The Next Frontier: Fostering Innovation by Improving Health Data Access and Utilization

KA Oye[1], G Jain[2], M Amador[3], R Arnaout[4,5], JS Brown[6], W Crown[7], J Ferguson[8], E Pezalla[9], JA Rassen[10], HP Selker[11], M Trusheim[12] and G Hirsch[2]

Beneath most lively policy debates sit dry-as-dust theoretical and methodological discussions. Current disputes over the EU Adaptive Pathways initiative[1,2] and the proposed US 21st Century Cures Act[3] may ultimately rest on addressing arcane issues of data curation, standardization, and utilization. Improved extraction of information on the safety and effectiveness of drugs-in-use must parallel adjustments in evidence requirements at the time of licensing. To do otherwise may compromise safety and efficacy in the name of fostering innovation.

To take stock of the current state of the art, this essay identifies sources of demand for better utilization of real-world medical data, highlights the need for improved data quality, data access, and analytic methods, and evaluates the US Sentinel Initiative and Optum Labs as examples of distributed and centralized data initiatives.

To engage with emerging needs, this essay offers an integrated research and policy agenda. Academic research topics focus on improving data quality and access and on developing hybrid observational and interventionist methods to enhance causal inference under less-than-ideal conditions. Policy agendas focus on the need for international coordination on data access, data standards, and evidentiary thresholds.

DEMAND FOR BETTER UTILIZATION OF MEDICAL DATA

Three trends are now driving demand for improved extraction of information on the safety and effectiveness of drugs from clinical trials data, registries, claims data, and health records.[4] These include: 1) increasing pressures to accelerate access for patients in need of new and better treatments for a wider range of medical conditions; 2) splintering of indications into smaller treatment groups as a consequence of advances in genetics and in translational and "precision medicine"; 3) growing demands from payers and health technology assessment officials for quantifiable measures of relative effectiveness of new drugs. Defining the path forward requires an understanding of each of these trends, and how they are shaping the evolution of evidence requirements.

First, with the advent of adaptive approaches to drug licensing, the need to leverage observational data to reassess licensing decisions is expanding from a few drugs that target life-threatening unmet needs to many drugs addressing a wider variety of medical needs.[5] Calls for rapid access to new treatments originally came from advocates for patients with fast-progressing conditions such as HIV, cancer, and many orphan conditions. Patients with chronic, slowly progressing diseases with unsatisfactory treatment options are now making the same plea for rapid access. The EU Adaptive Pathways initiative and the US 21st Century Cures initiative both seek to address patients' demands for early access to a broadening array of treatments, with an associated need for reassessment of initial licensing decisions in light of evolving real-world evidence (RWE) on drug safety and effectiveness.

For these initiatives to succeed, regulators must strike an appropriate balance between addressing needs through early access to a drug and accepting uncertainty on the safety, efficacy, and effectiveness of a drug. The issue is the extent to which uncertainties must be resolved at the time of initial licensing and coverage decisions or whether positive decisions may be based on the balance of acceptable uncertainty at the time of licensing with continuous monitoring after initial licensing. Both Adaptive Pathways and 21st Century Cures initiatives are designed to foster the progressive reduction of uncertainty through pre-agreed evidence generation plans, with restrictions on initial utilization, and monitoring in the marketplace. Should emerging evidence suggest that the benefit–risk trade-offs are not acceptable,

[1]Massachusetts Institute of Technology (MIT) Department of Political Science and Engineering Systems Division, Cambridge, Massachusetts, USA; [2]Center for Biomedical Innovation, MIT, Cambridge, Massachusetts, USA; [3]MIT Portugal Program, International Risk Governance Council Portugal, Portugal; [4]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA; [5]Department of Pathology, Harvard Medical School (HMS), Boston, Massachusetts, USA; [6]Department of Population Medicine, Harvard Pilgrim Health Care Institute and HMS, Boston, Massachusetts, USA; [7]Optum Labs, Boston, Massachusetts, USA; [8]Shire, Lexington, Massachusetts, USA; [9]Aetna, Inc., Hartford, Connecticut, USA; [10]Aetion, Inc., New York, New York, USA; [11]Tufts Clinical and Translational Science Institute, Tufts University, and Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts, USA; [12]Sloan School of Management, MIT, Cambridge, Massachusetts, USA. Correspondence to: KA Oye (oye@mit.edu)

Received 20 July 2015; accepted 26 July 2015; advance online publication 3 August 2015. doi:10.1002/cpt.191

regulators should be legally authorized and politically empowered to restrict or withdraw a product even in the face of resistance from patients and sponsors. Together, these postlicensing components of these initiatives should improve the benefit–risk tradeoffs for patients relative to current approaches. These critical components at the back end of Adaptive Pathways and 21st Century Cures initiatives require improved postmarketing evidence generation and utilization.

Second, the recognition of population heterogeneity and the complexity of treatment effects has resulted in the splintering of indications and treatment groups. Precision medicines are directed at subpopulations defined by genotypic and phenotypic markers, with disease stratifications based on genotypic biomarkers accompanied by dedicated diagnostics. Acquiring subgroup-specific information on real-world benefits and risks will require improved postmarketing evidence generation and utilization.

With narrower definitions of indications, the recruitment of sufficient numbers of confounder-cleansed subjects for randomized controlled trials (RCTs) is becoming increasingly difficult. Conventional RCTs will be feasible for some identifiable subgroups with common mutations, but will not be feasible for rare mutations. For less common biomarkers, benefit–risk information based on real-world data accrued later in the lifespan of a product may be needed. Moreover, as more biomarkers are identified the trend is towards custom-made medicines. As patients receive individualized gene therapies based on modified patient-derived cells, antisense oligonucleotides, and other types of advanced therapies, treatment-eligible populations approach the limit of $n = 1$. What might be termed "basket licensing" of a family of medicines with individual variations may be the only viable route to market. However, even minor changes in the molecular structure of a drug could result in significant changes in toxicity profiles and treatment effects. An Adaptive Pathway approach with modification of initial basket licensing decisions grounded on rigorous observation of individual patient experience may become necessary. Uncertainty associated with long-term safety and effectiveness of targeted gene therapies and regenerative medicines has further accentuated the need for ongoing monitoring of therapies.

Third, payers, providers, and patients seek credible evidence on the relative effectiveness of new therapies. As the costs of new therapies continue to rise, demands for evidence-based pricing and coverage are increasing. Regulatory approval is a necessary but not sufficient condition for effective patient access. Payer decisions on the terms of coverage are keys to patient access to a new therapy, while the price of a therapy affects sponsor incentives for drug development and the willingness of private and public payers to extend coverage.

Payers are shifting from seeing decisions on reimbursement as a one-time binary decision, to seeing reimbursement decisions as an ongoing process aimed at providing greater certainty about value as evidence accumulates. Once a coverage decision has been made, payers have an interest in limiting initial use to subpopulations with the most favorable cost-effectiveness ratios, in improving patient adherence, in monitoring treatment outcomes, and in modifying conditions of reimbursement in light of RWE on effectiveness. All of these activities require improved generation and utilization of data on the safety and effectiveness of drugs in use.

Patients and providers are demanding more evidence than has traditionally been available to them. The evidence needed to inform treatment choices is both difficult to access and interpret. Furthermore, data are a principal source of influence in shaping regulatory licensing decisions and payer reimbursement decisions. Even when access to raw data is provided, patients and providers may lack the context-specific knowledge needed to interpret information, to combine varied sources of information, and to generate credible findings to inform treatment choices and to influence regulatory and payer decisions.

ELEMENTS OF NEED FOR EVIDENCE-BASED MEDICINE

The three trends above, taken together, have substantially increased demand for better understanding of the safety and effectiveness of drugs in use. The balance of this essay provides an overview of how those demands may be addressed in a more coordinated and timely manner, with emphasis on evidence generation through improved data quality and access; on analytical methods to enhance the reliability of findings on real-world safety and effectiveness under less-than-ideal conditions for robust causal inference; and on information tools and platforms to enable more rapid utilization of health data for decision making.

Data quality: integration and curation

RWE generation is now based on many types of data, including administrative claims data, electronic medical records, and patient and product registries. Furthermore, continuous readouts from medical instruments, continuous monitoring from mobile applications, and contextual information from social media are now available. With the exception of some registries, these data were not originally created for the evaluation of safety and effectiveness of drugs, so their value for such secondary uses is often unreliable. In addition, they embody varying degrees of quality, accuracy, and completeness, which adds further uncertainty to their utility in the generation of evidence. The integration of disparate and heterogeneous datasets that use different diagnostic categories and outcome measures, which include both genotypic and phenotypic information, and that are based on different principles of organization, is further a challenge. For example, acceptable evidence generation and use requires the ability to understand and adjust for possible bias in data sources related to such factors as data completeness, data capture incentives, clinical data workflows, and other factors.[6–8] Examples of selection biases which may affect data quality are referral filter, confounding by indication, and inclusion and exclusion criteria bias.

Data access: consent and ownership

Access to medical information is now limited by two types of constraints. Pharmaceutical companies own clinical trials data and some of the registries, public and private payers own claims data, and healthcare providers own medical records. Patients, on

the other hand, have varying degrees of access to their data and control over the uses of their own genotypic and phenotypic information. The combination of ownership and consent requirements presents formidable challenges to the generation of actionable evidence through the interrogation of disparate and distributed data sources across organizational boundaries. Furthermore, as the development for smaller subpopulation groups becomes more prevalent, like for ultra-orphan diseases, the need to interrogate data across organizational and international boundaries will increase. Intellectual property rights conventions, patient consent, and privacy protections vary markedly from nation to nation, further complicating effective utilization of medical data.

Data analysis: analytical methods and research designs to strengthen causal inference

All methods of evaluating the safety and effectiveness of drugs have limitations. The classic interventional studies using RCT provide a valid basis for inferring the safety and efficacy of a drug for strictly adhering confounder-cleansed populations, but real-world patient heterogeneity limits the external validity as a basis for predicting the safety and effectiveness on general populations with limited adherence. Observational studies based on real-world data can provide insights into the safety and efficacy of drugs, but the presence of selection effects, biases in data capture, and variable data completeness limit the internal validity of such studies. The issue is not to choose between interventional and observational studies, but rather to combine appropriate study designs for RCTs, observational studies, and pragmatic trials to accelerate evidence generation in a mathematically sound, statistically robust manner. This integrated evidence generation approach will accelerate delivery of targeted treatments to patients in need while simultaneously improving our understandings of safety, efficacy, and effectiveness. One critical operational issue reduces to the ability to work backwards from observational studies that may suggest safety problems or effectiveness issues to the definition of additional interventional studies.

STATE OF THE ART IN CURRENT DATA INITIATIVES

There are many data initiatives under way in the US and EU with varying approaches to addressing problems of data quality, data access, and analytical methods. **Table 1** provides an illustrative sample of data initiatives from the public sector, foundations, and professional societies, with information on lifespan stage, data types, target outputs, and regional emphasis. Some initiatives focus explicitly and narrowly on data quality. Clinical Data Interchange Standards Consortium (CDISC) and Public Health Data Standards Consortium (PHDC) focus on fostering data exchange standards in drug development and medical care delivery phases to improve interoperability, while the Quality Metrics initiative by the International Society for Pharmaceutical Engineering emphasizes standards for quality in discovery and development phases. Some established initiatives are engaging with data access, with goals, varying from an emphasis on the safety and effectiveness in oncology, as in ASCO's CancerLinQ, to consideration of safety across many indications, as in ENCePP and the Sentinel. The Innovative Medicines Initiative (IMI) in the EU is now supporting projects that will foster shared data access and the advancement of methods suited to improving understandings of the safety and effectiveness of drugs.

To anchor discussion of data quality, data access, and analytical methods, two examples involving different but potentially complementary data systems in the US are provided, each with its own distinctive architecture and mode of ownership. These include: the Sentinel Initiative, a public, distributed system, and Optum Labs, a private, centralized system. Both the Sentinel and Optum Labs are established initiatives with sufficient time-in-use to provide a practical basis for discussion.

Sentinel Initiative: a distributed data system[9]

The US Food and Drug Administration (FDA) created the Sentinel Initiative in response to a congressional mandate to enhance active postmarket surveillance for drugs. As part of the Sentinel Initiative, the FDA created the Mini-Sentinel project (now transitioning to "Sentinel") and the Mini-Sentinel operation center that coordinates data analysis, data-partner management, and distributed querying and analysis. The distributed nature of the network means that the data partners maintain control of their data behind their firewalls, and respond to queries sent by the operations center. The operations center collaborates with data partners to design algorithms and write distributed analytic code (in SAS) to be run by data partners. The Sentinel operations center sends out code to partners, partners run the code on their data, and partners send results of the analysis back to the Sentinel. As a matter of policy, Sentinel wants to bring as little data as possible from its data partners. The FDA covers the annual cost of around $14 million to maintain the data network as well as to support the operation center, maintenance of the data network, and query-related costs.[10] The network data partners in the Sentinel Initiative together maintain in a common data model electronic health data for more than 178 million US lives. The data include health insurance administrative and claims data coupled with clinical information for a subset of patients.[11]

Standards. Once SAS code is generated, no changes can be made by the user. The operations center develops, and data partners implement, the SAS programs. This is an example of standardized analytics, where the code is not changed for different data sources.[12] The analytics uses a common data model that leverages existing data standards commonly used by health insurers and clinical practices in the US (e.g., International Classification of Diseases, Clinical Modification [ICD-9-CM], American Medical Association's Current Procedural Terminology [CPT] codes, and National Drug codes [NDC]). Insurance companies or data partners maintain the patient-level data per the industry standards. Without access to patient-level data, Sentinel can identify potential biases and errors in the aggregate by comparing results across sources but data curation remains the responsibility of data partners.

Table 1 Illustrative examples of data initiatives

Name of data initiative	Lead/host organization	Lifespan stage			Data type					Target outputs				Geography		
		Discovery	Development	Delivery	Registry	Claims data	EMR	Clinical trials	Others	Biomarkers	Standards	Infrastructure	Methodology	US	EU	Other
Data Quality																
Clinical Data Interchange Standards Consortium (CDISC)	CDISC	x	x	x	x	x	x	x	x		x			x	x	x
Quality Metrics Initiative	International Society for Pharmaceutical Engineering	x	x						x		x			x		
Public Health Data Standards Consortium (PHDSC)	PHDSC		x	x	x	x	x				x			x		
Data Access																
CancerLinQ	American Society of Clinical Oncology		x	x	x	x	x	x		x		x		x		
Coalition Against Major Diseases (CAMD)	Critical PATH Institute		x			x				x		x	x	x		
ENCePP	European Medicines Agency		x	x			x	x				x			x	
European Medical Information Framework (EMIF)	Innovative Medicines Initiative			x	x	x	x				x	x	x		x	
Health Care Cost Institute (HCCI)	HCCI			x		x						x		x		
Optum Labs	Optum Labs		x	x		x	x		x			x		x		
PCORNet	PCORI		x	x		x	x	x				x		x		
Sentinel	US Food and Drug Administration		x	x		x						x	x	x		
Methods																
GetReal	Innovative Medicines Initiative		x	x	x	x	x					x	x		x	
Observational Medical Outcomes Partnership (OMOP)	Foundation of the National Institutes of Health		x	x		x	x					x	x	x		
PROTECT	Innovative Medicines Initiative		x	x	x	x	x	x					x		x	
Safer and Faster Evidence-based Translation (SAFE-T)	Innovative Medicines Initiative	x						x	x	x			x	x	x	

*Note: This is not an exhaustive list of data initiatives

ENCePP: The European network of centers for pharmacoepidemiology and pharmacovigilance; PCORI: The patient-centered outcomes research institute; PROTECT: pharmacoepidemiological research on outcomes of therapeutics by a European consortium.

Ownership. The Sentinel is a distributed data network, and, as such, the FDA does not collect any data itself. Data partners use data routinely collected for administrative, billing, and clinical care purposes. The data partners are responsible for the protection and appropriate use of the data. The data partners can opt-in for any queries initiated by the FDA.

Privacy and consent. Currently, only the FDA may initiate queries in Sentinel. Institutional Review Board (IRB) approvals are not required for FDA queries initiated to protect public health. Sentinel is part of a program of active surveillance for adverse effects of drugs, and as such, this provides the legal foundation for treatment of queries under a public health exception doctrine. Data partners can use their data as they wish, but must secure approval for research from IRBs with customary provisions for informed consent. For example, Aetna, a data partner for the Sentinel, complies with the provisions of the Health Insurance Portability and Accountability Act of 1996 (HIPAA)[13] for operations-related research using patient-level data.

Analytical methods. The reliance on a distributed structure, with queries to data owners returning marginal counts or intermediate matrices rather than individual-level data, complicates the use of advanced analytical methods to control for selection effects and interaction effects, limits data curation, precludes follow-up queries for further information on individual cases. The inference, however, is limited for small population indications, in cases when only the pooling of data from each partner would provide adequate statistical power to the analysis. At the same time, the distributed architecture is the key to providing Sentinel access to a very large and heterogeneous set of data sources.

Optum Labs, a centralized data system[14]

Optum Labs, an affiliate of UnitedHealth Group, is the first of its kind of an open, collaborative research and innovation center designed to harness RWE. A key asset of Optum Labs is a rich, high-quality, integrated healthcare database, covering more than 150 million US lives with deidentified claims and clinical data from multiple health plans and provider groups. The data are integrated across care settings and longitudinally linked at the patient level. The database scale supports identification and sizing of important subgroups and rare occurrences.

The data are assimilated and stored in a secure central location, rather than being held at their contributing sources, as in a distributed approach. The centralized data facilitate rapid turnaround on analyses, and enables close investigation of rare events, small patient subgroups, and other important inquiries that are difficult to conduct with a distributed data model. The centralized model also facilitates direct queries of the full database to test project feasibility and generate preliminary descriptive results without needing to distribute common protocols to remote data holders, obtain iterative approvals, and aggregate results across sites postanalysis.

Standards. All claims data from the contributing plans are stored in a common data model and subjected to numerous edit checks to cleanse the data to support research. The same is true of the EMR data contributed by providers. However, in the case of the EMR data up to 85% of the clinical content originally resides in unstructured data. As a consequence, thousands of clinical data elements are extracted from EMRs through natural language processing (NLP), curated, normalized, and then linked to claims at an individual level to add great clinical depth to the claims population breadth.

Ownership. As with Sentinel, Optum Labs utilizes data routinely collected for administrative, billing, and clinical care purposes. Data rights to use deidentified claims and EMR data for research is obtained by Optum Labs from the contributing health plans and provider groups under its business associate relationships with these organizations. This enables the use of a centralized data system with significant benefits as noted below.

Privacy and consent. Only deidentified patient data are held by Optum Labs and used for research. The patient-protected health information is encrypted and double-hashed into unique identifiers that enable data to be linked across sources in a completely deidentified manner. Research enclaves are established for each research project. Data views are prepared for researchers that contain data elements required for their research problem. A statistical expert in deidentification approves all views and determines that the data elements in combination present a "very small" risk of reidentification, as specified by HIPAA.

Analytical methods. The reliance on a centralized structure, with direct access to anonymized individual data, permits the use of advanced analytical methods to control for selection effects and interaction effects and allows follow-up queries for further information on individual cases. At the same time, the centralized architecture limits Optum Labs access to a smaller set of data sources than Sentinel. Sentinel has access to larger claims data, whereas the scale of the linked clinical and claims data is larger in Optum Labs.

Extensions and alternatives

The strengths and weakness of Sentinel and Optum Labs are intrinsically related to the characteristics of distributed public and centralized private data models.

According to the Center of Biomedical Innovation at Massachusetts Institute of Technology (MIT), there is interest among global government agencies and foundations focused on biomedical innovation in the US, EU, Canada, and Asia for the potential global scale-up of the Sentinel model (unpublished survey). In addition, there is interest among payers and pharmaceutical companies to explore the use of the Sentinel model to answer both safety and effectiveness-related questions.[15] It is unclear if such work would require IRB approval as queries move beyond areas covered by the current public health exemption. Problems of data curation and causal inference as noted above would remain challenging for the use of a distributed data system to reach conclusions on effectiveness.

A broader version of the private centralized approach modeled after Optum Labs could be viable, with other private and public

healthcare networks nucleating around the Optum Labs network or forming on their own. By pooling individual-level information from multiple sources, such an expanded private network would have advantages in terms of inferring safety and effectiveness of drugs targeted at small populations and detect potential systematic biases in data sources. Most significantly, deanonymized individual level data could be used to detect and correct errors in databases and to test hypotheses from observational data by setting up focused interventions and pragmatic clinical trials. Although deanonymization of data provides potential for additional insights, it needs to balance the privacy requirements as specified by HIPAA.

Product, patient, disease, and treatment-specific registries provide an alternative to general distributed and centralized data models. In fact, sponsors and regulators are increasingly turning to product registries to address the immediate need for control over the quality of the data on the safety and efficacy of each product. Over time, though, the lack of interoperability of these registries will likely slow the evolution of knowledge about the disease at the individual and population levels due to the fragmentation of associated information. A carefully curated registry has clear advantages relative to general-purpose distributed and centralized data initiatives. However, registries do not eliminate the need for more general purpose data initiatives. First, product-specific registries do not address payer demands for comparative effectiveness data on a product and on alternative therapies. Second, developing product-specific evidence may require linking multiple registries, each of which captures data on narrow subpopulation; e.g., regenerative medicines are typically variations on a technology platform with slight modification for different indications requiring separate registries. Third, given specific enrollment criteria and sponsorship of registries, inclusion and exclusion biases are common, and generalizability is often problematic. Some combination of more targeted registries and general-purpose data initiatives may be needed to generate high-quality generalizable data at reasonable cost.

Various emerging software companies, like Aetion Inc., have developed tools to address the growing need for increased speed of evidence generation particularly using real-world data. These rapid-cycle analytic tools could be consistently applied to different populations or data sources (including the sources noted above), on a pooled- or source-by-source basis, which may help provide better understanding of the variations of treatment effects while using standardized analytics. Such tools can reduce the time to conduct epidemiological and cost-effectiveness research using large real-world datasets.

CONCLUSIONS: RECOMMENDED RESEARCH TOPICS AND POLICY AGENDAS

Research topics and policy agendas are typically decoupled, to be addressed by technical and policy experts working in separate spheres. Academic research topics focus on improving data quality and access and on developing hybrid observational and interventionist methods to enhance causal inference under less-than-ideal conditions. Policy agendas focus on the need for international coordination on data access, data standards, and evidentiary thresholds. These spheres are in fact connected. Failures in the realm of policy coordination and standard setting may underscore the need for development of technical methods for dealing with nonstandardized data, while successes may ease technical challenges. Conversely, technical work on estimation of biases in data may inform the policy problem of setting of evidentiary standards and thresholds.

Data quality

Existing data initiatives such as CDISC have rightly focused on the problems of data standardization and interoperability. The curation of heterogeneous datasets is an under-researched topic. At the most basic level, detecting and correcting for simple measurement and data entry errors is a challenge. Developing and improving analytical tools for recognition of inconsistent values and development of better testing of the fidelity of existing and emerging analytical methods will be an important contribution. At a more advanced level, all data sources have some elements of bias. Understanding and explaining the specific types of bias is important for developing and communicating RWE. For example, evaluating and adjusting for possible biases in diagnostic codes in claims data will be difficult. One idea for estimating biases in claims data entails exploiting natural experiments, where the strength of biases may be expected to differ. For example, in Singapore reimbursement for drug costs is done from medical savings accounts without fixed formularies. This weakens incentives for physicians to shade diagnostics codes to justify patient reimbursements. Statistically controlled comparisons of diagnoses and prescriptions in Singapore with diagnoses and prescriptions in healthcare systems with conventional formularies, like the US, may provide a basis for better estimating biases in diagnostic codes and in calculating rates of off-label use. These examples are illustrative, and extended brainstorming focused on these topics would be fruitful.

Data access

The Sentinel Initiative and Optum Labs overcame data access problems using different approaches. It is important to conduct analysis of intellectual property rights and privacy/informed consent requirements in consultation with specialists on those issues. Data ownership and access rules vary markedly even within OECD nations, with US and EU differences on public access to trials data and with substantial variation in protections for genomic and phenotypic information. An important issue for consideration, particularly in the context of the potential extension and/or scale-up of the Sentinel Initiative beyond its piloted use in addressing safety queries posed by the FDA, is the distinction between public health and research topics that serves public interests. In addition, it would be valuable to analyze alternative methods of defining and securing consent for utilization of linked genotypic and phenotypic information. And finally, a systematic analysis of differences between OECD and non-OECD nations would be useful in a world of globalized drug development, production, and distribution.

Analytical methods

Rather than recapitulating conventional debates on the relative merits of RCTs vs. observational data, researchers should focus on

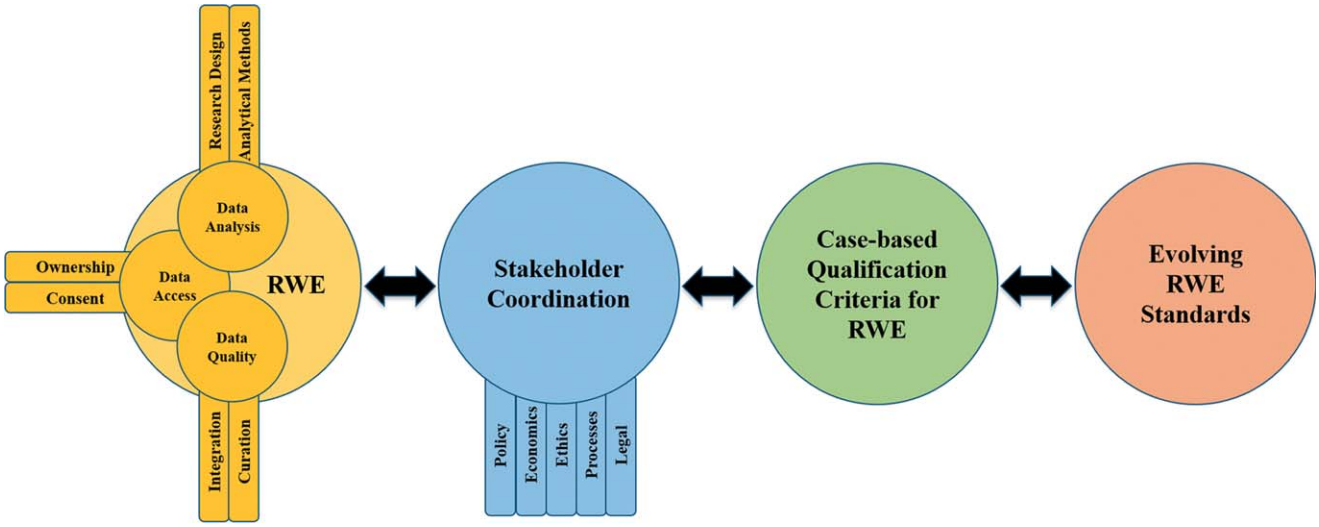

Figure 1 Framework for action: evolving pipeline for informing real-world evidence (RWE) standards. The schematic illustrates a framework to catalyze the generation and utilization of RWE and informing the development of RWE standards. As described in the lighter yellow circles, there are several components affecting the quality of the RWE: data quality, data access, and analytical methods (dark yellow circles and rectangles). National and international coordination of stakeholders is needed to not only improve the quality of RWE, through data curation and interoperability standards, intellectual property rights, subject patient consent, and evidentiary thresholds, but to agree on how to use RWE by addressing policy, economics, ethics, processes, and legal issues (blue circle). Insights and outputs from many public–private initiatives to improve the quality of RWE can be integrated by developing a case-based qualification criteria for RWE acceptable to stakeholders (green circle). This, in turn, will advance the discussion towards acceptable RWE standards by the key decision-makers (orange circle). The two-sided arrows represent the influence of one on another.

how observational and experimental methods could be leveraged in a complementary way to improve inferential validity. Specifically, how the integration of evidence using RCTs, observational studies, and pragmatic trials in an unconventional order could accelerate the delivery of targeted treatments to patients in need, while simultaneously improving our understandings of safety and efficacy. For example, observational methods might generate suggestive results on comparative effectiveness, with reservations based on the possible existence of selection biases. The random assignment of patients to treatment groups in a pragmatic trial could resolve uncertainty over observational results.

Policy coordination and standardization

Moving beyond research, progress in the generation and utilization of RWE would be fostered by international coordination on data curation and interoperability standards, intellectual property rights, subject patient consent, and evidentiary thresholds. Many public–private initiatives to explore the prospects for coordination on these issues are under way globally (**Table 1**), but integrating the insights and outputs from these efforts into practice will be challenging. Ultimately, standards are needed to clarify the requirements for acceptability by the key decision-makers. Elements of these standards are beginning to evolve in the US[16] and EU.[17] However, using the evolutionary timeline of nearly 60 years so far for Good Clinical Practice (GCP) guidelines as a benchmark,[18] it will take years for these standards to evolve. In the meantime, there are immediate actions that can catalyze and inform the development of standards (**Figure 1**), particularly when carried out in the context of collaborative multistakeholder efforts. For example, in IMI at least two relevant initiatives, GetReal[19] and Big Data for Better Outcomes,[20] are starting to explore the issues related to standards for RWE.

In addition, the MIT New Drug Development Paradigms (NEWDIGS) initiative's Data Program was launched in 2014 to provide an integrative test-bed environment to catalyze coordinated progress in this area. Structured as a precompetitive multistakeholder collaboration, and leveraging the systems engineering expertise of MIT, NEWDIGS is committed to building on the foundation it laid in its work on Adaptive Pathways that inspired the EMA pilot project launched in March 2014,[1,2,21] in order to address this critical set of innovation enablers. NEWDIGS uses a case-based collaborative learning and innovation methodology that proved valuable in its early work on Adaptive Pathways.[22] These confidential discussions happen in a safe haven environment, under the Chatham House rule,[23] which enhances candid dialog and sharing of proprietary data. Moving forward, this "scenario design" methodology, supported by an evolving open access modeling and simulation platform,[24] will be applied to explore case-based decision qualification criteria for RWE. In addition, the implications of these options for each stakeholder and for global health will be assessed.

The pace of change in the generation and exploitation of data on the real-world effects of drugs will be affected by the extent of intranational, international, and interregional cooperation. National initiatives in the US and EU are under way. In addition to the initiatives discussed in **Table 1**, the FDA just issued a request for a proposal for source data capture from electronic health records, using standardized clinical research data.[25] The internationalization of such initiatives is coming. Pooling of data will be useful in assessing the safety and efficacy of a drug targeting a small pool of

patients. Pooling of data from firms or nations with different degrees of care in limiting off-label uses will be useful in discovering second and third uses of previously licensed drugs and in identifying combination therapies. Pooling resources in the development of analytic methods to make better use of observational data, to overcome problems with data quality, and to integrate observational methods and trials may create the tools that will be needed to accelerate change in data generation and exploitation.

ACKNOWLEDGMENTS, DISCLAIMER

This article is the product of a multistakeholder collaboration under the NEWDIGS initiative of the MIT Center for Biomedical Innovation. All authors participated in a forum organized by NEWDIGS where the issues mentioned in this article were discussed. The views expressed are those of the authors and should not be understood or quoted as being made on behalf of or reflecting the position of the agencies or organizations with which the authors are affiliated.

AUTHOR CONTRIBUTIONS

K.A.O., G.J., M.A., R.A., J.S.B., W.C., J.F., E.P., J.A.R., H.P.S., M.T., and G.H. wrote the article.

CONFLICT OF INTEREST

All the authors have institutional affiliations, corporate affiliations, and/or financial interests in the subject matter discussed in this work as noted in the author list. The Center for Biomedical Innovation currently receives or has received financial support from the Sloan Foundation, Kauffman Foundation, Massachusetts Technology Collaborative, and Robert Woods Johnson Foundation, as well as consortium members Aetna, Allscripts, Amgen, Aquafine BioProcess, Asahi Kasei, Baxter, Biogen, BioMarin, bluebird bio, Boehringer Ingelheim, Bristol-Myers Squibb, CSL Behring, Eli Lilly, EMD Millipore, Genentech, Genzyme, GlaxoSmithKline, Histogenics, KBI Biopharma, Latham BioPharm Group, Life Technologies, MedImmune, Merck and Co., Merck Serono, Millennium Pharmaceuticals, Novartis, Pall Corporation, Pfizer, Regeneron, Sanofi, Sanofi Pasteur, Shire, and Sigma-Aldrich Fine Chemicals. J.A.R. is a co-owner and employee of Aetion, Inc.

1. EMA. Adaptive Licensing pilot project. <http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000601.jsp&mid=WC0b01ac05807d58ce> (2014). Accessed July 17, 2015.
2. Eichler, H.G. *et al*. From adaptive licensing to adaptive pathways: delivering a flexible life-span approach to bring new drugs to patients. *Clin. Pharmacol. Ther.* **97**, 234–246 (2015).
3. U.S. House of Representatives. Rules committee print 114-22. Text of H.R. 6, 21st Century Cures Act. <http://docs.house.gov/billsthisweek/20150706/CPRT-114-HPRT-RU00-HR6.pdf> (2015). Accessed July 17, 2015.
4. Oye, K.A., Pearson, M., Eichler, H.G., Mullin, T. & Hoos, A. Managing uncertainty in drug development and use: enhancing adaptability and flexibility in pharmaceuticals regulation and use. <http://www.irgc.org/wp-content/uploads/2014/10/OYE-IRGC-Rethinking-Risk-Regulation-Conference-2014.pdf> (2014). Accessed July 17, 2015.
5. Baird, L.G. *et al*. Accelerated access to innovative medicines for patients in need. *Clin. Pharmacol. Ther.* **96**, 559–571 (2014).
6. Kahn, M.G. *et al*. Transparent reporting of data quality in distributed data networks. *EGEMS (Wash. DC).* **3**, (2015).
7. Brown, J.S., Kahn, M. & Toh, S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med. Care.* **51**, S22–29 (2013).
8. Toh, S., Gagne, J.J., Rassen, J.A., Fireman, B.H., Kulldorff, M. & Brown, J.S. Confounding adjustment in comparative effectiveness research conducted within distributed research networks. *Med. Care.* **51**, S4–10 (2013).
9. Platt, R. *et al*. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol. Drug Saf.* **21** (suppl. 1), 1–8 (2012).
10. Platt, R. FDA's Mini-Sentinel program to evaluate the safety of marketed medical products. A look back, a look ahead. <http://www.brookings.edu/~/media/events/2014/1/14-sentinel-initiative-public-workshop/richard_platt_slides.pdf> (2014). Accessed July 10, 2015.
11. Curtis, L.H. *et al*. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol. Drug Saf.* **21** (suppl. 1), 23–31 (2012).
12. Sentinel. Routine querying tools (modular programs). <http://www.mini-sentinel.org/data_activities/modular_programs/default.aspx>. Accessed July 10, 2015.
13. US Department of Health and Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule. <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf> (2012). Accessed July 7, 2015.
14. Wallace, P.J., Shah, N.D., Dennen, T., Bleicher, P.A. & Crown, W.H. Optum Labs: building a novel node in the learning health care system. *Health Aff. (Millwood).* **33**, 1187–1194 (2014).
15. Baldziki, M. *et al*. Utilizing data consortia to monitor safety and effectiveness of biosimilars and their innovator products. *J. Manag. Care. Spec. Pharm.* **21**, 23–34 (2015).
16. FDA. Guidance for industry. Good pharmacovigilance practices and pharmacoepidemiologic assessment. <http://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM126834.pdf> (2005). Accessed July 15, 2015.
17. EMA. Guidelines on good pharmacovigilance practices (GVP). <http://www.ema.europa.eu/docs/en_GB/document_library/Other/2015/04/WC500186216.pdf> (2015). Accessed July 15, 2015.
18. Vijayananthan, A. & Nawawi, O. The importance of Good Clinical Practice guidelines and its role in clinical trials. *Biomed Imaging Interv. J.* **4**, e5 (2008).
19. IMI GetReal project website. <http://www.imi-getreal.eu/>. Accessed July 15, 2015.
20. IMI. Consultation paper for big data for better outcomes—concept for an IMI2 programme. <http://www.imi.europa.eu/sites/default/files/uploads/documents/Events/SF%202015/BigData_Concept_29May2015.pdf> (2015). Accessed July 15, 2015.
21. Eichler, H.G. *et al*. Adaptive licensing: taking the next step in the evolution of drug approval. *Clin. Pharmacol. Ther.* **91**, 426–437 (2012).
22. Baird, L. & Hirsch, G. Adaptive licensing: creating a safe haven for discussions. *Scrip Regulatory Affairs* 2013 August 20.
23. Chatham house rule. <http://www.chathamhouse.org/about/chatham-house-rule>. Accessed July 7, 2015.
24. Trusheim, M.R., Baird, L.G., Garner, S., Lim, R., Patel, N. & Hirsch, G. The Janus initiative: a multi-stakeholder process and tool set for facilitating and quantifying Adaptive Licensing discussions. *Health Policy Technol.* **3**, 241–247 (2014).
25. FDA. Request for proposal: source data capture from electronic health records: using standardized clinical research data. <http://www.regulations.gov/#!documentDetail;D=FDA-2015-N-1887-0001> (2015). Accessed July 17, 2015.