Economical Utilization of Health Information with Learning Healthcare System Data Commons

By Thomas O'Hara, PhD; Anil Saldanha, MS; Matthew Trunnell, MSc; Robert L. Grossman, PhD; Bala Hota, MD; and Casey Frankenberger, PhD

High-Level Statement

The ability to inventory institutional data assets, design research studies, and to share and analyze data has proven challenging for healthcare systems that are focused on delivery of care. Here, the use of data commons is described as a low-cost and low-staff approach to data utilization to facilitate quality improvement and research in such cases.

Abstract

Finding, accessing, sharing, and analyzing patient data from a clinical setting for collaborative research has continually proven to be a challenge in healthcare organizations. The human and technological architecture required to perform these services exist at the largest academic institutions but are usually under-funded. At smaller, less academically focused healthcare organizations across the United States, where the majority of care is delivered, they are generally absent. Here we propose a solution called the Learning Healthcare System Data Commons where cost is usage-based and the most basic elements are designed to be extensible, allowing it to evolve with the changing landscape of healthcare. Herein we also discuss our reference implementation of this platform tailored specifically for operational sustainability and governance using the data generated in a hospital setting for research, quality, and educational purposes.
Information management professionals within healthcare organizations navigate a high degree of complexity for each project and for each data source used for research and quality improvement services. Data and data policy must be governed tightly, consistently, and transparently to meet the expectations of patients and to comply with the high ethical and legal standards in the healthcare industry. Even prior to the pandemic, access and sharing of patient data has been of paramount importance to assess current status of medical knowledge, as well as to accelerate clinical research related to diagnosis, prognosis, and therapeutic intervention in the context of cancer care; complex, or rare disease; and in the face of rapidly changing technologies for telehealth, surveillance, engagement, and intervention.

The COVID-19 pandemic has highlighted the need for unified and harmonized data sets. Early in the pandemic, an urgent need to evaluate outcomes related to COVID-19, efficacy of treatments, risks for severe disease, and health equity differences was identified. The diversity of patients’ current health and medical history relative to various viral strains presents issues for all medical research institutions both in the capacity to access data in real time and the costs to maintain such flexible, agile analytics environments. The need to search, access, analyze, and share medically related data of patients in a manner that is reliable and secure has required separate teams to architect and manage data flow and to manage data governance. Additionally, interfaces that are accessible to experienced care providers/researchers are needed to integrate and contextualize results to the larger community.

Here we present the Learning Healthcare System Data Commons (LHSDC), a cloud-based orchestration environment and governance framework that meets the highest standards for security, cost efficiency, and platform extensibility, enabling scalable access to a FAIR computing environment. This platform is open source, pay-as-you-go, and cost efficient, making it interoperable across an ecosystem of National Institutes of Health (NIH)-supported commons-based data enclaves and supportive of big data initiatives within academic or public-private partnerships. The LHSDC attempts to realize the re-usability and easy upkeep of data-as-a-service made possible with a data commons.

**Methods**

**The Gen3 Platform: A Scalable Open-Source Platform for Data Storage and Governance**

The Gen3 platform ([https://gen3.org/](https://gen3.org/)), an open source data commons framework, is cloud-native and makes central control of data access and data use possible. It generates a FAIR environment. The Gen3 platform is currently used in different ways by different research communities. For example, some Gen3 commons are focused on data types, such as the BloodPAC Data Commons for liquid biopsy data and the National Institute of Biomedical Imaging and Bioengineering Medical Imaging and Data Resource Center (MIDRC) data commons.**
adoption, as this framework also serves public-private partnerships. The LHSDC differentiates itself from previous work in two important ways: 1) It is the first data commons in general, and Gen3 data commons in particular, designed to facilitate quality improvement and research for a healthcare provider. 2) It is also the first Gen3 implementation on the Microsoft Azure platform and was developed in collaboration with Microsoft.

Data volumes per site can range by orders of magnitude depending on the number of patients seen in practice as well as the nature of the data being investigated. A small practice can see hundreds of patients, but the largest hospital systems have seen millions of patients in aggregate. The overall data storage burden is not determined by patient count alone, but rather what data files are included. Typical data may include image files and genomic data in addition to medical records, and so even individual files can be “large.” This creates specific challenges in how to house the data in a manner that is readily accessible and to “focus” the data to deliver meaning at manageable storage volumes. The Gen3 commons framework links out to data objects for storing images, genomic data, wearable data, and other large files, in whatever commodity storage is most convenient (e.g., Azure Blob storage, Amazon Web Services (AWS) Simple Storage Service (S3) buckets). A Gen3-hosted Postgres database houses structured data and metadata, which consumes orders of magnitude less storage than the externalized raw data and makes it findable through a graph data model. Facets of the metadata are exposed for interactive exploration using an Elasticsearch index (pre-populated as configured). All metadata is granularly accessible, monitored, and governed through application programming interface (API)-based calls in the Jupyter Notebook. The centralization of the infrastructure of these resources means that the whole environment can be managed by a small team or even one individual. Additionally, the analytic team can be greatly streamlined, as they can reuse data sources, methods, and code.

**Implementation**

**Data Assets, and Assets Loaded (Counts of Files by Type)**

Rush University, operating as a major medical hospital in a diverse major city, is home to diverse troves of multimodal (i.e., wholly different information categories: medical images, genomic sequences, and clinical records) diagnostic and medical treatment outcomes data assets. Historically, much of this data has been stored in on-premises data centers. More recently, data has been stored in external data centers and in the cloud.

Rush Research Analytics provides access to clinical data records from the electronic health record (EHR). Medical images from CT/PET/MR, etc., are available for research (via a picture archiving and communication system (PACS) server in midst of migration to Flywheel). Biospecimens at Rush are extensive. For example, in the last two years, tens of thousands of COVID-19 sputum samples and thousands of cancer biopsy specimens have been collected.

This website is using cookies

This website uses cookies to ensure you get the best experience on our website. To learn more about cookies and how we use them, please view our privacy policy.

Agree
sensors involved in patient care. Rush collects and analyzes data related to social determinants of health, an active area of ongoing research funded by the Robert Wood Johnson Foundation.

The data assets described above are all related to individual patients, which can be de-identified carefully following contemporary best practices for personal health information definition and removal. Importantly, the assets are stored and managed separately and, thus far, there has been no coordinated effort to bring together a holistic view of care and outcomes with an eye toward better care with efficient use of resources. However, this cohesion is possible at Rush and similar medical centers, facilitated by efforts like the LHSDC.

**Deployed Components and Improvements Advanced by this Implementation**

The LHSDC is the first deployment of the Gen3 framework in the Microsoft Azure Cloud. Gen3 microservices were forked from open-source versions available from University of Chicago’s Center for Translational Data Science (github.com/uc-cdis), and they were modified to include support for Azure-specific resources. This includes support for Azure Active Directory for user authentication and authorization (analogous to AWS Identity and Access Management (IAM)), and also support for Azure Blob storage (analogous to AWS S3). We strengthened the security of communication between services and the backend database by introducing support for secure sockets layer (SSL) with Postgres. These improvements have been merged back into the main/master branches of the open repositories for the Gen3 components (specifically, indexd, sheepdog, tube, and fence). (Visit gen3.org and github codebase ([https://github.com/uc-cdis/gen3.org](https://github.com/uc-cdis/gen3.org)) for detailed information on these microservices and the roles they play in a Gen3 data commons.)

We introduce a novel deployment paradigm that utilizes Terraform (infrastructure as code language and framework by HashiCorp, [https://www.terraform.io/](https://www.terraform.io/)) and/or Azure DevOps pipelines to stamp out instances of the commons. To orchestrate the configuration of Gen3 components (aka microservices), a high-level-definition (HLD) is supplied (in yaml) that defines which services to deploy, which tagged version to use, and where to find that build. The specific build can be a custom version, as was the case during development of the LHSDC before improvements/updates for Azure listed above were publicly available. These custom images were built by a pipeline after testing and were stored in Azure Container Registry, where they could be referenced by an HLD configuration. Now that the microservices with the updates are public, the images in a configuration can be references to tagged public builds (quay.io/organization/cdis). Open-source publication of HLD orchestrations and Terraform scripts for Azure deployments, to be made available among Gen3 resources (github.com/cdis), are forthcoming. Our specific realization of the HLD concept for continuous deployments is based on the general-use freeware offered by Microsoft as “bedrock” ([https://github.com/microsoft/bedrock](https://github.com/microsoft/bedrock)).
Data Governance and Cybersecurity

There are numerous opportunities to perform research and collaborate within or across an organization; however, prior to the data commons implementation, each new project requires its own new data preparations and its own new governance. Data can be pulled from the electronic health record into a clinical data warehouse, processed to remove any identifying information, and potentially combined with other data resources and prepared for research investigation. These activities currently operate on a per-project manner with the data decentralized and put into the hands of investigators at the earliest stage. The current procedures for sharing prepared research data can vary widely from project to project. Access to files for a research team is managed ad-hoc, with users granted credentials to view to files in shared directories (e.g., Google Drive, Dropbox, Box) or with users sent email attachments. Though these various modes of sharing can be accomplished securely, the lack of centralized and uniform control over the sharing procedure prevents meaningful audits and provides opportunity for access leakage or data leakage. The data can be copied, and the copies and user list is not tracked. This creates a problem for data governance, as usage and compliance are largely self-reported. Because of the transition to cloud computing, all of this can be done in an environment that is accessible to researchers but controlled centrally. With highly centralized data sharing, research analytics can become a self-service enterprise, with authorized scientists logging in to view and work with already available data. The data will have already been approved by institution review board (IRB)/governance entities. The oversight of the resources will be operated by a small team of data software developers whose operational practices are monitored and approved by institutional cybersecurity and treated very much like any other website (albeit one with sensitive data).

Governance pre-approval of data for centralization in a general purpose (or narrow focus) database is a growth and maintenance process, with small but regular updates to be made. This is in contrast to the current state of data request processes, where decentralized data requests are considered in isolation from each other in a case-by-case manner. Even if there exist redundancies or overlap, the cases each require separate and end-to-end judgements without coordination around a unifying data agenda and without exploiting efficiencies of a once-for-all approach for shared needs. The use of the cloud also enables centrally managed authentication and authorization, which eliminates the need for additional outlets or copies of data. This results in an overall more closed and, therefore, more secure, system. This data flow is depicted in Figure 1A, from research concept to centralized sharing.

Data Sharing and Cohort Discovery

Commonly, biomedical research is a multi-institutional enterprise. Data use and data sharing agreements can be slow to take full shape and take time to gain full support and buy-in. Following the establishment of such agreements among parties, there remains the challenge of data harmonization, which can often take time and require debate and eventual endorsement.
With centralized data aggregation and sharing across institutions from the start, as has been the case for recent COVID-19-related databases, it is possible to provide a single shared resource with a single set of standards and interrelationships among data in a data model. The data model can be managed as open-source code and versioned as it evolves. The debate and buy-in over this standard presentation of the data in this data model is integral to its very existence as code in a shared repository (e.g., github) where users can create branched/forked versions with special features that can evolve on their own or be merged back into the main (or master) branch once consensus is reached and the model is tested by the community of interested and authorized developers.

**Streamlined Logistics and Pay-As-You-Go Economics**

Currently, research projects can begin with IRB approval of the proposed undertaking. This is followed by a formal data request, which is queued for data extraction (and de-identification). Once the research analytics team is engaged on a request, data is identified and accessed, and an ETL (extract, transfer, load) operation is performed. The resulting records are presented in a usable research database. This process is often iterative; once analysis begins in earnest, non-obvious deficiencies in the original request may be revealed. Scientists connect their database with a provisioned and reserved virtual machine where allowed site licensed software is installed and/or custom code is developed, tested and refined, and executed. Again, this is iterative, and the data request or even the IRB may need updates as the research matures toward conclusions.\(^9\)

The above process takes weeks to months before actual science can begin (Figure 1B). The back-and-forth communication between researchers and analytics and the IRB is typically accomplished over email and scheduled meetings, which can be slow. The data extraction results in a static and single-use database that is typically highly specific and not readily reusable or applicable to other projects. The provisioned virtual machine(s) are reserved and cost money even when idle.

In contrast, with the LHSDC, general purpose (de-identified) multimodal data can be presented in a general purpose data model to a general audience of scientists, exploration can begin right away for those with access. The IRB approval to serve this data to the intended audience can be earned early, once and for all, and approved by any other comprehensive research data governance body. A sufficiently large and general purpose multimodal dataset will not suffer from incompleteness or deficiencies for preliminary exploration. Any need for more specialized, PHI-containing, or otherwise excluded information will reveal itself upon preliminary exploration, and not after rounds of iteration from (request, acquisition, analysis, repeat). The general purpose database can be explored where it sits, in the cloud and utilizing cloud-native processing power with preinstalled software and off-the-shelf analytics (e.g., Jupyter Notebook). This compute resource can be a shared cluster, provisioned when needed with pay-as-you-go.
Rush University Reference Implementation

At Rush University, to improve the auditability, sustainability, and efficiency of research informatics, we have developed a reference implementation for data storage and access using Gen3. Our implementation consists of a three-tiered model for data governance and operations: 1) a data governance committee, 2) a research informatics core, 3) a Gen3 learning healthcare system data commons. We have observed an increase in projects that have come to our organization through both academic and commercial channels. A data governance committee was developed to have a diverse and representative team review each project with respect to legal, ethical, and practical aspects and to ensure documentation and consistency of the projects approved, the parameters by which the projects are defined, and, where applicable, the standards to which the project and involved parties will be held. Once projects are defined, documented, and ready for operationalization, our internal team, defined as the research informatics core, will either provide access to data sets already contained in the data commons or load data into the data commons and provide access to the newly loaded data. These data can be as diverse as medical history data derived from our electronic health records combined with raw and processed files related to genomics or imaging. Multimodal data has long been a Holy Grail for precision medicine analytics, but the diversity of data elements can present problems for housing data in traditional databases. The Gen3 data commons framework is an ideal intermediary for multimodal data. Essentially, it holds data in a data lake, has a searchable index of metadata for each data point, and holds each data element using interoperable data formats where possible and where there exists some degree of consensus for what the interoperable format can and should be. In the Rush pilot, we have included EHR patient data, genomics files, pathology files, PACS image files, and biorepository data in our initial instantiation (Figure 2A). We have plans to incorporate IOT hub reporting and clinical trial management system (CTMS) integration over time and as usage becomes more widespread.

One example of the power of this approach is the development and use of a common data model for clinical data. A common trope is the phrase “if you have seen one instance of Epic, you have seen one instance of Epic,” which alludes to the significant customizations for each institution using EHRs. This can be largely resolved by the use of a common data model like OMOP (Observational Medical Outcomes Partnership), PCORI (Patient-Centered Outcomes Research Institute), I2B2 (Informatics for Integrating Biology & the Bedside), and others. We have selected Fast Healthcare Interoperability Resources (FHIR) as the basis for our EHR data due its relatively low level of loss and its ability to be converted to other data models such as OMOP or PCORI with minimal loss of resolution. FHIR presents numerous advantages in that it is a required resource for many hospital organizations and that it can be accessed through a pull mechanism across organizations. In addition to the standard Gen3 components, which include a user interface or website front end, a flexible data model (as code) with graphical representation, faceted search cohort discovery (rapid Elasticsearch), a SQL command window, and workspaces that allows for custom analytics using custom code (Python, R, etc.), the LHSDC
that can be important for controlling and recouping (or directly charging) for operating costs. Data exploration and analysis interfaces are shown in Figure 2B.

The LHSDC project has developed a data model that makes use of standard naming from HL7.org to facilitate data sharing outward via the FHIR API. The exploration tabs for faceted search in Gen3 is customizable, and we have designed tabs with features and plots tailored to an idea of user types (e.g., researchers interested in genomics versus those interested in social determinants of health) get separate specific tabs that surface data of expected high relevance. A periodic or event-based automatic data loading feature has been developed which facilitates the incorporation of summarized (or directly streamed) data from IoT devices.

**Conclusion**

Using the principles of the Learning Healthcare System Data Commons framework to support research in quality improvement, clinical and translational investigation and training and educationally focused activities, here we present an open source technology reference architecture based on the infrastructure-as-a-service cloud computing, which greatly reduces inefficiencies in operational costs and greatly increases findability, accessibility, interoperability, and reusability of data and code. This Learning Healthcare System Commons solved an unmet need, made clear during the COVID-19 pandemic, for rapid, self-service access to health data sets for our researchers, with appropriate governance over use. Rush University Medical Center served as an ideal laboratory for the trial of this platform, as it has multiple types of healthcare data and multiple stakeholders across its academic mission, but also values efficiencies to broaden access to data. Figure 2C illustrates the expansion of this model from this reference (pilot) to outside institutions, and integration into existing data ecosystems (i.e., commons of commons).

Organizational digital transformation around research data in healthcare is an inevitability. The data used in clinical decision-making is increasingly heterogeneous, ranging from electronic medical records that can be flat image files; structured and unstructured fields; notes; medical image files, including high resolution flat images of tissue pathology or three dimensional radiological images; biorepository specimens; genomics files at various stages of processing and annotation; and other ancillary data from Internet of Things devices or status updates from clinical trial management systems. Accessing the sheer scale and heterogeneity of data became a clear issue over the past two years, as the COVID-19 pandemic has presented many clinically complex situations where urgency for data-driven insights were clear and the need to aggregate and share data responsibly was paramount. Here we present an open source technology reference Gen3 architecture based on the infrastructure-as-a-service cloud computing to execute a “data first” strategy through a holistic technology-enabled approach to data planning, governance. and use.
The Learning Healthcare System Commons was funded through a generous donation from the Searle family and by donated time from the Microsoft Cloud Acceleration Program.

Acknowledgements

The authors would like to extend their appreciation to the Cloud Acceleration Program at Microsoft and to the scientific computing team at BioTeam Inc. for support in further developing and porting the Gen3 platform to the Azure environment and for developing aspects of the Gen3 framework applied to an academic medical center research and learning environment.

Notes


8. Shneiderman B. Plaisant C. “Sharpening analytic focus to cope with big data volume and...


Author Biographies

Thomas O’Hara (tom.ohara@gmail.com) was the lead full stack engineer at Rush University System for Health. O’Hara is currently Python Software Engineer at Altoida Inc.

Anil Saldanha (anil_j_saldanha@rush.edu) is the chief cloud officer at Rush University System for Health.
chief information officer of the Fred Hutchinson Cancer Research center where he led a team of clinical informaticists and software engineers producing clinical data reports and developing reusable clinical data products.

Robert L. Grossman (rgrossman1@uchicago.edu) is the Frederick H. Rawson Distinguished Service Professor in Medicine and Computer Science and the Jim and Karen Frank Director of the Center for Translational Data Science at the University of Chicago. He is also the chair of the Open Commons Consortium, a nonprofit that develops and operates data commons to support research in science, medicine, health care, and the environment.

Bala Hota (bala.hota@gmail.com) was the chief analytics officer at Rush University System for Health. Hota is currently a senior vice president and chief informatics officer at Tendo Systems Inc.

Casey Frankenberger (frankenberger@gmail.com) was the chief research informatics officer at Rush University System for Health. Frankenberger is currently a vice president of business development at C2i-Genomics Inc.