, and novel disease models linked to RWD

The availability of medical imaging in RWD has also increased in recent years, facilitated greatly by the development of digital image analysis to increase the accuracy of diagnostics[34] and conduct passive screening on large databases of medical images using machine-learning (ML) algorithms.[35] Radiomics involve advanced algorithms to better detect lesions or other events within medical images that are associated with biological or clinical endpoints (e.g., molecular biomarkers).[36,37] Algorithms can also help identify additional diagnostic tests of value from medical images with pathology.

Across a diverse set of imaging modalities, digital images typically include metadata and/or annotations that may include protected health information (PHI) (e.g., patient name, date of birth). Although diagnostic images generally do not warrant the same level of privacy concerns as genomic data, researchers must also remove facial characteristics or other features that could identify a patient.[38] There are now standard medical imaging datasets that can be leveraged to evaluate the performance of both existing and novel de-identification algorithms.[39]

Digital image analysis can be used to support R&D by analyzing large volumes of tissue specimens or other medical images to run molecular screens on PDXs that model biomarkers and treatment responses by transplanting a portion of a patient's tumor into humanized mice[40] or 3D tissue cultures derived from stem cells that resemble miniature organs.[41-43] These models allow researchers to conduct controlled laboratory experiments that can inform treatment approaches and link predicted treatment response to actual clinical outcomes by linking this data to EHR, claims, and other sources of RWD. Similarly, preclinical studies can be informed by safety assessments conducted in animal models or studies of animal molecular biomarkers or anatomic abnormalities to minimize the burden on human study participants.[44] Findings can also inform clinical trial optimization by stratifying participants according to predicted response and determining appropriate eligibility criteria.[44,45]

PDXs can also further our understanding of signal transduction mechanisms, acquired tumor resistance, and identify potential combinatorial therapies to overcome this resistance. For example, it was found that PDX models of ovarian cancer resistant to platinum chemotherapy and poly (ADP-ribose) polymerase (PARP) inhibitors (PARPi) had increased activity for ataxia telangiectasia and Rad3-related kinase (ATR)-CHK1. Based on these findings, investigators reported a significant increase in survival when using a combination of PARP and ATR inhibitors (ATRi) to treat patients with this type of cancer.[30]

Insights from PDX models linked to other sources of RWD can also influence the design of studies that explore the potential benefits and risks of combination therapies. For example, a nonrandomized study examined 28 patients with high grade serous ovarian cancer and homologous recombination deficient mutations who had progressed on PARPi monotherapy.[46] Researchers tested the hypothesis that ATRi agents may help overcome resistance to PARPi and reported that the combination of Olaparib (PARPi) and ceralasertib (ATRi) had a synergistic clinical effect.[46] Taken together, these studies suggest that RWD can help R&D leaders make informed business decisions on the rationale and design for larger scale clinical trials.

USE OF RWD TO INFORM CLINICAL DEVELOPMENT

Informing trial design

Changing regulation, policy, and healthcare delivery have created increased pressure for R&D companies to deliver assets faster, often for competitive indications that have multiple existing therapies.[10] While RWE often cannot supplant "gold standard" RCTs,[47] it can nonetheless provide insights needed to streamline RCTs to reduce their duration and costs.[48-50] Although clinical trials are complex and can fail for many reasons, uncertainty at the design stage (e.g., optimal sample size, endpoints, follow-up) is an important contributing factor. Analyses of RWD can reduce some of this uncertainty and help inform the design of more efficient clinical trials by better informing global target enrollment sizes needed, selecting more productive trial sites, enriching trial populations with predicted responders, and increasing the diversity of trial participants.[51] The benefits of these newer use cases for RWE are mainly supported by myriad industry white papers, but this may change within the next 1–2 years as researchers share their experiences. The various uses of RWE to provide insights needed for clinical development are summarized in **Table 2**.

A recent study proposed an approach that uses all available data—including from RCTs and RWD—on the efficacy of therapies for a disease of interest to inform the design of future clinical trials.[51] By including estimates of efficacy derived from RWE in a network meta-analysis, researchers estimated that the required sample size in future clinical trials would decrease by at least 40% compared with estimates derived only from RCTs. Such reductions

Table 2 Uses of RWE to provide insights needed for clinical development

Theme	Understanding patient population	Understanding health care utilization	Understanding disease
Components	• Prevalence • Incidence • Population size • Comorbidities • Temporal trends • Diagnostic journey	• Quantity/quality of health care • Standard of care • Unmet needs • Clinical trial sites • Adherence/persistence	• Natural history • Disease progression • Disease segmentation • Endpoints • Sample size
Potential uses	• Viability of: ○ Clinical development regulatory pathway ○ Commercialization	• Developing value proposition • Benchmarking against competitors • Identifying health care disparities	• Trial feasibility • Trial modeling • Trial design • Generating hypotheses • Effect size

RWE, real-world evidence.

in sample size were estimated to yield time savings of at least 6 months in the conduct of clinical trials, which could represent millions of dollars in saved costs for trial execution, or potentially hundreds of millions of dollars in product revenue. While such examples are enticing, they currently represent isolated cases and areas of opportunity for RWE in the future, rather than being standard business practice.

Understanding a disease often begins with studying its natural history, which is a conceptual framework to illustrate how an individual may progress through different stages of a disease (e.g., normal/healthy, preclinical manifestation, clinical onset, mild/moderate/severe clinical presentation, partial/complete resolution, and death) in the absence of any healthcare intervention.[52] Elements required to understand natural history include the number of relevant disease stages, how the disease manifests (e.g., signs and symptoms) in each stage, the amount of time spent in each stage, risk factors for speed of progression, reversibility of progression, and heterogeneity within and between stages. For example, a study examined the natural history study of amyotrophic lateral sclerosis (ALS) using retrospective RWD on 175 patients with the A4V SOD1 genotype for ALS at 15 medical centers in North America.[53] The study reported that patients within this subgroup of SOD1 ALS had a clinically homogenous natural history, with a median survival of 1.2 years. Data from this study were also used to inform future sample size calculations. Findings suggested that a study with only 52 participants per group and a 2-year follow-up would be sufficiently powered to detect a clinically meaningful difference in survival (hazard ratio 0.5) among patients with A4V SOD1 ALS vs. 88 participants per group when including all patients (i.e., A4V and non-A4V) with SOD1 ALS.

Publicly available incidence/prevalence data are an important component of a TPP. Rather than relying solely upon published studies, which may not include specific information about a target indication's subpopulation rather than the general population of all patients who have a condition, TPPs can be supplemented with proprietary data from internal trials, natural history studies, and patient journey RWD for that specific subpopulation. In addition, mid- and late-phase trial designs may be informed by the standard of care (SOC) illustrated within RWD, including elements such as frequency of visits, selection and timing of diagnostic laboratory tests, and selection of endpoints. However, RWD may not always be informartive for this purpose. For rare conditions or for indications in which patient identification requires deep curation on a high volume of patients (e.g., non-muscle invasive bladder cancer (NMIBC), which has no clear set of diagnostic codes to unambiguously identify the correct patients), limited RWD exist, so significant investment ($M USD) may be required, and return on investment must be carefully considered. Practical factors can also limit the impact of RWD, such as trials for solid tumors using Response Evaluation Criteria in Solid Tumors (RECIST) criteria which are rarely used in routine clinical care and therefore not available in RWD, or dermatology trials with provider assessments of efficacy requiring a much higher number of patient/provider interactions than occur in a real-world setting.

Another source of RWI that can be utilized to understand patients when defining a TPP are treatment patterns. Real-world medical use of therapies can differ significantly from guidelines for a number of reasons, so a deeper understanding of what the real SOC is within a given country or geographic region, type of care delivery site (i.e., community hospital or clinic, integrated delivery system, academic medical center, or other site of care), and subcohort of patients (e.g., by line of therapy, patient age, or race) is critical. This seemingly simple concept can be complex in practice. The framework we use to define SOC in terms of health care quantity, type, sequencing, quality, concordance with evidence-based guidelines, persistence, adherence, location, combinations, switching patterns, and other attributes, can teach us to categorize seemingly heterogenous patients into clinically meaningful and informative subgroups. Understanding and targeting the appropriate subgroups that are most likely to benefit from an innovative medicine or other intervention can be critical for the strategic viability of an early research indication or candidate asset, as well as for downstream trial success. Companies who license multiple sources of RWD can readily derive treatment patterns and SOC for many conditions. Determining line of therapy, drug holidays, cessation, and other clinical concepts requires a nuanced understanding of how care is delivered and may require additional data, such as curated EHR information on dates of progression for progression-based line of therapy definitions rather than treatment switch-based definitions.

For example, a study analyzed claims from the IQVIA PharMetrics database linked to the Modernizing Medicine EHR database to examine treatment patterns for patients with psoriasis who received biologic therapies.[54] It reported that 24.8% of patients received combination therapy, of whom 12.2% switched therapy and 24.4% discontinued therapy, suggesting treatment failure, toxicity, or other event. By understanding and characterizing the patients who had sustained therapies, life science companies are better able to target specific subgroups of patients with psoriasis.

Selection of endpoints and identification of surrogate endpoints

Clinical trials often consider regulatory guidance, regulatory precedents, competitive landscape, scientific literature, and subject matter experts to select endpoints, including biomarkers, patient-reported outcomes, and other clinical outcome assessments. However, the endpoints favored by one stakeholder (e.g., regulatory agencies) may not always be meaningful to other stakeholders (e.g., patients, payers, and HCPs).[10] This misalignment is particularly challenging with biopharmaceutical products that receive accelerated regulatory approval based on changes in surrogate endpoints (e.g., biomarkers). Additional studies are then required to confirm the clinical benefits of these changes; such "confirmatory" studies increasingly involve novel clinical trial designs (e.g., pragmatic trials) that incorporate RWE.[55,56] Analyses of RWD can be used to assess the feasibility of pragmatic RWE trials before they are discussed with regulators.[57]

For example, a systematic literature review analyzed published data from multiple sources (e.g., clinical trials and RWE from registries) to understand how hemoglobin concentration (a biomarker) is related to various clinical endpoints, such as stroke, cerebrovascular disease, kidney disease, pulmonary vasculopathy, and mortality in patients with sickle cell disease.[58] Based on a meta-analysis of these data, researchers concluded that changes in hemoglobin concentration are a validated intermediary measure of disease progression in patients with sickle cell disease.

Changes observed in clinical presentation (e.g., signs and symptoms) at different stages of disease progression can provide insights about clinical endpoints to measure in clinical trials, as well as the magnitude of changes required in those endpoints to determine if a patient is deviating meaningfully from their expected prognosis without intervention. Studying disease progression can also lead to identifying relevant biomarkers, which can be used to predict treatment response at different stages of the disease, or as surrogate endpoints when other measures are unavailable or difficult to interpret.[52]

For example, a study examined the natural history of prostate cancer to identify the stage at which patients may benefit most from treatment.[59] Researchers developed a disease progression model to predict how patients would flow through eight stages of prostate cancer based on tumor status, metastasis, therapy, and a biomarker (serum testosterone), using data from clinical trials, literature, and RWD to create this model. The model concluded that new therapies aimed at slowing the progression from stage 6 (non-metastatic castration-resistant prostate cancer) to stages 7 and 8 would have the greatest potential benefit on morbidity and mortality in patients with prostate cancer. Such findings could be used to design clinical trials targeting a patient subgroup with the greatest unmet need through the identification of surrogate markers of progression.

Further, RWD, including EHRs and registries, can be a rich source of information for these insights, particularly in rare diseases where biopharmaceutical companies can struggle to recruit large study populations. For example, a study analyzed RWD from registries focused on Huntington's disease (HD) to better understand disease progression in HD by comparing 3 subgroups: (1) pre-symptomatic HD (gene carriers); (2) early HD; and (3) healthy controls.[60] Researchers examined changes in brain imaging, cognitive function, quantitative motor evaluation, oculomotor evaluation, and neuropsychological assessments over 24 months to understand how each measure changed over time, and how these changes correlated with overall disease progression. The study concluded that for a 2-year clinical trial, findings in serial brain imaging would likely be more sensitive to changes in clinical presentation than measures related to quality of life, which showed minimal change over this period. Such insights can be very informative when selecting endpoints for a proposed clinical trial.

Optimizing clinical trial execution

Beyond informing TPPs, RWI can also lead to actionable insights to help plan and conduct clinical trials, which are one of the most challenging aspects of an R&D program, prone to delays, high costs, and failures.[61] Even with robust trial designs, site selection is another source of uncertainty. It is costly to activate trial sites, particularly when some may accrue no or very few patients compared with their enrollment targets, resulting in potential expenditures of millions of dollars.

Historically, industry utilized some RWD sources that identify US HCPs using a National Provider Identifier (NPI). Analyses of RWD can identify HCPs who have recently provided care to patients with a disease of interest and—based on claims or EHR data—appear to meet eligibility criteria for a proposed trial, which can inform clinical trial site selection. This approach has key limitations, including masked or missing data that are removed to protect patient confidentiality when NPIs or locations are made available and lack of data elements, such as laboratory values that are routinely part of trial eligibility criteria. More recently, RWD aggregation tools and trial optimization platforms, such as TriNetX, are being used to identify participating global healthcare sites that provide care to patients who may be eligible for a trial. Some, although not the majority, of participating sites have elected to make laboratory and some biomarker data available to more accurately identify sites with eligible patients. Raw overall patient counts are available, as are more sophisticated assessments of the number of potential patients who have been diagnosed or treated over a recent time interval (e.g., 6 months) to identify sites most likely to rapidly enroll patients. As with any tool, there are limitations, including low availability of some types of RWD (e.g., biomarkers), lack of robust patient diversity and SDoH data, and, for some indications, difficultly in accurately defining a cohort of interest from available diagnostic or procedure codes, alone (e.g., NMIBC as described above).

Enhancing diversity in trials

Another dimension of RWE's value for identifying the appropriate patients for a clinical trial is demographic diversity (e.g., race and ethnicity) and inclusion of historically under-represented groups. It is well-documented that clinical trials tend to include more older, White male patients than the general population of patients with a given condition. Pfizer recently published an assessment of diversity for its interventional clinical trials from 2011 through 2020.[62] This study compared age, sex, race, and ethnicity for participants in Pfizer-sponsored clinical trials to demographic data from the US Census. Authors concluded that Pfizer clinical trials had similar proportions of Black or African American individuals and females as the broader US population, but a lower proportion of Hispanic or Latino participants than expected. The study also estimated the proportion of clinical trials that met or exceeded targets for representation of different population subgroups based on national demographic data. Authors concluded that whereas approximately half (51.4% to 56.1%) of trials had adequate representation of Black or African American, Hispanic or Latino, and White individuals, these figures were much lower (8.5–16.0%) for the representation of American Indian or Alaska Native, Asian, and Native Hawaiian or Pacific Islander populations. There is more progress to be made for representative and inclusive trials, and RWD is one of the tools available to achieve this.

RWE can also help identify potential disparities in healthcare utilization or outcomes related to patient race, ethnicity, gender, age, comorbidities, payer type, employment status, or other socioeconomic factors, including those that are SDoH. Social determinants of health are the conditions, systems, and circumstances in which people are born, live, work, and age that relate to conditions of daily life, such as structural racism, level of educational attainment and quality, language proficiency, socioeconomic status, employment status, income level and inter-generational wealth, physical environment (e.g., localized level of public safety or environmental exposures to toxins), housing status, access to personal or public transportation, diet and access to healthful foods, social support networks, access to health care and implicit biases in some healthcare providers.[63,64] Recognizing such disparities is one of the first steps in correcting them, a goal shared by many biopharmaceutical companies in clinical development. Efforts can then be made with RWD to target and enroll participants in clinical trials who may otherwise be overlooked, limiting the generalizability of findings and perpetuating disparities based on availability of supporting evidence.

Although biopharmaceutical companies do not intend to exclude patients from participating in sponsored clinical trials on the basis of race or ethnicity, the application of trial eligibility criteria may have unequal or unintended effects in different populations. For example, a study in 2017 applied eligibility criteria from phase III clinical trials for multiple myeloma and estimated that 40% of patients—52.7% of African American patients—in a disease registry for multiple myeloma would not qualify for any clinical trials for various reasons (e.g., disease staging and results of laboratory tests).[65] Companies confronted with such findings should re-evaluate their study design to ensure that specific eligibility criteria are not disproportionally excluding patient subgroups. In light of these and other similar findings, clinical trial design teams now regularly use RWD to evaluate the impact of proposed inclusion exclusion criteria as part of the study design process.

Key limitations still exist in the use of RWD and SDoH to increase trial diversity. EHR sources typically lack accurate and complete race, ethnicity, and other non-medical information required to better understand under-represented patients and their disease. Most SDoH factors are not routinely captured in clinical research, although some surrogates, such as insurance status, are routinely used. While RWD can help identify disparities, use of RWD alone is not yet sufficient to meet diversity goals for clinical trials. Real-world understanding of patient behaviors, decision making, and motives are needed to design more successful solutions.

Examples of studies from the literature based on RWD to inform product development strategy and clinical trial design are summarized in **Table 3**.

USE OF TOKENIZATION AND ADVANCED ANALYTICS TO HARNESS RWD

Data tokenization

It is now possible to use tokens to link different sources of patient-level RWD (e.g., claims, EHR, registries, clinical trials, "omic" data, molecular biomarkers, laboratory, and SDoH) to provide a more comprehensive understanding of health and health care. Patient "tokens" are unique identifiers created by companies, such as HealthVerity Inc. (Philadelphia, PA) and Datavant (San Francisco, CA) to recognize a patient who appears across multiple sources of RWD. Tokens do not contain PHI (e.g., date of birth and social security number), are not derived from PHI, and are intended to protect against reidentification of patients. Tokenization vendors can either serve as a matchmaker to link patient data across datasets available to biopharmaceutical companies or provide access to a centralized marketplace with multiple databases that are already linked. For example, research questions at the onset of COVID-19 far outpaced insights that could be drawn from single data sources (e.g., total number of new cases requiring hospitalization). In response, HealthVerity used tokenization to link multiple sources of RWD into a cohesive dataset that enabled biopharmaceutical companies to better understand medication use, hospital-based mechanical therapies, disease progression, and re-infection.[71,74,75,76]

Although this technology is gaining traction, it is currently available almost exclusively in the US. Questions remain about the transparency of linking methods (e.g., automatic vs. manual), potential for inaccurate matching (e.g., that may link one patient's laboratory data to a second patient's clinical information), and assurances about maintaining patient confidentiality and consent with linked RWD.[77] Another limitation is the cost for tokenizing large datasets at scale, which can put this technology out of reach for some pre-registrational programs that may lack sufficient funding. Evolution in regulation and policy may impact tokenization opportunities moving forward. In its most recent draft guidance,[92] the FDA identified key considerations regarding the potential suitability of tokenization in RWE.

Advanced analytics

The increasing availability of biomedical big data within biopharmaceutical companies has stimulated the development of advanced analytics (e.g., semi-automated biomedical curation

Table 3 Case examples of RWE for drug development strategy and clinical trial design

Use	Citation	Study Objective	Data Source(s)	Insight
Understanding patient populations	Broder *et al*. (2018)[17]	Estimate prevalence and incidence of neuroendocrine tumors	IBM MarketScan and IQVIA PharMetrics claims databases	Prevalence and incidence increasing over time.
	Dellon *et al*. (2014)[66]	Estimate prevalence of EE	IQVIA PharMetrics claims	Updated estimates for number of patients with EE in the United States following the introduction of a new ICD-9 diagnosis code specific to EE.
	Wallin *et al*. (2019)[16]	Estimate national prevalence for MS by analyzing multiple US databases, covering different population segments.	Optum, IBM, Kaiser Permanente, Department of Veterans Affairs, and the Centers for Medicare and Medicaid claims databases	The 3-year prevalence of MS was 309.2 per 100,000, with an estimated 727,344 cases in the United States, higher than previous studies.
	Halpern *et al*. (2019)[67]	Estimate prevalence of agitation among patients with AD	Optum EHR database	Prevalence of agitation over a 2-year period was 44.6%. NLP was used to analyze unstructured data for keywords related to agitation.
	Chehade *et al*. (2021)[68]	Describe patient journey for individuals with EG/EoD	Symphony Health Patient Source claims database	Many EG/EoD patients initially diagnosed with irritable bowel syndrome or dyspepsia, highlighting the need for improved diagnosis.
	Morgan *et al*. (2021)[69]	Describe diagnostic journey of patients with PSP	Patient interviews and physician chart reviews in France, Germany, Italy, Spain, the United Kingdom, and the United States	Diagnostic delays may be related to patients first presenting to primary care providers before being evaluated by movement disorder specialists.
Understanding treatment patterns	Zhu *et al*. (2019)[70]	Characterize current treatment patterns for AA in China	Disease Registry in China	Only 1 in 5 AA patients were receiving first-line care concordant with evidence-based guidelines
	Stewart *et al*. (2021)[71]	COVID-19: understand medication use, hospital-based mechanical therapies, disease progression, and re-infection	HealthVerity used tokenization to link multiple data sources	Use of hydroxychloroquine with or without azithromycin among hospitalized patients with COVID-19 was described.
	Murage *et al*. (2019)[54]	Examine treatment patterns for patients with psoriasis receiving biologic therapies.	IQVIA PharMetrics database linked to the Modernizing Medicine EHR database	Results on combination therapy, switching, adherence, and discontinuation are valuable for biopharmaceutical companies developing therapies targeting specific patient subgroups (i.e., treatment failures)
	Shah *et al*. (2017)[65]	Applied eligibility criteria from phase III clinical trials for MM to assess the proportion of patients being excluded from trials.	Disease Registry	Estimated that 40% of MM patients – 52.7% of African American patients – would not qualify for any clinical trials
	Foerster *et al*. (2021)[72]	Describe the diagnostic journey for women with breast cancer in Sub-Saharan Africa	Prospective Cohort Study	White patients in Nigeria had a median diagnostic journey of only 2.4 months, compared with 11.3 months for patients in Uganda.
	Bakouny *et al*. (2021)[73]	Effect of COVID-19 pandemic on cancer screening and diagnosis	EHRs from one integrated delivery network	Cancer screening procedures decreased 60%-82% from 2019 to 2020. New cancer diagnoses decreased 19%–78%.
Understanding diseases	Bali *et al*. (2017)[53]	Natural history study of ALS with A4V SOD1 genotype	EHRS from 15 North American medical centers	Genotype is adequately defined and understood to study in clinical trials. Data on disease course used to inform future trial sample size calculations.

(Continued)

Table 3 (Continued)

Use	Citation	Study Objective	Data Source(s)	Insight
	Scher *et al.* (2015)[59]	Build a dynamic progression model for prostate cancer	NCI-SEER	Findings could be used to design clinical trials targeting a patient subgroup with the greatest unmet need.
	Tabrizi *et al.* (2012)[60]	Understand disease progression in HD	Disease Registries	Endpoint selection for future trials should use serial brain imaging rather than measures related to quality of life. Former was more sensitive to changes in clinical presentation.
	Ataga *et al.* (2020)[58]	Understand how hemoglobin concentration is related to stroke, cerebrovascular disease, kidney disease, pulmonary vasculopathy, and mortality in patients with SCD	Meta-Analysis including disease registries	Changes in hemoglobin concentration is a validated intermediary measure of disease progression in patients with SCD.

AA, Aplastic Anemia; AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; COVID-19, coronavirus disease 2019; EE, eosinophilic esophagitis; EG/EoD, eosinophilic gastritis or duodenitis; EHR, electronic health record; HD, Huntington's disease; ICD-9, International Classification of Disease 9th revision; MM, multiple myeloma; MS, multiple sclerosis; NLP, natural language processing; PSP, progressive supranuclear palsy; RWE, real-world evidence; SCD, sickle cell disease; SEER, Surveillance, Epidemiology and End Results.

pipelines, ML, and deep learning) to harness these data.[78,79] For example, R&D teams have traditionally identified and summarized scientific literature manually, which is increasingly difficult in fields with a rapid growth in the quantity of new studies published. In response, tools using natural language processing have been developed to identify and summarize abstracts[80] and even identify genetic targets of high interest based on publication frequency.[81,82] Other tools leverage human curation and allow researchers to execute queries across multiple data sources (e.g., QIAGEN Ingenuity Pathway Analysis). In one study, ML methods to predict a variety of medical events were successfully applied to EHR data from different health care centers without having to first harmonize data across sites.[83]

ML can also be applied to analyze data from wearable devices to detect abnormalities in gait and heart rate[84] and even predict seizures in patients whose disease is poorly managed, among other applications.[85] Advanced algorithms are also being applied to RWD to better inform clinical trial design and increase generalizability. Using the Trial Pathfinder computational framework to analyze EHR data from 61,000 patients with non-small cell lung cancer, researchers determined that commonly used trial eligibility criteria often excluded patients who may have benefited from study interventions.[86] Nevertheless, the use of advanced analytics in RWD faces a number of major hurdles, including limitations in data quality, accompanying metadata, data access and data sharing, as well as the supply of skilled data scientists.[87] Studies have also cautioned that potential gains in the speed or costs of preclinical research facilitated by artificial intelligence could be dwarfed by subsequent failure in clinical trials or drugs with unexpected toxicity.[88]

CONSIDERATIONS WHEN USING RWD TO INFORM INTERNAL DECISIONS

Ensuring high quality data, study design, and analysis

Although RWD can be analyzed to provide rich insights that inform internal decisions at biopharmaceutical companies throughout the drug development process, it must be remembered that the validity of these insights is highly dependent on the quality of the underlying data, and enthusiasm about RWD should be tempered by awareness of its limitations. Because RWD is not collected for research purposes, it will not readily conform to the quality standards expected of data collected in prospective RCTs conducted according to Good Clinical Practices; expectations should be set accordingly. Nevertheless, the quality of RWD must be sufficient to support the insights drawn from it. Because this concept is still evolving, it may be helpful to review how regulators are approaching the notion of RWD quality, even when RWE is intended for internal decision making.

For example, the first guidance issued by the FDA on RWD/RWE after the 21st Century Cures Act stated that the suitability of RWD should be based on its: (1) relevance and (2) reliability.[89] Whereas the former is concerned with determining if the RWD directly addresses its intended use (e.g., contains data elements for specific endpoints of interest in a representative patient population for the target indication), the latter is concerned with more traditional aspects of data quality (e.g., completeness, timeliness, and accuracy). The 2018 framework on RWE from the FDA reiterated the importance of determining the suitability of RWD to its proposed regulatory use, which involves assessing its: (1) relevance ("fit for purpose" or "fit for use"); (2) reliability, and (3) use of appropriate statistical and research methods to analyze RWD.[1] However, the framework did not specify how to conduct these assessments and referred to future guidance that will be issued on these topics. The FDA also cautions that it does not currently endorse any specific type or source of RWD and that any assessment of RWD quality should be tailored to a specific intended regulatory use.

Insights into how the FDA may approach "regulatory-grade" RWE in the future might be gleaned from opinions expressed by current and former FDA employees. For instance, presentations by Jacqueline Corrigan-Curay at the FDA stated "Quality RWE can't be built without quality RWD" and proposed that this be evaluated according to whether: (1) RWD is fit for use; (2) RWE study design answers the regulatory question, and (3) RWE meets regulatory requirements (e.g., standards for "substantial" evidence).[90,91] Assessing RWD fitness for use should be based on data reliability (e.g., accrual, control, precision, consistency, missingness, and availability of covariates) and data relevance (i.e., for intended regulatory question), and can be guided by good research practices for observational studies using RWD (e.g., Joint ISPOR-ISPE Task Force).

Prior to joining the FDA, Amy Abernethy co-authored an article that proposes a checklist to ensure "regulatory-grade" data quality.[92] It states that credible RWE should be developed from RWD that is obtained from sources relevant to the intended use, that is cleaned, harmonized, and linked to address any gaps, and that includes relevant endpoints. Key requirements of "regulatory-grade" RWD highlighted in this article included: (1) quality (e.g., clarity, traceability, and auditability); (2) completeness (e.g., using predefined rules and compared to appropriate benchmarks); (3) transparency (e.g., study aims, eligibility criteria, and design); (4) generalizability (e.g., identifying and addressing biases); (5) timeliness (e.g., recent); and (6) scalability (e.g., clear definitions that can be applied in larger datasets); these principles seem consistent with those highlighted by the FDA and provide a useful starting point to guide internal quality assessments of RWD.

These principles were reinforced in a recent draft guidance from the FDA on how to evaluate EHR and claims data to support a proposed regulatory decision.[93] This guidance emphasized the importance of ensuring the reliability (e.g., accuracy, completeness, provenance, and traceability) and relevance (e.g., availability of data elements related to exposure, outcomes, and covariates) of RWD. It highlighted common limitations to data from EHRs (e.g., data limited to HCPs using same system) and claims (e.g., clinical coding primarily for reimbursement purposes, and loss to follow-up when patients change health plans) and emphasized the importance of study design to mitigate these limitations.

More recently, the FDA issued a draft guidance on data standards for RWD in regulatory submissions.[94] It highlights the various challenges when attempting to standardize RWD from different sources (e.g., claims, EHRs, and registries), providers (e.g., third-party payers and health systems), and file formats (e.g., XML) into common data standards (e.g., Clinical Data Interchange Standards Consortium (CDISC)). The FDA advises sponsors to discuss their plans to submit RWD with their review divisions early in the process to align on methods, including data standards, analytical plans, and study methods. It also encourages sponsors to explain any challenges encountered when mapping RWD to common data standards (e.g., claims data based on patient sex vs. EHR data based on gender, and differences in number of categories available to classify patient race across sources), provide detailed documentation related to data provenance, curation, transformation, and cleaning, and a comprehensive data dictionary.

No sources of RWD contain data that is "fit for purpose" for every type of RWI or RWE study. Best practice dictates that every engagement with RWD begin with a feasibility assessment to ensure that the data of interest are indeed relevant and reliable, which is determined by examining the vendor, their processes for data normalization, curation and quality control, the data collection setting, and its relevance to the study question, as well as technical components of data quality, such as formatting, missingness, and validity of specific data elements. This evaluation process may take up to several months to complete at significant internal cost in terms of personnel time and loss of their productivity on other projects, and may result in a biopharmaceutical company deciding to not pursue a RWE study. A source of RWE might fail a feasibility assessment for many reasons, such as not finding the cohort of interest (e.g., diagnosis and procedure codes unable to define a specific patient population), lack of generalizability (e.g., patients are from a single payer or practice setting), not containing measures or endpoints of interest (e.g., new biomarker not widely available in routine practice), and masking of sensitive endpoints, (e.g., cause of death).

Transparency

Another aspect that should be considered by biopharmaceutical companies when deriving insights from RWD—even when such insights are intended primarily to inform internal decisions—is transparency. A key difference between analyses of RWD and analyses of data collected in prospective RCTs is that there is minimal third-party oversight in RWD. Biopharmaceutical companies can readily obtain large sources of RWD through commercial licensing agreements and begin to analyze this data without interacting with an ethics review board. Although the accessibility of RWD is one of its salient features, concerns have been raised that analyses of RWD may be prone to bias if researchers adjust their methodology to obtain the desired results.[95] While such concerns were voiced primarily in the context of analyses conducted by biopharmaceutical companies and submitted to support external decisions (e.g., regulatory approvals and payer coverage), the principle of transparency in analyses of RWD are equally valid when intended to support internal decisions.

The Professional Society for Health Economics and Outcomes Research (ISPOR) and International Society for Pharmacoepidemiology (ISPE) created a task force to address concerns about the lack of transparency in RWE studies and assure the integrity of analyses conducted using RWD.[95,96] This effort focused on Hypothesis Evaluating Treatment Effectiveness (HETE) studies that seek to test a hypothesis; exploratory RWD analyses were excluded. Key Task Force recommendations for those conducting HETE studies included: (1) declaring at onset if the goal is to conduct an HETE study; (2) public registration of study protocol and statistical analysis plan (SAP) prior to analyses; (3) explanation of any deviations from the registered protocol and SAP in publications; (4) facilitating the replication of HETE studies; (5) conducting HETE studies in different RWD sources than those used to inform development of protocol and SAP; (6) publicly addressing any criticisms raised about RWE studies; and (7) involving key stakeholders in the design and dissemination of HETE studies.

A companion project by ISPE described the steps needed to provide transparency that is sufficient to enable other researchers to reproduce or replicate RWE study findings, which is considered an important element to increase public trust in RWE.[97] Key recommendations to achieve greater transparency when reporting

RWE studies included fully describing the: (1) RWD used (e.g., provider, data type, and dates); (2) data processing (e.g., cleaning or transformation prior to analyses); (3) study design using a diagram (e.g., patient flow); (4) study inclusion/exclusion criteria; providing operational definitions of (5) study exposure (e.g., medication); (6) follow-up period (e.g., time from first exposure); (7) outcomes of interest (e.g., occurrence of events); (8) covariates used (e.g., comorbidity scores); (9) control groups (e.g., matching method); and (10) any statistical software (e.g., name, version, and packages) used. Researchers also encouraged the public sharing of source data and programs but acknowledged the potential barriers to doing so (e.g., prohibitions in data use agreements).

A related project termed the RWE Transparency Initiative included representatives from ISPOR, ISPE, the National Pharmaceutical Council, and the Duke-Margolis Center for Health Policy, and proposed ways to implement task force recommendations and promote transparency in RWE.[5] Recommendations included using existing clinical trial platforms (e.g., ClinicalTrials.gov) to register HETE studies, developing a standardized template for HETE protocols, and enlisting regulators, third-party payers, policy makers, and journal editors to promote the registration of HETE studies until it is as routine as the registration of clinical trials. Such efforts to promote transparency in RWE are expected to increase the overall quality of RWE studies over time.

Researchers then attempted to implement the ISPOR/ISPE task force and RWE Transparency Initiative recommendations by developing a reporting checklist and standardized template for RWE studies.[7] This initiative—termed the "Structured template and reporting tool for real-world evidence" or STaRT-RWE—is modeled after similar checklists and templates for reporting other study types (e.g., CONSORT for RCTs). Although this effort is focused on improving reporting to increase transparency and replicability, STaRT-RWE could also help researchers in designing RWE studies by addressing in their protocol all the elements highlighted as critical for future reporting. Researchers within biopharmaceutical companies who analyze RWD should consider how these principles can be applied to analyses intended for internal decisions.

CONCLUSIONS

Both RWD and RWE can offer valuable insight to guide the numerous decisions that must be made within biopharmaceutical companies throughout the product development cycle. Although recent literature has often focused on the potential regulatory uses of RWE, such uses are nascent and prone to change as regulators develop more conclusive guidance and experience in this domain. In the interim, RWD and RWE can continue to help to answer research questions related to the patient population of interest, healthcare utilization, and SOC in those patients, as well as natural history and disease progression. These insights can inform the feasibility, design, and efficiency of clinical trials. Internal stakeholders should consider the quality of the RWD used to develop these insights and should endeavor to increase the transparency of analyses based on RWD.

FUNDING

No funding was received for this work.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

1. Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). *Framework for FDA's Real World Evidence Program*. (U.S. Food and Drug Administration, Silver Spring, MD, 2018).
2. Rudrapatna, V.A. & Butte, A.J. Opportunities and challenges in using real-world data for health care. *J. Clin. Invest.* **130**, 565–574 (2020).
3. U.S. Food and Drug Administration (FDA). *Sentinel System: Five-Year Strategy 2019–2023*. (U.S. Food and Drug Administration (FDA), Silver Spring, MD, 2019).
4. Franklin, J.M. *et al*. Nonrandomized real-world evidence to support regulatory decision making: process for a randomized trial replication project. *Clin. Pharmacol. Ther.* **107**, 817–826 (2020).
5. Orsini, L.S. *et al*. Improving transparency to build trust in real-world secondary data studies for hypothesis testing-why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Value Health* **23**, 1128–1136 (2020).
6. Baumfeld Andre, E., Reynolds, R., Caubel, P., Azoulay, L. & Dreyer, N.A. Trial designs using real-world data: the changing landscape of the regulatory approval process. *Pharmacoepidemiol. Drug Saf.* **29**, 1201–1212 (2020).
7. Wang, S.V. *et al*. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ* **372**, m4856 (2021).
8. O'Donnell, J.C. *et al*. Evolving use of real-world evidence in the regulatory process: a focus on immuno-oncology treatment and outcomes. *Future Oncol.* **17**, 333–347 (2021).
9. Naidoo, P. *et al*. Real-world evidence and product development: opportunities, challenges and risk mitigation. *Wien. Klin. Wochenschr.* **133**, 840–846 (2021).
10. LoCasale, R.J. *et al*. Bridging the gap between RCTs and RWE through endpoint selection. *Ther. Innov. Regul. Sci.* **55**, 90–96 (2021).
11. Abrahami, D., Pradhan, R., Yin, H., Honig, P., Baumfeld Andre, E. & Azoulay, L. Use of real-world data to emulate a clinical trial and support regulatory decision making: assessing the impact of temporality, comparator choice, and method of adjustment. *Clin. Pharmacol. Ther.* **109**, 452–461 (2021).
12. Baumfeld Andre, E. & Honig, P.K. Overcoming regulatory aversion to novel methods of evidence generation. *Clin. Pharmacol. Ther.* **107**, 1057–1058 (2020).
13. Schuhmacher, A., Gassmann, O. & Hinder, M. Changing R&D models in research-based pharmaceutical companies. *J. Transl. Med.* **14**, 105 (2016).
14. Rassen, J.A., Bartels, D.B., Schneeweiss, S., Patrick, A.R. & Murk, W. Measuring prevalence and incidence of chronic conditions in claims and electronic health record databases. *Clin. Epidemiol.* **11**, 1–15 (2019).
15. Nelson, L.M. *et al*. A new way to estimate neurologic disease prevalence in the United States: illustrated with MS. *Neurology* **92**, 469–480 (2019).
16. Wallin, M.T. *et al*. The prevalence of MS in the United States: a population-based estimate using health claims data. *Neurology* **92**, e1029–e1040 (2019).
17. Broder, M.S., Cai, B., Chang, E. & Neary, M.P. Incidence and prevalence of neuroendocrine tumors of the lung: analysis of a US

commercial insurance claims database. *BMC Pulm. Med.* **18**, 135 (2018).

18. Yee, S.W. *et al*. Drugs in COVID-19 clinical trials: predicting transporter-mediated drug-drug interactions using in vitro assays and real-world data. *Clin. Pharmacol. Ther.* **110**, 108–122 (2021).
19. Bower, D. & Wisniowska, A. *Evolving a Market Access Strategy to Improve Patient Access*. (MJH Life Sciences, Cranbury, NJ, 2021) <https://www.pharmexec.com/view/evolving-a-market-access-strategy-to-improve-patient-access>.
20. Schwarze, K., Buchanan, J., Taylor, J.C. & Wordsworth, S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet. Med.* **20**, 1122–1130 (2018).
21. CMS Expands Coverage of Next Generation Sequencing as a Diagnostic Tool for Patients with Breast and Ovarian Cancer [press release]. Centers for Medicare & Medicaid Services (2020).
22. NICUSeq Study Group *et al*. Effect of whole-genome sequencing on the clinical management of acutely ill infants with suspected genetic disease: a randomized clinical trial. *JAMA Pediatr* https://doi.org/10.1001/jamapediatrics.2021.3496. [e-pub ahead of print].
23. Jarada, T.N., Rokne, J.G. & Alhajj, R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *J. Cheminform.* **12**, 46 (2020).
24. Khan, A., Miller, W.R. & Arias, C.A. Mechanisms of antimicrobial resistance among hospital-associated pathogens. *Expert Rev. Anti. Infect. Ther.* **16**, 269–287 (2018).
25. Marine, J.C., Dawson, S.J. & Dawson, M.A. Non-genetic mechanisms of therapeutic resistance in cancer. *Nat. Rev. Cancer* **20**, 743–756 (2020).
26. Galluzzi, L., Humeau, J., Buqué, A., Zitvogel, L. & Kroemer, G. Immunostimulation with chemotherapy in the era of immune checkpoint inhibitors. *Nat. Rev. Clin. Oncol.* **17**, 725–741 (2020).
27. Niedrig, D.F. *et al*. Clinical relevance of a 16-gene pharmacogenetic panel test for medication management in a cohort of 135 patients. *J. Clin. Med.* **10**, 3200 (2021).
28. Prelaj, A., Tay, R., Ferrara, R., Chaput, N., Besse, B. & Califano, R. Predictive biomarkers of response for immune checkpoint inhibitors in non-small-cell lung cancer. *Eur. J. Cancer* **106**, 144–159 (2019).
29. Gordon, D.E. *et al*. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science* **370, eabe9403** (2020).
30. Kim, H. *et al*. Combining PARP with ATR inhibition overcomes PARP inhibitor and platinum resistance in ovarian cancer models. *Nat. Commun.* **11**, 3726 (2020).
31. Shabani, M. & Marelli, L. Re-identifiability of genomic data and the GDPR: Assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation. *EMBO Rep* **20**, e48316 (2019).
32. Byrd, J.B., Greene, A.C., Prasad, D.V., Jiang, X. & Greene, C.S. Responsible, practical genomic data sharing that accelerates research. *Nat. Rev. Genet.* **21**, 615–629 (2020).
33. Evans, B.J. *et al*. How can law and policy advance quality in genomic analysis and interpretation for clinical care? *J. Law Med. Ethics* **48**, 44–68 (2020).
34. Aeffner, F. *et al*. Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. *J. Pathol. Inform.* **10**, 9 (2019).
35. Arazi, O. AI won't replace radiologists, but it will change their work. Here's how: World Economic Forum (2020) <https://www.weforum.org/agenda/2020/10/how-ai-will-change-how-radiologists-work/>.
36. Bedrikovetski, S. *et al*. Artificial intelligence for pre-operative lymph node staging in colorectal cancer: a systematic review and meta-analysis. *BMC Cancer* **21**, 1058 (2021).
37. Avanzo, M., Stancanello, J. & El Naqa, I. Beyond imaging: the promise of radiomics. *Phys. Med.* **38**, 122–139 (2017).
38. Moore, S.M. *et al*. De-identification of medical images with retention of scientific research value. *Radiographics* **35**, 727–735 (2015).
39. Rutherford, M. *et al*. A DICOM dataset for evaluation of medical image de-identification. *Sci. Data* **8**, 183 (2021).
40. Byrne, A.T. *et al*. Interrogating open issues in cancer precision medicine with patient-derived xenografts. *Nat. Rev. Cancer* **17**, 254–268 (2017).
41. Rossi, G., Manfrin, A. & Lutolf, M.P. Progress and potential in organoid research. *Nat. Rev. Genet.* **19**, 671–687 (2018).
42. Prasad, M., Kumar, R., Buragohain, L., Kumari, A. & Ghosh, M. Organoid technology: a reliable developmental biology tool for organ-specific nanotoxicity evaluation. *Front. Cell Dev. Biol.* **9**, 696668 (2021).
43. Hu, L.F., Yang, X., Lan, H.R., Fang, X.L., Chen, X.Y. & Jin, K.T. Preclinical tumor organoid models in personalized cancer therapy: Not everyone fits the mold. *Exp. Cell Res.* **408**, 112858 (2021).
44. Lara, H. *et al*. Quantitative image analysis for tissue biomarker use: a white paper from the digital pathology association. *Appl. Immunohistochem. Mol. Morphol.* **29**, 479–493 (2021).
45. Goulooze, S.C. *et al*. Beyond the randomized clinical trial: innovative data science to close the pediatric evidence gap. *Clin. Pharmacol. Ther.* **107**, 786–795 (2020).
46. Shah, P.D. *et al*. Combination ATR and PARP Inhibitor (CAPRI): A phase 2 study of ceralasertib plus olaparib in patients with recurrent, platinum-resistant epithelial ovarian cancer. *Gynecol. Oncol.* **163**, 246–253 (2021).
47. Bartlett, V.L., Dhruva, S.S., Shah, N.D., Ryan, P. & Ross, J.S. Feasibility of using real-world data to replicate clinical trial evidence. *JAMA Netw. Open* **2**, e1912869 (2019).
48. DiMasi, J.A., Grabowski, H.G. & Hansen, R.W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.* **47**, 20–33 (2016).
49. Wouters, O.J., McKee, M. & Luyten, J. Estimated research and development investment needed to bring a new medicine to market, 2009–2018. *JAMA* **323**, 844–853 (2020).
50. Paul, S.M. *et al*. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9**, 203–214 (2010).
51. Martina, R., Jenkins, D., Bujkiewicz, S., Dequen, P., Abrams, K. & GetReal, W. The inclusion of real world evidence in clinical development planning. *Trials* **19**, 468 (2018).
52. Jewell, N.P. Natural history of diseases: statistical designs and issues. *Clin. Pharmacol. Ther.* **100**, 353–361 (2016).
53. Bali, T. *et al*. Defining SOD1 ALS natural history to guide therapeutic clinical trial design. *J. Neurol. Neurosurg. Psychiatry* **88**, 99–105 (2017).
54. Murage, M.J. *et al*. Treatment patterns, adherence, and persistence among psoriasis patients treated with biologics in a real-world setting, overall and by disease severity. *J. Dermatolog. Treat.* **30**, 141–149 (2019).
55. Koehler, M., Donnelly, E.T., Kalanovic, D., Dagher, R. & Rothenberg, M.L. Pragmatic randomized clinical trials: a proposal to enhance evaluation of new cancer therapies with early signs of exceptional activity. *Ann. Oncol.* **27**, 1342–1348 (2016).
56. Selker, H.P. *et al*. A proposal for integrated efficacy-to-effectiveness (E2E) clinical trials. *Clin. Pharmacol. Ther.* **95**, 147–153 (2014).
57. Gamerman, V., Cai, T. & Elsäßer, A. Pragmatic randomized clinical trials: best practices and statistical guidance. *Health Serv. Outcomes Res. Method.* **19**, 23–35 (2018).
58. Ataga, K.I., Gordeuk, V.R., Agodoa, I., Colby, J.A., Gittings, K. & Allen, I.E. Low hemoglobin increases risk for cerebrovascular disease, kidney disease, pulmonary vasculopathy, and mortality in sickle cell disease: A systematic literature review and meta-analysis. *PLoS One* **15**, e0229959 (2020).
59. Scher, H.I., Solo, K., Valant, J., Todd, M.B. & Mehra, M. Prevalence of prostate cancer clinical states and mortality in the United States: estimates using a dynamic progression model. *PLoS One* **10**, e0139440 (2015).
60. Tabrizi, S.J. *et al*. Potential endpoints for clinical trials in premanifest and early Huntington's disease in the TRACK-HD study: analysis of 24 month observational data. *Lancet Neurol.* **11**, 42–53 (2012).

61. Rogers, J.R., Lee, J., Zhou, Z., Cheung, Y.K., Hripcsak, G. & Weng, C. Contemporary use of real-world data for clinical trial conduct in the United States: a scoping review. *J. Am. Med. Inform. Assoc.* **28**, 144–154 (2021).
62. Rottas, M. *et al*. Demographic diversity of participants in Pfizer sponsored clinical trials in the United States. *Contemp. Clin. Trials* **106**, 106421 (2021).
63. Weir, R.C., Proser, M., Jester, M., Li, V., Hood-Ronick, C.M. & Gurewich, D. Collecting social determinants of health data in the clinical setting: findings from national PRAPARE implementation. *J. Health Care Poor Underserved* **31**, 1018–1035 (2020).
64. Churchwell, K. *et al*. Call to action: structural racism as a fundamental driver of health disparities: a presidential advisory from the American Heart Association. *Circulation* **142**, e454–e468 (2020).
65. Shah, J.J. *et al*. Analysis of common eligibility criteria of randomized controlled trials in newly diagnosed multiple myeloma patients and extrapolating outcomes. *Clin. Lymphoma Myeloma Leuk.* **17**, 575–83 e2 (2017).
66. Dellon, E.S., Jensen, E.T., Martin, C.F., Shaheen, N.J. & Kappelman, M.D. Prevalence of eosinophilic esophagitis in the United States. *Clin Gastroenterol Hepatol.* **12**, 589–596 (2014).
67. Halpern, R., Seare, J., Tong, J., Hartry, A., Olaoye, A. & Aigbogun, M.S. Using electronic health records to estimate the prevalence of agitation in Alzheimer disease/dementia. *Int J Geriatr Psychiatry.* **34**, 420–31 (2019).
68. Chehade, M., Kamboj, A.P., Atkins, D. & Gehman, L.T. Diagnostic delay in patients with eosinophilic gastritis and/or Duodenitis: A population-based study. *J Allergy Clin Immunol Pract.* **9**, 2050–2059 (2021).
69. Morgan, J.C., Ye, X., Mellor, J.A., Golden, K.J., Zamudio, J., Chiodo, L.A., et al. Disease course and treatment patterns in progressive supranuclear palsy: A real-world study. *J Neurol Sci.* **421**, 117293 (2021).
70. Zhu, X.F., He, H.L., Wang, S.Q., Tang, J.Y., Han, B., Zhang, D.H., et al. Current treatment patterns of aplastic anemia in China: A prospective Cohort Registry Study. *Acta Haematol.* **142**, 162–70 (2019).
71. Stewart, M. *et al*. COVID-19 evidence accelerator: a parallel analysis to describe the use of hydroxychloroquine with or without azithromycin among hospitalized COVID-19 patients. *PLoS One* **16**, e0248128 (2021).
72. Foerster, M., McKenzie, F., Zietsman, A., Galukande, M., Anele, A., Adisa, C., et al. Dissecting the journey to breast cancer diagnosis in sub-Saharan Africa: Findings from the multicountry ABC-DO cohort study. *Int J Cancer.* **148**, 340–351 (2021).
73. Bakouny, Z., Paciotti, M., Schmidt, A.L., Lipsitz, S.R., Choueiri, T.K. & Trinh, Q.D. Cancer screening tests and cancer diagnoses during the COVID-19 pandemic. *JAMA Oncol.* **7**, 458–460 (2021).
74. Burn, E. *et al*. Use of dialysis, tracheostomy, and extracorporeal membrane oxygenation among 240,392 patients hospitalized with COVID-19 in the United States. *medRxiv* preprint. https://doi.org/10.1101/2020.11.25.20229088. [e-pub ahead of print].
75. Murk, W., Gierada, M., Fralick, M., Weckstein, A., Klesh, R. & Rassen, J.A. Diagnosis-wide analysis of COVID-19 complications: an exposure-crossover study. *CMAJ* **193**, E10–E18 (2021).
76. Harvey, R.A. *et al*. Association of SARS-CoV-2 seropositive antibody test with risk of future infection. *JAMA Intern. Med.* **181**, 672–679 (2021).
77. *Lessons From ISPOR 2021: The Growing Influence of Real-World Data and Real-World Evidence*. (TriNetX, Cambridge, MA, 2021) <https://trinetx.com/lessons-from-ispor-2021-the-growing-influence-of-real-world-data-and-real-world-evidence/> [October 15, 2021].
78. Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R.K. & Kumar, P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol. Divers.* **25**, 1315–1360 (2021).
79. Vamathevan, J. *et al*. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
80. Gates, A. *et al*. Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. *Syst. Rev.* **8**, 278 (2019).
81. Serrano Najera, G., Narganes Carlon, D. & Crowther, D.J. TrendyGenes, a computational pipeline for the detection of literature trends in academia and drug discovery. *Sci. Rep.* **11**, 15747 (2021).
82. Ochsner, S.A. *et al*. The Signaling Pathways Project, an integrated 'omics knowledgebase for mammalian cellular signaling pathways. *Sci. Data* **6**, 252 (2019).
83. Rajkomar, A. *et al*. Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* **1**, 18 (2018).
84. Jourdan, T., Debs, N. & Frindel, C. The contribution of machine learning in the validation of commercial wearable sensors for gait monitoring in patients: a systematic review. *Sensors (Basel)* **21**, 4808 (2021).
85. Beniczky, S., Karoly, P., Nurse, E., Ryvlin, P. & Cook, M. Machine learning and wearable devices of the future. *Epilepsia* **62**(Suppl 2), S116–S124 (2021).
86. Liu, R. *et al*. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* **592**, 629–633 (2021).
87. *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development*. (GAO USGAO, Washington, DC, 2019).
88. Bender, A. & Cortes-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discov. Today* **26**, 1040–1052 (2021).
89. Center for Devices and Radiological Health (CDRH), Center for Biologics Evaluation and Research (CBER). *Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: Guidance for Industry and Food and Drug Administration Staff.* (U.S. Food and Drug Administration (FDA), Silver Spring, MD, 2017).
90. Corrigan-Curay, J. *Framework for FDA's Real-World Evidence Program*. (U.S. Food and Drug Administration (FDA), Silver Spring, MD, 2019).
91. Corrigan-Curay, J. *The FDA, Real-World Evidence (RWE) Framework and Considerations for Use in Regulatory Decision-Making* (U.S. Food and Drug Administration (FDA), Silver Spring, MD, 2021).
92. 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).
93. Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Oncology Center of Excellence (OCE). *Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision Making for Drug and Biological Products* (U.S. Food and Drug Administration, Silver Spring, MD, 2021).
94. Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). *Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry.* (U.S. Food and Drug Administration, Silver Spring, MD, 2021).
95. Berger, M.L. *et al*. Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Pharmacoepidemiol. Drug Saf.* **26**, 1033–1039 (2017).
96. Berger, M.L. *et al*. Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE special task force on real-world evidence in health care decision making. *Value Health* **20**, 1003–1008 (2017).
97. Wang, S.V. *et al*. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies v1.0. *Pharmacoepidemiol. Drug Saf.* **26**, 1018–1032 (2017).