Journal of the American Medical Informatics Association, 29(4), 2022, 581–584
https://doi.org/10.1093/jamia/ocac024
Editorial

Editorial

Research data warehouse best practices: catalyzing national data sharing through informatics innovation

Shawn N. Murphy[1,2], Shyam Visweswaran [3,4], Michael J. Becich [3,4], Thomas R. Campion [5,6], Boyd M. Knosp[7], Genevieve B. Melton-Meaux [8,9], and Leslie A. Lenert[10,11]

[1]Research Information Science and Computing, Mass General Brigham, Somerville, Massachusetts, USA, [2]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA, [3]Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA, [4]Clinical and Translational Science Institute, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA, [5]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA, [6]Clinical and Translational Science Center, Weill Cornell Medicine, New York, New York, USA, [7]Roy J. and Lucille A. Carver College of Medicine and the Institute for Clinical & Translational Science, University of Iowa, Iowa City, Iowa, USA, [8]Department of Surgery, University of Minnesota, Minneapolis, Minnesota, USA, [9]Institute for Health Informatics (IHI), University of Minnesota, Minneapolis, Minnesota, USA, and [10]Biomedical Informatics Center (BMIC), Medical University of South Carolina, Charleston, South Carolina, USA and [11]Health Sciences South Carolina, Columbia, South Carolina, USA

Corresponding Author: Shyam Visweswaran, MD, PhD, University of Pittsburgh, The Offices at Baum, 5607 Baum Blvd., Suite 523, Pittsburgh, PA 15206, USA; shv3@pitt.edu

Received 4 February 2022; Editorial Decision 8 February 2022; Accepted 14 February 2022

Key words: research patient data repository, electronic health records, cloud technology, research data warehouse, Fast Healthcare Interoperability Resources (FHIR), Trusted Exchange Framework and Common Agreement (TEFCA)

Research Patient Data Repositories (RPDRs) have become essential infrastructure for traditional Clinical and Translational Science Award (CTSA) programs and increasingly for a wide range of research consortia and learning health system networks.[1–5] Almost every institution with a CTSA or Clinical Translational Research (CTR) program (found in states with lower amounts of National Institutes of Health funding) hosts an RPDR for the benefit of affiliated researchers. These repositories aim to enable healthcare research based upon the patient populations they serve. Within the institution, RPDRs are valuable for a range of research activities. They are used to identify patients for clinical trial recruitment using privacy-preserving methods to search and extract specific cohorts of trial-eligible patients.[6] They aid in developing and validating computable phenotypes that are increasingly important for accurately identifying patient cohorts in a reproducible fashion.[7] RPDRs provide de-identified patient data for population health research and support a growing body of artificial intelligence to predict patient outcomes.[8] Further, clinical studies can often be simulated using data from an RPDR.[9] Beyond the institution, aggregates of de-identified datasets from multiple institutions linked with privacy-preserving hash codes provide an unprecedented opportunity to conduct population health research, perform comparative effectiveness analyses and apply artificial intelligence methods over large and diverse populations.[10] The data contained within the RPDR vary across institutions, based on institutional strengths and weaknesses; the papers published in this issue reflect that variability (see Table 1). Data are commonly acquired from local electronic health records (EHRs) and other clinical information systems that capture information during clinical care. Data consist of diagnoses, problem lists, procedures, prescribed medications, laboratory exams, and many types of free-text reports. Overall, the benefits of the RPDR for accelerating translational research can be significant. For example, at Harvard, in 2006, between $94 and $136 million in annual research funding was linked to the use of data from the RPDR.[11]

This focus issue of *JAMIA* describes some of the current research, approaches, applications, and best practices for RPDRs comprising 11 research and applications papers[12–22] and 4 case reports[23–26] (see Table 1). Ten of the papers describe RPDRs, and 5 describe governance, regulatory and technical issues related to RPDRs. The scope of the papers ranges from a single site to regional to US-wide (2, 7, and 6 articles, respectively). The number of patients in the RPDRs ranges from 125K to 24M, of which 7 include privacy-preserving features, and 1 contains data from natural language processing (NLP). Commonly used data models (CDMs) in the RPDRs include the Observational Medical Outcomes Partnership (OMOP) CDM,[27] the National Patient-Centered Clinical Research Network's (PCORnet's) CDM,[5] and the Accrual to Clinical Trials (ACT)[4] and Tri-

Table 1. Selected features of RDPRs and practices related to RPDRs

			National EHR data sharing			Special features					Common data models supported			
Lead author	Patients	Scope	CTSA	PCORI	All of Us	Cancer Registry	Claims	NLP	Privacy Preserving	SDoH	OMOP CDM	PCORnet CDM	ACT, TriNetx	Comments
Hogan	17.2M	Regional	X	X		X	X		X	X	X	X	X	Early SDoH data (geocoding)
Pfaff	6.4M	US wide	X	X	X				X		X	X	X	Focus on COVID-19 data quality
Waitman	24M	Regional		X		X	X					X	X	Cloud RPDR for multi-institutional collaboration
Loomba	Not RDPR	Regional	X						X					Regional NIH data commons resource
Meeker	600K	Regional	X						X	X				Public health system data governance
Visweswaran	5M	Single site	X	X	X		X		X		X	X	X	RPDR case study
Khan	224K	US wide			X				X	X	X			Height and weight normalization
Barnes	400K	US wide	X											Federated collection
Campion	Not RDPR	Regional	X	X									X	Tools matching investigators approaches
Castro	125K	Regional	X					X	X				X	i2b2 based biobank linking multiple data types
Nelson	Not RDPR	US wide	X								X	X	X	RIC's EHR cohort assessment process
Walji	4.4M	US wide											X	Multi-institutional dental data warehouse
Knosp	Not RDPR	US wide	X	X										National survey of best practices
Kahn	7.3M	Regional	X											Large-scale migration of RPDR to the cloud
Walters	Not RDPR	Single site	X											Model for governance of RPDR data requests

NIH: National Institutes of Health; PCORI: Patient-Centered Outcomes Research Institute; RIC: Recruitment Innovation Center; SDoH: social determinants of health .

NetX[9] CDMs that are based on the Informatics for Integrating Biology & the Bedside (i2b2) platform.[7]

A key emerging innovation is the adoption of cloud technology for RPDRs. Knosp et al[21] surveyed 20 CTSA hubs and found that 2 hubs had completely migrated their RPDRs to the cloud and several others were considering moving their RPDRs to the cloud. Three other papers describe approaches, advantages, and challenges of implementing RPDRs in the cloud.[14,15,17] Barnes et al[17] offer an approach to RPDRs that is focused on sharing and integrating data for large-scale research projects, using the Amazon Web Services (AWS) to create a distributed data commons. Common workspaces can be created where datasets from multiple sources can be accessed through common authentication and analyzed with preconfigured tools, including Jupyter and R notebooks. A limitation of this approach is that the researchers must harmonize data across the different data models, although the datasets contain common data elements, use controlled vocabularies, and adhere to other standards.

Anticipating what may become a common architecture for RPDRs, Kahn et al[15] describe opportunities and challenges of migrating a large RPDR with administrative, clinical, genomic, and population-level data from on-premises infrastructure to the Google Cloud Platform. While the cloud offers advantages such as inexpensive storage, automatic backups, and secure analytic environments, a variety of issues have to be carefully evaluated to enable smooth migration from on-premises infrastructure to the cloud. The Extract, Transform and Load (ETL) processes may need redesigning due to movement of large data volumes across routers and networks, and realizing cost savings requires organizational changes that may be difficult to implement.

Waitman et al[14] describe how cloud technology facilitates multi-institutional research. The Greater Plains Collaborative (GPC) Reusable Observable Unified Study Environment (GROUSE) is implemented on AWS and integrates EHR, claims, and tumor registry data from 7 healthcare systems. Using GROUSE, the authors demonstrate that clinical data may sometimes allow for more precise inferences than coded data; for example, obesity is more accurately inferred from body mass index measures compared to diagnostic (ICD-10) codes. However, comorbidities associated with obesity such as diabetes and sleep apnea are more accurately inferred from diagnostic codes. This article outlines GROUSE's governance, architecture, and compliance components and describes interagency agreements that facilitate health system collaboration, and that ensures security and privacy policies align with federal requirements.

The papers in this issue aptly illustrate that RPDRs are a diverse, vibrant ecosystem that collaboratively and progressively enhances national health research infrastructure. This infrastructure has been invaluable in investigating the COVID-19 pandemic.[8,28–30] What are the future directions for RPDRs? Assuming support for the current funding for data curation at individual site RPDRs is continued by the 2 primary funding agencies for these activities, the Patient-Centered Outcomes Research Institute (supports PCORnet)[5,31] and the National Center for Accelerating Translational Science (NCATS) at the National Institutes of Health (supports N3C[2] and ACT[4]) one would expect expansion in the depth and breadth of data available in these networks. PCORnet[32] and N3C are in the process of expanding the deployment of privacy-preserving record linkage systems that will allow the integration of data from individual RPDRs across networks using the encrypted hashed identifiers. Even so, the data in RPDRs could be broader and more representative of the national healthcare system. Advances in application programming interfaces to access data EHRs brought about by the 21st Century Cures Act[33] and expansion of the United States Core Data for Interoperability (USCDI) Standards[34] to reflect research data needs may make it possible for a broader range of health systems to contribute data to RPDRs. One area that requires further policy development is expanding health information exchange for research. Currently, the governance for the National Health Information Network (NHIN) acknowledges the importance of health information exchange for research, but does not support it within its Trusted Exchange Framework and Common Agreement (TEFCA).[35] Access to data from multiple providers through a TEFCA process for research studies could remove many gaps that limit the completeness of patient-level health information in RPDRs. However, further policy development is needed by the Office of the National Coordinator for Health Information Technology (ONC) and TEFCA's Recognized Coordinating Entity (the Sequoia Project), to achieve this capability.

Paradoxically, national standards that improve access to data for research from the health system might seem to obviate the case for RPDRs, where they may seem less needed when EHR data are universally available in standardized formats and by protocols such as bulk Fast Healthcare Interoperability Resources (FHIR).[33] In this setting, funders might want to centralize data resources to reduce costs, creating a monoculture based on cloud infrastructure. The N3C Data Enclave illustrates this approach on the cloud, which uses central resources to normalize data and provide access to data sets and analytics in a cloud environment operated by a government contractor.[2] This "monoculture," particularly if controlled by a private contractor, might stifle the types of innovative work detailed in this issue. Furthermore, much of the benefit of the RPDR is achieved through local hospital connections. RPDRs greatly assist recruitment of patients for clinical trials through processes local to the hospitals where the trials are being conducted. Engagement of clinical researchers from hospitals and medical centers occurs mostly at the local level, where they can decide on priorities for data ETL and data aggregation. Taking Protected Health Information (PHI) outside of hospital entities is greatly limited by the Health Insurance Portability and Accountability Act but necessary to validate data in the EHR through chart review. A centralized architecture may or may not be more efficient but is certainly less diverse and provides fewer opportunities for research in RPDR methods than alternative federated approaches used in PCORnet and ACT.

Further, many technical challenges remain in the curation and delivery of healthcare system data for research which might be best addressed initially in a diverse competitive ecosystem and greatly enhance the capabilities and potential health impacts of RPDRs. Further development is required to integrate NLP technology, and corresponding integration of NLP abstracted data into RPDRs requires further development. While many NLP systems are being developed in the context of RPDRs, there are few standards for representing data that is the product of NLP systems. Broad dissemination of NLP technologies may require further algorithm research, standardized tool kits, and standards for target concepts for abstraction. NLP abstracted data, being derived from algorithms, may also require the representation of the precision of abstraction within RPDRs to fully support its use in research studies. Integrating EHR data with hospital clinical trials and clinical studies is a further area of research that requires new methods and development. Such methods may overcome some of the limitations in data collection from case report forms and provide new ways to conduct the studies.

The representation of genomic data with clinical data in RPDRs is another area where additional development is needed. Papers published in this issue describe the use of i2b2 ontologies for the representation of genomic data variation and association data.[18,22] The size and complexity of representation of gene variant data and single nucleotide polymorphism associational data as well as other 'omics' data, in association with clinical data on phenotypes, makes standardization of data representations for queries difficult. While there is evolving work

on architectures[36] and standards supporting this,[37] the models for representation may need further maturation to support standardized data queries and federation of data across RPDRs in a network.

Overall, the collection of papers in this issue demonstrates the value of a diverse program supporting institutional level RPDR development. Ongoing support for diversity in RPDRs at individual institutions creates opportunities to advance that field that would be difficult to achieve in a more centralized monoculture. As also shown in the paper by Pfaff et al,[20] integration of these data resources, when necessary for specific national-level programs, is feasible and strengthens the ecosystem of RDPRs as a whole.

FUNDING

This work was funded in part by the University of Rochester Center for Leading Innovation and Collaboration (CLIC), under Grant U24TR002260.

AUTHOR CONTRIBUTIONS

All authors contributed to the manuscript, made critical revisions, and approved the final version for submission.

ACKNOWLEDGMENTS

We thank Dr. Suzanne Bakken for the insightful comments and suggestions on the draft manuscript.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

No new data were generated or analyzed in support of this research.

REFERENCES

1. Brat GA, Weber GM, Gehlenborg N, *et al.* International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *NPJ Digit Med* 2020; 3: 109.
2. Haendel MA, Chute CG, Bennett TD, *et al.*; N3C Consortium. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* 2021; 28 (3): 427–43.
3. Denny JC, Rutter JL, Goldstein DB, *et al.*; All of Us Research Program Investigators. The "All of Us" research program. *N Engl J Med* 2019; 381 (7): 668–76.
4. Visweswaran S, Becich MJ, D'Itri VS, *et al.* Accrual to clinical trials (ACT): a clinical and translational science award consortium network. *JAMIA Open* 2018; 1 (2): 147–52.
5. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc* 2014; 21 (4): 578–82.
6. Claerhout B, Kalra D, Mueller C, *et al.* Federated electronic health records research technology to support clinical trial protocol optimization: evidence from EHR4CR and the InSite platform. *J Biomed Inform* 2019; 90: 103090.
7. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. *J Am Med Inform Assoc* 2012; 19 (2): 181–5.
8. Weber GM, Zhang HG, L'Yi S, *et al.*; Consortium For Clinical Characterization Of COVID-19 By EHR (4CE). International changes in COVID-19 clinical trajectories across 315 hospitals and 6 countries: retrospective cohort study. *J Med Internet Res* 2021; 23 (10): e31400.
9. Topaloglu U, Palchuk MB. Using a federated network of real-world data to optimize clinical trials operations. *JCO Clin Cancer Inform* 2018; 2: 1–10.
10. Hripcsak G, Ryan PB, Duke JD, *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci USA* 2016; 113 (27): 7329–36.
11. Nalichowski R, Keogh D, Chueh HC, Murphy SN. Calculating the benefits of a Research Patient Data Repository. *AMIA Annu Symp Proc* 2006; 2006: 1044.
12. Visweswaran S, McLay B, Cappella N, *et al.* An atomic approach to the design and implementation of a research data warehouse. *J Am Med Inform Assoc* 2022; 29 (4): 601–8.
13. Khan MS, Carroll RJ. Inference-based correction of multi-site height and weight measurement data in the All of Us research program. *J Am Med Inform Assoc* 2022; 29 (4): 626–30.
14. Waitman LR, Song X, Walpitage DL, *et al.* Enhancing PCORnet Clinical Research Network data completeness by integrating multistate insurance claims with electronic health records in a cloud environment aligned with CMS security and privacy requirements. *J Am Med Inform Assoc* 2022; 29 (4): 660–70.
15. Kahn MG, Mui JY, Ames MJ, *et al.* Migrating a research data warehouse to a public cloud: challenges and opportunities. *J Am Med Inform Assoc* 2022; 29 (4): 592–600.
16. Nelson SJ, Drury B, Hood D, *et al.* EHR-based cohort assessment for multicenter RCTs: a fast and flexible model for identifying potential study sites. *J Am Med Inform Assoc* 2022; 29 (4): 652–9.
17. Barnes C, Bajracharya B, Cannalte M, *et al.* The Biomedical Research Hub: A federated platform for patient research data. *J Am Med Inform Assoc* 2022; 29 (4): 619–25.
18. Castro VM, Gainer V, Wattanasin N, *et al.* The Mass General Brigham Biobank Portal: an i2b2-based data repository linking disparate and high-dimensional patient data to support multimodal analytics. *J Am Med Inform Assoc* 2022; 29 (4): 643–51.
19. Loomba JJ, Wasson GS, Chamakuri RKR, *et al.* The iTHRIV Commons: a cross-institution information and health research data sharing architecture and web application. *J Am Med Inform Assoc* 2022; 29 (4): 631–42.
20. 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): 609–18.
21. Knosp BM, Craven CK, Dorr DA, Bernstam EV, Campion TR Jr. Understanding enterprise data warehouses to support clinical and translational research: enterprise information technology relationships, data governance, workforce, and cloud computing. *J Am Med Inform Assoc* 2022; 29 (4): 671–6.
22. Campion TR, Sholle ET, Pathak J, Johnson SB, Leonard JP, Cole CL. An architecture for research computing in health to support clinical and translational investigators with electronic patient data. *J Am Med Inform Assoc* 2022; 29 (4): 677–85.
23. Hogan WR, Shenkman EA, Robinson T, *et al.* The OneFlorida Data Trust: a centralized, translational research data infrastructure of statewide scope. *J Am Med Inform Assoc* 2022; 29 (4): 686–93.
24. Walji MF, Spallek H, Kookal KK, *et al.* BigMouth: development and maintenance of a successful dental data repository. *J Am Med Inform Assoc* 2022; 29 (4): 701–6.
25. Meeker D, Fu P Jr, Garcia G, *et al.* Establishing a research informatics program in a public healthcare system: a case report with model documents. *J Am Med Inform Assoc* 2022; 29 (4): 694–700.
26. Walters KM, Jojic A, Pfaff ER, *et al.* Supporting research, protecting data: one institution's approach to clinical data warehouse governance. *J Am Med Inform Assoc* 2022; 29 (4): 707–12.
27. Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015; 216: 574–8.
28. Sharafeldin N, Bates B, Song Q, *et al.* Outcomes of COVID-19 in patients with cancer: report from the National COVID Cohort Collaborative (N3C). *J Clin Oncol* 2021; 39 (20): 2232–46.
29. Turk MA, Landes SD, Formica MK, Goss KD. Intellectual and developmental disability and COVID-19 case-fatality trends: TriNetX analysis. *Disabil Health J* 2020; 13 (3): 100942.
30. Visweswaran S, Samayamuthu MJ, Morris M, *et al.* Development of a coronavirus disease 2019 (COVID-19) application ontology for the Accrual to Clinical Trials (ACT) network. *JAMIA Open* 2021; 4 (2): ooab036.
31. Forrest CB, McTigue KM, Hernandez AF, *et al.* PCORnet® 2020: current state, accomplishments, and future directions. *J Clin Epidemiol* 2021; 129: 60–7.
32. Bian J, Loiacono A, Sura A, *et al.* Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. *JAMIA Open* 2019; 2 (4): 562–9.
33. Mandl KD, Gottlieb D, Mandel JC, *et al.* Push button population health: the SMART/HL7 FHIR bulk data access application programming interface. *NPJ Digit Med* 2020; 3 (1): 151.
34. United States Core Data for Interoperability (USCDI). https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi. Accessed May 5, 2022.
35. ONC TEFCA RCE. 2019. https://rce.sequoiaproject.org/. Accessed May 5, 2022.
36. Murphy SN, Avillach P, Bellazzi R, *et al.* Combining clinical and genomics queries using i2b2—three methods. *PLoS One* 2017; 12 (4): e0172187.
37. Alterovitz G, Warner J, Zhang P, *et al.* SMART on FHIR genomics: facilitating standardized clinico-genomic apps. *J Am Med Inform Assoc* 2015; 22 (6): 1173–8.