Translational Research, Design and Analysis Perspective

Real-world data for 21st-century medicine: The clinical and translational science awards program perspective

Pablo Cure[1] ⬤, Sanae ElShourbagy Ferreira[1], Joshua P. Fessel[1] ⬤, Deborah Ossip[2,3], Martin S. Zand[2,3,4] ⬤, Scott J. Steele[2,3,†] ⬤, Kenneth Gersing[1] ⬤ and Christopher M. Hartshorn[1]

[1]National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, MD, USA; [2]Center for Leading Innovation and Collaboration (CLIC), Clinical and Translational Science Program National Coordinating Center, University of Rochester Medical Center, Rochester, NY, USA; [3]Department of Public Health Sciences, University of Rochester Medical Center, Rochester, NY, USA and [4]Department of Medicine, Division of Nephrology, University of Rochester Medical Center, Rochester, NY, USA

Background

Real-world data (RWD) refers to data related to patient health status and/or the delivery of healthcare routinely collected from a variety of sources. The clinical evidence derived from analysis of RWD regarding the utilization and potential benefits or risks associated with a medical product or intervention is known as real-world evidence (RWE) [1,2]. Real-world data includes a variety of data sources such as electronic health records (EHR), laboratory and imaging, claims/billing, vital records, digital health technologies (DHT), and other modes of remote data collection that can be obtained retrospectively and/or prospectively during the patient's continuum of care. In addition, other types of data sources such as environmental exposures, pollution levels (indoor and outdoor), geolocation, text messaging, social media, economic measures, and other sources have the potential to "enrich" health-related data for patients and populations. In the last decade, several advances have brought the potential for RWD applications to a new level: an exponential increase in the sheer quantity of data that can be collected from single and multiple sources; data integration and "ingesting" capabilities from multiple Common Data Models (CDMs); new analytical tools and novel methods such as artificial intelligence (AI) and machine learning (ML); and higher capacity to store and maintain information locally and/or centrally within secure cloud environments.

To keep up with evolving types and uses for RWD, the Food and Drug Administration (FDA) has been developing a series of RWD guidance documents [3] and has launched an Advancing Real-World Evidence Program [4]. In addition, the National Institutes of Health (NIH), the National Cancer Institute (NCI), the National Center for Advancing Translational Sciences (NCATS), and the National Library of Medicine have been addressing the need for standardization around RWD collection and use. For example, since 2017, the FDA-led Common Data Models Harmonization and Open Standards for Evidence Generation project [5] has worked to ensure infrastructure is in place to support standardized data generated across the translational spectrum. These CDMs allow for a unified database model to help in the integration of various RWD sources according to the same standard, facilitating the interoperability of the data from multiple data sources for the conversion of RWD to RWE. During the COVID-19 pandemic, these efforts further materialized in the development of the National COVID-19 Cohort Collaborative (N3C) [6], one of the largest repositories of HIPAA-defined limited data sets in the country that, as of March 2023, includes data from over 77 sites in the U.S. and > 7 million COVID + cases and > 11 million controls, "ingesting" data from a variety of CDMs [6]. Through partnerships among many organizations that provide clinical data, and by securely making data accessible to more than 3,000 researchers, N3C has helped to answer critical questions about COVID-19 biology, clinical behavior, and treatment strategies. For example, one of the first RWD-driven characterizations of the post-acute sequelae of COVID-19 ("long COVID") was accomplished using N3C [7]. N3C offers one model of how to create a large de-identified database through collection, curation, and analysis of multisite data in a single protected data enclave that can facilitate rapid response to a public health emergency. Recently, to further the potential of these types of RWD approaches, NCATS launched the N3C Public Health Answers to Speed Tractable Results to deliver fast, actionable analyses on pressing COVID-19 questions [8]. Efforts such as N3C can continue to model how institutions, government, and other stakeholders can work together in developing large-scale RWD. In addition, combining clinical/health-related data from EHR-based datasets such as N3C with non-clinical data (e.g., sociodemographic, environmental, geolocation, etc.) can provide a more comprehensive picture of the health and well-being of individuals and communities.

Table 1. Challenges and opportunities in the development and implementation of RWD for RWE in the CTSA program

Key RWD Area	Challenges	Opportunities
Data Infrastructure, Sources, and Quality	Data quality control Combining datasets from multiple sources High quality datasets identification Data confidentiality, privacy, and security	Develop and disseminate good quality control practices and use cases Develop guidance on data sources and combining datasets Develop and disseminate AI and ML techniques to improve data quality Sharing best practices and develop novel consent platforms and processes for identifiable participant data
Integration, Harmonization and Analysis	Data sharing infrastructure and tools Health needs in RWD/RWE Data privacy and security for multisite data Data harmonization	Disseminate lessons learned from N3C on central data sharing and analytical tools (e.g., Enclave) Local and CTSA consortium-wide health needs in RWD/RWE Best practices for data sharing, privacy, and security standards for the CTSA program Data harmonization best practices and use of CDMs
Diversity, Equity, Inclusion, and Accessibility	Data representativeness and access to RWD technology sources Unintended consequences of RWD when evaluating large datasets Priorities in RWD/RWE based on patient, community and population needs Measuring impact of RWD in the health of individuals and populations	Develop local and national standards and strategies for diverse representation based on sociodemographic and geographic variables Lessons learned when using AI and ML in large datasets Community Engagement efforts at CTSA hubs to identify population needs in RWD Using RWD to improve trial recruitment and representation of minorities
Training, Education, and Career Development in RWD/RWE	Training and education competencies for next generation Training and resources in data science Real-time dissemination Diversity of workforce in RWD	Develop role-based training competencies and materials Data management training for researchers Provide access to RWD and tools locally and consortium wide to scholars/trainees Disseminate best practices and tools in RWD that can assist researchers RWD competencies and career development opportunities accessible to all Develop standards for Team Science in RWD

CDMs = common data models; CTSA = clinical and translational science awards; N3C = national covid cohort collaborative; RWD = real-world data; RWE = real-world evidence.

Through its Clinical and Translational Science Awards (CTSA) Program, NCATS supports a national network of biomedical research institutions that accelerate the translation of scientific observations into innovative health solutions [9]. Advancing translational science and improving the efficiency and effectiveness of translation requires a coordinated, collaborative effort. Incorporating RWD into translational science requires combining lessons learned, processes, and know-how from experts and from the communities directly impacted by advances in RWD. As the field rapidly evolves and more data from a wide variety of sources are incorporated into clinical and translational science (CTS) research and day-to-day patient care, the challenges and opportunities presented herein (Table 1) represent a possible blueprint for how the CTSA program could make RWD and RWE a reality to help transform 21st-century medicine.
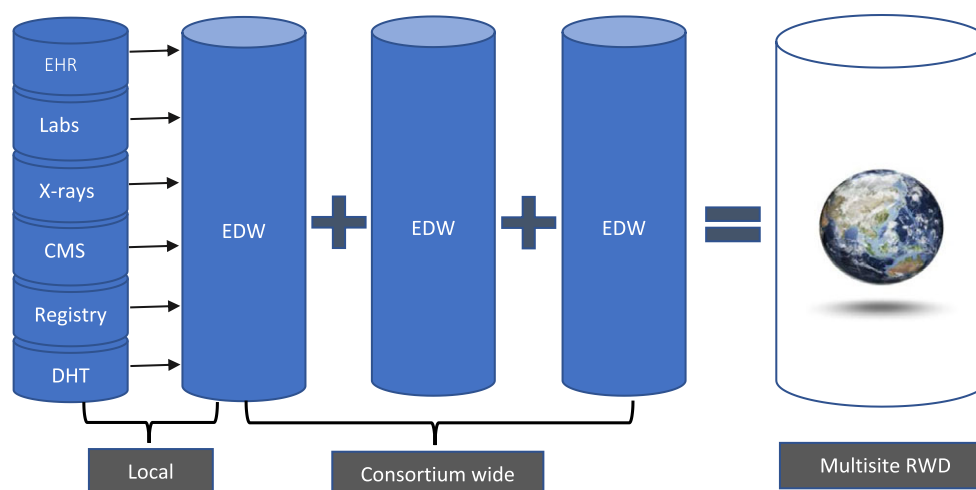
Infrastructure, Sources, and Quality

Data infrastructure, sources, quality, and reliability can be highly variable between individual data ecosystems [10–12]. Further, ethical aspects of use of patient-derived data for research and protection of privacy and confidentiality continue to pose challenges [13]. Informed consent, de-identification, data ownership, linkages, and sharing of data are issues inherent to any RWD effort [14]. Methods to store data and control levels of access for researchers and analysts to specific datasets are as important as the data itself. These challenges are further compounded by continuous and rapid advances in data science and adjacent fields. Together, the challenges require data experts and institutions to constantly re-visit their policies and procedures to maintain

the highest standards of quality, reliability, privacy, and security of data and their sources. In addition, data governance and provenance are important parts of the data ecosystem to ensure accuracy and quality of the data collected [15].

Capabilities to simultaneously collect/aggregate, securely link, and analyze data from several RWD sources are in great demand to provide a more complete assessment of patients' and communities' health (Fig. 1). Currently, EHRs are one of the main data sources for RWD; however, challenges with the quality, reliability, heterogeneity, and utility of the data collected in the EHR continue to be an issue [16–18]. Indeed, EHR systems were developed for clinical documentation, administrative, and billing purposes, so their "re-purposing" as RWD sources for CTS research requires adaptations in data collection, integration, validation, and analysis. Furthermore, overlaying "traditional" sources of RWD with other contextual information can potentially impact data robustness and privacy. Environmental exposures, geolocation, economic measures, and other data extracted from publicly and non-publicly available data sets can enrich analyses. Yet, sources not typically thought of as health-related have different regulations and policies governing their use that must be considered.

Some data sources are well-suited to RWD/RWE applications. For example, disease or specialty-specific registries and databases have successfully supported new indications for existing drugs [19–21]. These registries require strategic planning – incorporating adequate governance, infrastructure, resources, and expertise beyond the registry itself – and flexibility, allowing for dynamic evolution to bring in novel RWD data sources and methods that bolster the registry's utility. One example is the NCI Surveillance, Epidemiology, and End Results cancer registry [22,23], which

Figure 1. Overall data infrastructure, sources, and integration. CMS = centers for medicare & medicaid services claims data; DHT = digital health technologies; EDW = enterprise data warehouse; HER = electronic health records; Labs = laboratory values; X-ray images; RWD = real-world data.

provides a high-quality population base (state or metropolitan area) from central cancer registries that enable monitoring of disease prevalence, cancer population needs, and health disparities and their impact in these communities. Bringing RWD together from various sources is only the first step. Generating RWE requires making RWD research-ready and employing adequate and sometimes sophisticated analytical methods.

Data collected from patients and/or populations requires the highest ethical, confidentiality, and privacy standards. Techniques such as de-identification, aggregation, data pooling, and other methods to maintain privacy should be in place both locally (at institutions) and centrally (within a central enclave environment). Further advances in obtaining informed consent, such as e-consent, should be up to speed with the pace of technologies used to collect participant data and their willingness to be contacted for future participation in prospective trials.

Integration, Harmonization, and Analysis

Combining data from multiple sources and systems requires harmonization of the data being collected to facilitate future evaluations and analyses. As the ability to integrate data from multiple sources improves, evidence from single data sources may become less preferable. However, the process of data harmonization and curation is currently resource-intensive. Harmonizing CDMs of various networks or data sources also allows researchers to ask specific data questions of potentially much larger (combined) sources of RWD. It is tempting to combine as much data as possible, but it is critical to first identify the purpose, timepoints, and desired impact of data and collect it in a way that minimally burdens patients, healthcare providers, and others at the data generation source. Further, it is important to adhere to the highest possible data quality standards. RWD data from multiple sources often improves data richness, but not all data will be equally useable without proper integration and harmonization.

When combining data from multiple data sources, maintaining privacy and security must be a focus of RWD/RWE efforts. Privacy-Preserving Record Linkage (PPRL) is one way of connecting records that refer to the same individual across

different data sources using secure, pseudonymization processes while maintaining the individual's privacy [24]. NCATS is piloting PPRL technology in the N3C to determine if linking multiple data sets enhances usability of COVID-19 RWD. Identifying high-quality data and mapping their provenance allows for selection of the best and most representative data for multisite analyses.

Despite rapid advances, there are important remaining challenges for broader use of multimodal RWD. For example, errors – inherent to each source and/or arising during harmonization – could confound the process of RWD integration and must be addressed. More robust interoperability for CDMs and codes of conduct to reduce misuse of research findings from RWD and research data in general is necessary. Additionally, newly formed federated or centrally kept data environments must incorporate tools to enable use of completely de-identified or HIPAA-limited data for additional flexibility and adaptation to both institutional and multisite data aggregation needs. Efficient and meaningful integration of data from multiple sources and institutions could help create the next generation of data systems for evidence-based medicine and real-time clinical support [25–27].

RWD can also come from more specific – disease-based – fully consented registries which include strict policies and procedures for data linkages and patient-identified data requests in order to protect patient privacy [28]. In these cases, data harmonization is mostly unnecessary with all data points and variables specified from the outset allowing for more specific questions and outcomes of interest (such as mortality, disease burden, laboratory results, and other variables) to be followed over time. However, as with well-designed clinical trials, research with registries must include *a priori* definition of meaningful outcomes to evaluate and pre-defined analytical methods to prevent or diminish bias.

Novel analytical techniques utilizing AI and ML can provide additional capacity to help analyze large amounts of data and identify "subtle" risk factors that may not be easily identified through more conventional analysis. At the same time, utilizing both automated and manual data review techniques can be used for data quality control and assurance in large multisite data efforts to improve RWD obtained locally and in aggregate with automatically maintained data provenance as a requirement [29].

Diversity, Equity, Inclusion, and Accessibility (DEIA)

Insufficient diversity, equity, inclusion, and accessibility (DEIA) considerations in planning, analyzing, and collecting data can introduce bias and limit the ability of RWD to develop meaningful RWE. Gichoya et al., developed an AI algorithm that accurately identified self-reported race from imaging data only, a phenomenon that could not be replicated by human radiologists and that was not readily explainable [30]. This deliberate example highlights the fact that AI algorithms can make race-specific conclusions based on factors invisible to human evaluators. If those cryptic conclusions are erroneous, or worse, perpetuate biases, they could lead to healthcare decisions that perpetuate health disparities. Of course, these kinds of models may also accelerate beneficial discoveries that would be otherwise difficult to achieve. To ensure the most beneficial patient and public health outcomes, the potential for introduction and amplification of biases or structural inequities must be reckoned with proactively.

Equitable RWE begins with high-quality, representative RWD. Addressing the "digital divide" and fair access to DHTs is also central to discussions of DEIA in RWD. Technologies are becoming more widely used to assess health parameters – e.g., continuous glucose levels, vital signs, and physical activity. Such strategies can adapt to users across the lifespan and expand testing of interventions or therapeutics to more diverse or rural communities or even monitor activity and health in space [31]. Low (e.g. text messaging)- and high-tech (e.g. smart apps) options need to be available and tailored to users to promote equitable access to and representativeness of RWD for all populations intended to benefit.

To advance DHTs, other DEIA-related roadblocks and access questions must be addressed. Services and support tools built and validated to promote inclusivity and equity in data collection and analytical technologies are crucial. Building trust in the use of novel technologies for the benefit of patients and communities is part of the providers' role in the development and implementation process. At the same time, patient/user input in the development and optimization of new technologies plays an important role in improving utilization, user satisfaction, and adherence. The million-dollar question and challenge for CTS researchers remains: how do we align 21st-century RWD with 21st-century medicine for the benefit of all? The answer requires the enterprise to focus on integrated approaches intentionally attuned to DEIA considerations.

Training, Education, and Career Development in RWD/RWE

Training the next generation of scientists is one of the main goals of NIH. For example, the CTSA program supports trainees and scholars through a number of funding opportunities such as the KL2 program (now K12), TL1 program (now T32 pre-doctoral and T32 post-doctoral), as well as other programs such as diversity, reentry, and reintegration supplements [32]. Each program is geared towards a specific phase in the training and career development of the trainee/scholar. Identifying training opportunities, activities and resources tailored to the specific trainee/scholar needs in data management, analytics, and reporting can be crucial in the development of a highly skilled/data-driven workforce of the future. Recently, the CTSA program diversity was published showing areas where we can improve to attract a more diverse workforce in clinical and translational science [33]. Specialized pieces of training in data management and novel

analytical methods using AI and ML can also help in the development and retention of professionals focused on health and related RWD in this fast-growing field. Facilitating an environment where cross-communication between clinicians, data managers, data analysis experts, regulatory authorities, and the community can provide the catalyzing force to maximize efficiency, impact, and return on investment of RWD-oriented projects. As in the device development "world" where clinicians, bioengineers, technology transfer, and marketing experts come together to solve unmet medical needs, we need to develop a similar team science approach where access to specialized resources and expertise in RWD can bring data managers, experts in data analysis (including AI and ML data experts) and others, closer to their clinical/scientific counterparts to help answer meaningful health questions. All needs to be done in parallel with identifying pressing community and population health needs and questions that if answered through RWD could help to significantly improve individual and public health locally and nationally.

Discussion

To make RWD and RWE a reality within the CTSA Program, investigators, research participants, clinicians, patient advocates, funding agencies, regulatory agencies, industry, and many others need to collaboratively identify strategic priorities that maximize impact of data on scientific knowledge and health outcomes. Developing generalized, intervention-agnostic approaches driven by translational science could be an area for consideration by the CTSA consortium. Several key areas already align with CTSA infrastructure and resources (Fig. 2).

Furthermore overlaying "traditional" sources of RWD with other information that can potentially impact health and health outcomes, such as environmental exposures, geolocation, place of residence, economic measures, and other data extracted from publicly and non-publicly available data sets (e.g., from social networking sites such as Facebook, LinkedIn, Instagram, etc., or even data from the "Internet of Things") can add richness (and further complexity) to the data. Yet, these datasets derive from sources not typically thought of as being health-related and that may have different regulations and policies governing their use could create additional challenges in the successfully translating RWD into RWE. Integration of RWD from smaller, well-designed, fit-for-use disease registries as well as other prospective hypothesis-generating types of observational datasets requires *a priori* definition of meaningful outcomes to evaluate using pre-defined analytical methods to minimize bias.

Several limitations must be taken into consideration when utilizing electronic health records data from efforts such as N3C, including overrepresentation of certain populations such as patients with more access to health services, high utilizers of health care, patients with more severe symptoms, and conditions and inpatients. In addition, clinical follow-ups outside of the health system, as it happens in community hospitals or other outside-of-the-health-system settings (e.g. private doctors' offices) can also be missing/not recorded and therefore outcomes are limited to data within the enclave [7]. It is therefore important to acknowledge the limitations of these large RWD sources before making broad population-based conclusions. Data privacy, security, and consent for future contact remain cornerstones to make sure the research performed using RWD through large and broad datasets or smaller and more controlled datasets (e.g., registries) can have all the
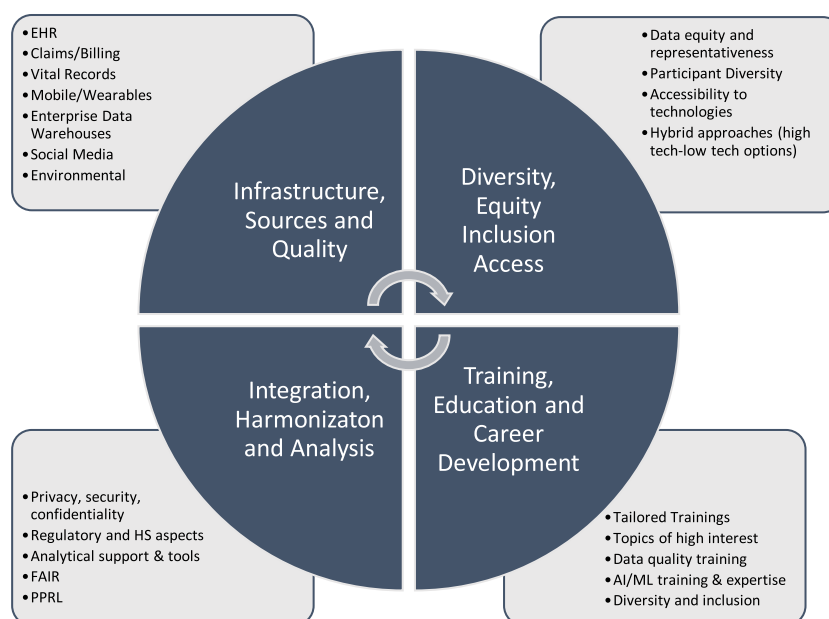
Figure 2. Key areas in real-world data within CTSA programs. D&I = dissemination and implementation; FAIR = findable, accessible, interoperable, reproducible; HER = electronic health records; HS = human subjects; PPRL = privacy-preserving record linkages.

necessary safeguards to allow for an ethical and scientifically rigorous process.

Federated or centralized approaches both provide advantages and disadvantages when it comes to data sharing, integration, harmonization, and quality [29]. For example, efforts to improve data quality in N3C and providing those data to institutions are now part of the feedback received by institutions sharing data in the N3C data enclave. Using federated and centralized approaches to collect RWD can also help with both data quality, by providing feedback to institutions contributing data in a centralized platform/data enclave, or by performing participant screening for future trials in a federated environment, under the right participant consent for future contact. Data standardization at the collection source also represents a major deficiency of large, integrated datasets. Establishing standards for data entry at the outset (during the clinical encounter or shortly after) as well as quality control methods while maintaining data provenance remains critical.

Training the next generation of data managers, statisticians, clinical informaticians, data clinicians and other experts within the CTSA collaborative/team science approach can greatly enhance the capability of the consortium to timely respond to current and future public health needs using RWD. Well-developed, customizable, complementary, and competency-based training programs may represent one of the biggest opportunities for developing the field of RWD and data science within the CTSA consortium. In addition, basic principles of data management and novel methods training could bring great added value to current and future clinical researchers.

Combining these CTSA assets and developing innovative approaches to improve the quality, utilization, and reproducibility of RWD findings under FAIR (Findable, Accessible, Interoperable, and Reproducible) guiding principles [34], can bring new and critical scientifically sound programmatic activities to fruition. Designing approaches to validate and implement new technologies and analytical tools can accelerate RWE advances. Issues of DEIA, including equitable access to technologies and representativeness

of data, need to be addressed from the start and not as an afterthought. Further, applying an equity lens to securely source, integrate, and harmonize clinically relevant, high quality, representative RWD could result in RWE-based approaches that transform healthcare and enhance patient health – as the ultimate goal of all RWE is to answer real-world questions and deliver real-world returns, to all.

References

1. **FDA**. Real-World Evidence: Food and Drug Administration; 2022. [FDA RWE Definitions and Selected Guidance]. Available from: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Accessed December 22, 2022.
2. **Concato J, Corrigan-Curay J.** Real-world evidence - where are we now? *New Engl J Med.* 2022;**386**(18):1680–1682.

3. **FDA**. FDA Issues Draft Guidances on Real-World Evidence, Prepares to Publish More in Future: Food and Drug Administration; 2022. Available from: https://www.fda.gov/drugs/news-events-human-drugs/fda-issues-draft-guidances-real-world-evidence-prepares-publish-more-future. Accessed December 22, 2022.

4. **FDA**. Advancing Real-World Evidence Program: Food and Drug Administration 2022. Available from: https://www.fda.gov/drugs/development-resources/advancing-real-world-evidence-program. Accessed December 22, 2022.

5. **HL7 International. Common Data Models Harmonization - Biomedical Research and Regulation Work Group** 2019. Available from: https://build.fhir.org/ig/HL7/cdmh/#guidance-to-the-readers. Accessed December 22, 2022.

6. **NCATS. National Center for Advancing Translational Sciences, National Institutes of Health** 2023. About the National COVID Cohort Collaborative N3C. 2023. Available from: https://ncats.nih.gov/n3c/about. Accessed March 23, 2023.

7. **Pfaff ER, Girvin AT, Bennett TD**, *et al.* Identifying who has long COVID in the USA: a machine learning approach using N3C data. *Lancet Digit Health.* 2022;**4**(7):E532–E41.

8. **N3C**. N3C Public Health Answers to Speed Tractable Results (PHASTR): National COVID Cohort Collaborative, NCATS, DHHS; 2023. Available from: https://covid.cd2h.org/phastr. Accessed March 15, 2023.

9. **NCATS**. Clinical and Translational Science Awards (CTSA) Program: National Center for Advancing Translational Sciences, National Institutes of Health 2022. Available from: https://ncats.nih.gov/ctsa. Accessed December 22, 2022.

10. **Spengler H, Lang C, Mahapatra T, Gatz I, Kuhn KA, Prasser F.** Enabling agile clinical and translational data warehousing: platform development and evaluation. *JMIR Med Inform.* 2020;**8**(7):e15918.

11. **Schneeweiss S.** Learning from big health care data. *N Engl J Med.* 2014;**370**(23):2161–2163.

12. **Fowe IE, Wallace NT, Rissi JJ.** The evolution and typology of learning health system hospitals and other health information interested hospitals in the US. *Perspect Health Inf Manag.* 2022;**19**(3):1b.

13. **Faden RR, Kass NE, Goodman SN, Pronovost P, Tunis S, Beauchamp TL.** An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. *Hastings Cent Rep.* 2013;**43**(s1):S16–S27.

14. **Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K.** Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med Care.* 2012;**50**(Suppl):S82–101.

15. **Haendel MA, Chute CG, Bennett TD**, *et al.* The national COVID cohort collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assn.* 2021;**28**(3):427–443.

16. **Ta CN, Weng C.** Detecting systemic data quality issues in electronic health records. *Stud Health Technol Inform.* 2019;**264**:383–387.

17. **Jetley G, Zhang H.** Electronic health records in IS research: quality issues, essential thresholds and remedial actions. *Decis Support Syst.* 2019;**126**:113137.

18. **Weiner MG, Embi PJ.** Toward reuse of clinical data for research and quality improvement: the end of the beginning? *Ann Intern Med.* 2009;**151**(5):359–360.

19. **FDA**. FDA approves new use of transplant drug based on real-world evidence: Food and Drug Administration; 2021. Available from: https://www.fda.gov/drugs/news-events-human-drugs/fda-approves-new-use-transplant-drug-based-on-real-world-evidence. Accessed December 22, 2022.

20. **Erdman J, Wolfram J, Nimke D**, *et al.* Lung transplant outcomes based on immunosuppressive regimen at discharge: data from the US scientific registry of transplant recipients (SRTR). *J Heart Lung Transpl.* 2021;**40**(4):S165–S.

21. **Fitzsimmons W, Erdman J, Wolfram J**, *et al.* Lung transplant outcomes based on immunosuppressive regimen at discharge: data from the US scientific registry of transplant recipients (SRTR). *Am J Transplant.* 2021;**21**:800–801.

22. **Cronin KA, Ries LAG, Edwards BK.** Preface. *Cancer.* 2014;**120**(S23):3755–3757.

23. **Penberthy LT, Rivera DR, Lund JL, Bruno MA, Meyer AM.** An overview of real-world data sources for oncology and considerations for research. *CA Cancer J Clin.* 2022;**72**(3):287–300.

24. **Vatsalan D, Christen P, Verykios VS.** A taxonomy of privacy-preserving record linkage techniques. *Information Systems.* 2013;**38**(6):946–969.

25. **Subbiah V.** The next generation of evidence-based medicine. *Nature Medicine.* 2023;**29**(1):49–58.

26. **Califf RM.** Now is the time to fix the evidence generation system. *Clin Trials.* 2023;**20**(1):17407745221147689–12.

27. **Wynia M, Beaty L, Bennett T**, *et al.* Real-world evidence of neutralizing monoclonal antibodies for preventing hospitalization and mortality in COVID-19 outpatients. *Chest.* 2022;**163**(5):1061–1070.

28. **Leppke S, Leighton T, Zaun D**, *et al.* Scientific registry of transplant recipients: collecting, analyzing, and reporting data on transplantation in the United States. *Transplantation Reviews.* 2013;**27**(2):50–56.

29. **Pfaff ER, Girvin AT, Gabriel DL**, *et al.* Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative. *J Am Med Inform Assoc.* 2022;**29**(4):609–618.

30. **Gichoya JW, Banerjee I, Bhimireddy AR**, *et al.* AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health.* 2022;**4**(6):E406–E14.

31. **Crucian B, Valentine R, Calaway K**, *et al.* Spaceflight validation of technology for point-of-care monitoring of peripheral blood WBC and differential in astronauts during space missions. *Life Sciences in Space Research.* 2021;**31**:29–33.

32. **NCATS. National Center for Advancing Translational Sciences, National Institutes of Health. Training and Career Development Programs** 2018. https://ncats.nih.gov/ctsa/about/training/programs. Accessed March 23, 2023.

33. **Rubio M, Baker HL, Doyle JM.** Demographic composition of national institutes of health clinical and translational science awards (CTSA) program principal investigators, scholars, and trainees. *J Clin Transl Sci.* 2023;**7**(1):e19.

34. **Wilkinson MD, Dumontier M, Aalbersberg IJ**, *et al.* The FAIR guiding principles for scientific data management and stewardship. *Scientific Data.* 2016;**3**(1):160018.