



Research and Applications

An atomic approach to the design and implementation of a research data warehouse

Shyam Visweswaran ^{1,2,†}, Brian McLay^{1,†}, Nickie Cappella¹, Michele Morris¹, John T. Milnes¹, Steven E. Reis², Jonathan C. Silverstein^{1,2,3}, and Michael J. Becich ^{1,2}

¹Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, ²Clinical and Translational Science Institute, University of Pittsburgh, Pittsburgh, Pennsylvania, USA and ³Chief Research Information Officer, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

[†]Shyam Visweswaran and Brian McLay contributed equally to this article.

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

Received 30 May 2021; Revised 27 July 2021; Editorial Decision 29 August 2021; Accepted 10 September 2021

ABSTRACT

Objective: As a long-standing Clinical and Translational Science Awards (CTSA) Program hub, the University of Pittsburgh and the University of Pittsburgh Medical Center (UPMC) developed and implemented a modern research data warehouse (RDW) to efficiently provision electronic patient data for clinical and translational research.

Materials and Methods: We designed and implemented an RDW named Neptune to serve the specific needs of our CTSA. Neptune uses an atomic design where data are stored at a high level of granularity as represented in source systems. Neptune contains robust patient identity management tailored for research; integrates patient data from multiple sources, including electronic health records (EHRs), health plans, and research studies; and includes knowledge for mapping to standard terminologies.

Results: Neptune contains data for more than 5 million patients longitudinally organized as Health Insurance Portability and Accountability Act (HIPAA) Limited Data with dates and includes structured EHR data, clinical documents, health insurance claims, and research data. Neptune is used as a source for patient data for hundreds of institutional review board-approved research projects by local investigators and for national projects.

Discussion: The design of Neptune was heavily influenced by the large size of UPMC, the varied data sources, and the rich partnership between the University and the healthcare system. It includes several unique aspects, including the physical warehouse straddling the University and UPMC networks and management under an HIPAA Business Associates Agreement.

Conclusion: We describe the design and implementation of an RDW at a large academic healthcare system that uses a distinctive atomic design where data are stored at a high level of granularity.

Key words: research patient data repository, research data warehouse, secondary use, electronic health records

INTRODUCTION

The passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act by the US federal government

led to the widespread adoption of electronic health record (EHR) systems that capture patient data at an ever-increasing pace.¹ The availability of large amounts of EHR data provides new opportuni-

ties for their secondary use to support clinical and translational science. Furthermore, EHR data in combination with other patient data from research studies, patient-reported outcomes, mobile health, and social media are progressively becoming important in biomedical research.

Data warehouses containing EHR data exist in large healthcare systems for a variety of operational, reporting, quality improvement, and financial purposes.² However, such warehouses often do not support research effectively due to the heterogeneity of EHR data, regulatory complexity such as the requirement for deidentification,³ and the need for research-project-specific data management. A common approach to efficient and large-scale reuse of EHR data for research is a dedicated research patient data repository or research data warehouse (RDW) that integrates and harmonizes EHR data and is architected, implemented, and operated by personnel with informatics expertise. Funded by the National Center for Advancing Translational Sciences (NCATS), the Clinical and Translational Science Awards (CTSA) Program hubs have developed RDWs for the efficient and widespread use of EHR and other data for research, with 94% of all hubs providing such services.⁴

Dedicated RDWs have enabled a wide range of research efforts such as clinical trial recruitment, large-scale characterization of treatment pathways,⁵ generation of real-world evidence for clinical decision-making,⁶ pharmacovigilance,⁷ rapid cohort identification,⁸ and phenome-wide association studies.⁹ Furthermore, harmonized data in RDWs unlock future opportunities for large-scale application of machine learning for biomedical discovery¹⁰ and clinical decision support that can support order entry,¹¹ smart prioritization of data in EHR systems,¹² anomaly detection,¹³ and precision medicine.¹⁴

RDWs have evolved along 2 broad pathways.¹⁵ Several large academic health centers have developed a single institutional RDW that is architected specifically based on local EHR systems and needs. Examples of single-institution RDWs are those at Northwestern University,^{16,17} Duke University Health System,¹⁸ Stanford University,^{19,20} and Vanderbilt University.²¹ Other institutions have implemented RDWs based upon analytics-oriented data models designed for multi-institutional consortia and data networks. Examples of such data models include the Informatics for Integrating Biology and the Bedside (i2b2),²² the Observational Medical Outcomes Partnership (OMOP) Common Data Model,²³ and the National Patient-Centered Clinical Research Network (PCORnet) Common Data Model.²⁴

In this article, we describe the design and implementation of a single institutional RDW, called Neptune, at the University of Pittsburgh (Pitt). Neptune is architected to ingest patient data from a multitude of sources, to store data at the level of granularity that exists in the sources, and from which data are subsequently transformed into analytics-oriented data models and research data sets. Beyond patient data, knowledge for mapping to standard terminologies and definitions for standardizing clinical concepts are also stored in Neptune. Because of the multiple sources of data, including multiple EHR systems, Neptune uses an atomic design. An atomic data warehouse²⁵ contains data at a high level of granularity and preserves data from the source systems with minimal filtering or summarization. This is in contrast to a warehouse that implements a common data model; such a warehouse stores significantly transformed and harmonized data after extraction from the source systems. We provide a brief description of the large health system associated with Pitt, details of the architecture of Neptune, and some of the distinctive aspects of Neptune related to the technical infrastructure.

MATERIALS AND METHODS

Setting and history

The University of Pittsburgh Medical Center (UPMC) is one of the largest healthcare systems in the United States. UPMC serves western, central, and western Pennsylvania and parts of Ohio, West Virginia, and New York, and comprises 40 hospitals with 8400 licensed beds, more than 700 doctors' offices and outpatient facilities, and 23 nursing homes. Annually, UPMC has 388K inpatient admissions, 1.1M emergency room visits, 5.5M outpatient visits, and 260K surgical procedures. The University of Pittsburgh School of Medicine (UPSOM), located in the city of Pittsburgh, is the medical college and the clinical research facility that, together with UPMC, comprises a top 5 NIH-supported leading academic medical center. UPSOM supports an academic staff of nearly 2500 physicians and educators and trains approximately 600 medical students and 1900 medical residents and clinical fellows yearly.

UPMC has evolved as a merger of previously independent hospitals and practices, and its clinical information systems reflect this heritage, including a range of legacy and modern systems. UPMC has deployed several EHR systems from different vendors. In most outpatient facilities, UPMC uses the EpicCare system (Epic, Verona, Wis.), while in the inpatient and emergency settings, UPMC has deployed the Cerner system. The UPMC Children's Hospital of Pittsburgh has an independent installation of the Cerner Millennium system. Additional EHR and ancillary systems are used in various specialty settings such as inpatient psychiatry, the cancer center, the perioperative setting, and radiological imaging. UPMC has created multiple interfaces among the clinical information systems to enable clinical workflows that require data from multiple systems; however, this has led to the replication of patient data across these systems.

As early as 1982, Pitt and UPMC developed a clinical data warehouse called the Medical ARchival Retrieval System (MARS) that integrated data from EHR systems and administrative claims systems.²⁶ MARS was developed as a file-based database system that archived both structured and clinical document data in text files. As UPMC grew with the acquisition of hospitals and their clinical information systems, data integration was achieved by sending data in Health Level Seven (HL7) format through a message router to MARS.

Since MARS was implemented more than 3 decades ago, UPMC has grown substantially and has deployed several modern EHR systems. The need for a modern dedicated RDW emerged over the past several years. As a long-standing CTSA hub, Pitt needed a modern RDW and robust informatics services for efficient and effective support of investigators at Pitt and UPMC.

Organization

The Biomedical Informatics Core (BIC) of the University of Pittsburgh Clinical and Translational Science Institute, the Center for Clinical Research Informatics (CCRI),²⁷ and the Research Informatics Office (RIO)²⁸ lead the development of Neptune and provisioning of patient data for research. BIC, CCRI, and RIO are each housed in the Department of Biomedical Informatics and are each led by informatics faculty. On behalf of UPMC, 3 informatics faculty members oversee long-term planning, implementation of new features, and maintenance of Neptune and its downstream analytics-oriented data marts for national data-sharing efforts.

The team that supports Neptune and the data marts perform a range of functions. The *technical group* identifies use cases, reviews

and selects technologies, implements extract-transform-load processes based on data standards and terminologies established by a data harmonization group, and develops processes for aligning and integrating research data with the EHR data. The *data harmonization group* establishes data standards for different types of EHR and research data, determines which standard terminologies to use, and maintains and updates mappings of local terms to terminologies. The *data quality group* establishes statistics, uncovers data anomalies by periodically measuring these statistics in the data, and returns discoveries to the technical group for changes in the data pipelines. The *user support group* communicates with local users, provides training and support, and solicits feedback from the user community.

Architecture of Neptune

The Neptune RDW consists of 3 main layers: (1) an *identity management layer* for managing personally identifiable information, (2) a *data layer* that contains EHR and other patient data, both identified and as limited data with preserved timestamps and zip codes, and (3) a *semantic layer* that consists of business logic such as mappings between local terms and standard terminologies (see Figure 1). As mentioned earlier, Neptune uses an atomic design where source system data are stored with little to no filtering, summarization, or transformation. This design preserves data with no loss of information and permits data to be transformed and harmonized in unforeseen ways in the future without re-extracting from the source systems.

The RDW is implemented using the Oracle database management system. The warehouse physically straddles the UPMC and the Pitt networks. For example, the identity management layer of Neptune resides within the UPMC network, and the semantic layer and most of the data layer reside in the Pitt network (see Figure 1). Though the warehouse is split across 2 distinct networks, members of the technical group can seamlessly view tables from both components of Neptune and run processes across all layers of Neptune.

Identity management layer

The identity management layer resides in the UPMC network and contains personally identifiable information of all patients. A key function in Neptune is to assign and maintain a unique research enterprise identifier to each patient. This research enterprise identifier is distinct from patient identifiers that are used in the healthcare system, including clinical enterprise identifiers, medical record numbers, and other healthcare identifiers. The research enterprise identifier is linked to healthcare system patient identifiers and is also linked to participant identifiers of research data sets that are integrated into Neptune. This 3-layer identity management exceeds best practices for Health Insurance Portability and Accountability Act (HIPAA) honest brokerage and helps ensure participant identifiers cannot be shared across projects.

Identity management and linking of patient identifiers are performed in a staging area. During monthly ingestion of data from clinical systems, new patients in the health system are identified and assigned new research enterprise identifiers. For existing patients who may have been assigned new clinical identifiers, the new identifiers are linked to the existing research enterprise identifier. Any merges of clinical enterprise identifiers and medical record numbers are also processed. Identity management enables patient data from any data domain and linked to any patient identifier to be accurately linked to the enterprise research identifier. Neptune's identity management achieves a key goal of Neptune: to create a comprehensive longitudinal record for each patient by integrating clinical and non-clinical data from multiple sources.

Data layer

The data layer resides mostly at Pitt and contains both structured and text EHR data as well as other types of data, such as imaging data (see Figure 2). The data layer stores atomic patient data; that is, the data are at the level of granularity in the source system with minimal transformation. The structured data include core data domains such as demographics, visits, diagnoses, procedures, laboratory test results,

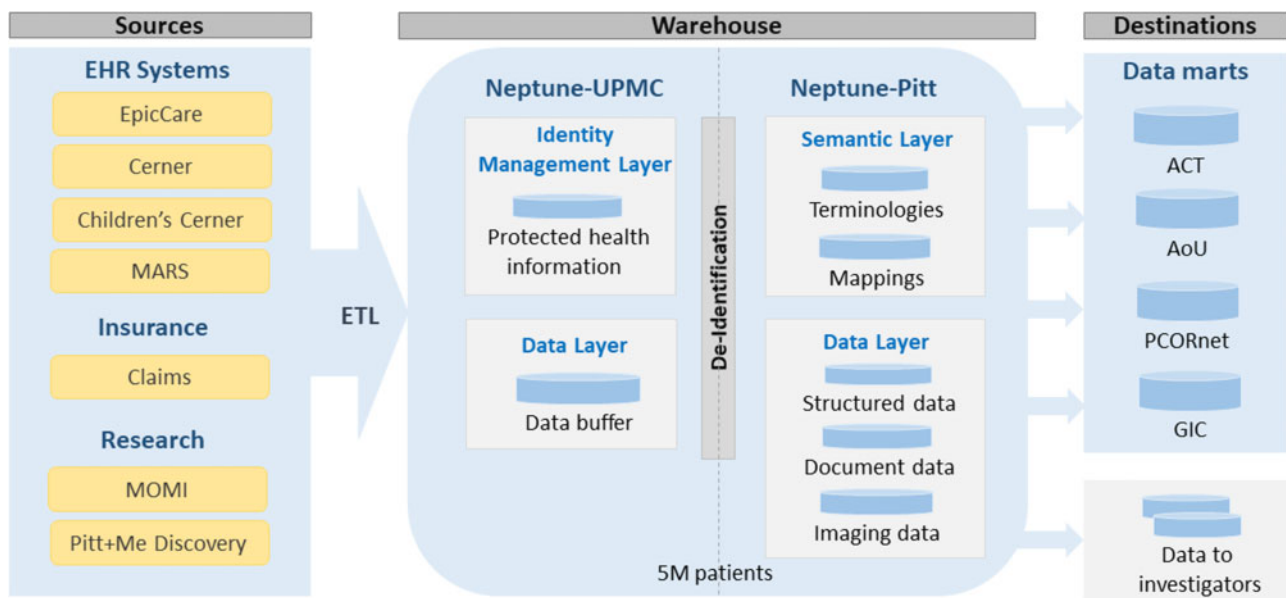


Figure 1. The architecture of Neptune with sources and destinations. The identity management layer resides at UPMC, the semantic layer resides at Pitt, and the data layer resides mostly at Pitt. ACT: Accrual to Clinical Trials; AoU: All of Us Research Program; GIC: Genomic Information Commons; PCORnet: National Patient-Centered Clinical Research Network.

Total	Monthly	Data destinations
5M patients Structured data 460M visits 311M diagnoses 130M procedures 1.42B lab test results 96M drug prescriptions 185M drug dispenses Clinical document data 371M documents Clinical imaging data 3M mammograms Insurance 1.3M patients	21K patients Structured data 4M visits 5M diagnoses 1M procedures 11M lab test results 1M drug orders 1M drug dispenses Clinical document data 4.5M documents	Data marts for national projects ACT 4.9 M patients AoU 26K participants PCORnet 3.5 M patients GIC 1.5M patients Local data marts Alzheimer's disease 13,000 patients Antibiotic usage 200,000 patients Local investigators 150 projects year

Figure 2. Total volume of data in Neptune, monthly data inflow, and data volumes in destinations served by Neptune.

medication orders, other orders (for laboratory tests, clinical imaging, procedures, etc.), and medication dispenses. Additional data domains include allergies, vaccine administrations, and metadata of clinical documents. The text data consist of deidentified content for all document types such as history and physical, progress, consultation, procedure, and discharge notes; radiology and pathology reports; electrocardiogram and electroencephalogram reports; and many more. In every domain, for each data item, the source system from which it was extracted is recorded to maintain data provenance.

At the time of extraction of data from source systems, the extracts are staged in the data layer at UPMC, where deidentification is performed before data are moved to the data layer at Pitt. A copy of the latest extract of data for all domains is maintained in the staging data layer. For most domains, data are extracted and processed monthly. Deidentification consists of removing all HIPAA-specified personally identifiable information with the exception of dates and zip codes to create a Limited Data Set, and the data are linked to the patient only through the research enterprise identifier. Deidentification is straightforward for structured data domains where database columns containing personally identifiable information are removed.

Deidentification of clinical documents is done using NLM Scrubber²⁹ that has been adapted for our use. We chose NLM Scrubber since it is an efficient tool that runs at scale on Linux (our preferred platform) on the health system side to perform best-effort deidentification on all clinical documents on extraction. NLM Scrubber, as well as other text deidentification tools, cannot guarantee perfect deidentification as HIPAA Safe Harbor method and Limited Data Sets require. For clinical documents, we generally use Data Use Agreements under HIPAA Waiver of Authorization, unless the clinical documents can also be hand-audited for the project.

Semantic layer

The semantic layer resides at Pitt and contains the knowledge and logic to harmonize data that may be represented in a heterogeneous fashion across different hospitals and source systems. For example, for the laboratory test of hematocrit, each hospital at UPMC uses a different local code, and the semantic layer contains a list of all local hematocrit codes that are mapped to the relevant LOINC code for standardization. Mappings are obtained from several sources. One source is reference data obtained from source systems like EpicCare that contains mappings between local terms and standard terminologies that are created and maintained by the clinical enterprise. However, the clinical enterprise does not necessarily create mappings to legacy data or mappings to standard terminologies that are not mandated by federal regulations. The data harmonization group creates

and updates mappings for legacy data and mappings that are useful in research. In addition, we import a comprehensive set of medical terminologies that are contained in the Unified Medical Language System (UMLS)³⁰ and maintain all versions of the terminologies going back to 2004 UMLS releases. We harmonize diagnoses to ICD-9 and ICD-10, procedures to ICD-9, ICD-10, CPT-4, and HCPCS, medications to RxNorm and NDC, and laboratory tests to LOINC. We also use the UMLS to validate the source system data that are increasingly coded with standard terminologies. In addition to mappings, the semantic layer contains value sets that have been collected from several sources, such as the NIH's Value Set Authority Center³¹ and value sets that have been defined by national patient data research networks. We typically update the terminologies and value sets twice a year, in June and in December, following the biannual UMLS releases. However, some terminologies and value sets are updated more frequently; for example, at the onset of the coronavirus disease 2019 (COVID-19) pandemic, we updated the LOINC codes for COVID-19 laboratory tests as frequently as weekly.

The mappings and value sets of the semantic layer are leveraged at the time data are delivered from Neptune to downstream data marts and to individual projects. Standard terminology codes and values are applied at that time to produce standardized data; this late binding approach provides efficient and timely standardization of data to constantly changing terminologies.

Extract, transform, and load processes

Most data domains in Neptune are updated at the beginning of each month. A series of extract, transform, load (ETL) database operations are implemented using the Pentaho Data Integrator (PDI) and run overnight to extract a month of data from the source systems. More than 70 workflows process the monthly incremental data updates. The PDI programs perform extraction and loading of warehouse tables, including data validation checks, error handling, auditing, and control processing. Linux scripts are used to call PDI programs. The data in the source systems are transactional, and some transactions may take several weeks to be finalized. At the beginning of each month, data are extracted from the source systems with a lag period of 1 month, at which time, almost all clinical transactions have been finalized in our health system. The use of the lag period accrues large savings in both human effort and system performance since the ETL processes are relatively simple, resulting in only the addition of rows to the warehouse tables rather than performing reconciliation and edits of changing values. However, a small amount of data are held back because they are incomplete; for example, data that are related to inpatient stays extending through the last day of the month and laboratory test orders that are not yet resulted by the last day of the month. The decision of monthly updates arose partly from the fact that at the onset, we had a small group of warehouse personnel who performed a wide range of tasks, including data provisioning to a large clientele, and partly from the fact that in the beginning, substantial time and effort was incurred in the design and implementation of ETL processes with the step-wise addition of key data domains over more than a year.

Extension of Neptune for COVID-19

The emergence of coronavirus disease 2019 (COVID-19) necessitated a more frequent update of COVID-19-related data in Neptune to support surveillance and research needs. Since UPMC is a large health system with millions of active patients and billions of transactions, rather than changing the monthly ETL processes, we devel-

oped a new parallel ETL process that updates data on COVID-19 twice a week. For efficiency, the new process extracts data only between the latest monthly Neptune update and the current day; thus, the data lag the source systems at most by 4 days. This COVID-19 component of Neptune serves national projects that added more frequent data requirements for COVID-19 serving the COVID-19 needs of local investigators.

Regulation of Neptune

Neptune was initially developed under an institutional review board (IRB) protocol, but the regulatory framework was later changed to an HIPAA Business Associates Agreement (BAA). Regulation under an IRB protocol was limiting since the addition of new data sources and technical changes need repeated changes to the protocol. The BAA is overseen by the Chief Research Information Officer (CRIO) at Pitt and the Chief Medical Information Officer (CMIO) at UPMC. This arrangement, allowing Neptune to function as an operational research system of UPMC under the BAA, enabled rapid expansion and development of new functionality in Neptune.

The warehouse is used in several configurations, including some free services, mostly grant-supported activities, and a substantial recharge center. Investigators have access to limited data in downstream self-service systems such as the i2b2, patient-level data as extracts performed by honest brokers, and regularly updated patient data as study-specific marts. Monthly meetings between the CRIO on the Pitt side with the CMIO on the UPMC side ensure robust and extensive use of Neptune for both Pitt and UPMC research that is highly valued by both organizations.

UPMC and Pitt have a shared regulatory compliance structure, including a shared IRB. The BAA rather than an IRB protocol for the warehouse ensures that full-time Pitt informatics personnel in the CRIO's office can operate on behalf of UPMC while directly supporting investigators, and also provides a structure for ensuring that Pitt investigators' responsibilities to UPMC are fulfilled. The BAA requires all uses of the warehouse have IRB protocols with the exception of preparatory-to-research queries.

RESULTS

Designed as an RDW that integrates patient data from varied sources, Neptune contains EHR data (structured, document, and imaging), insurance data, and research data. EHR data in Neptune goes back to 2004 when UPMC completed the implementation of electronic clinical information systems. Every month, a large volume of EHR data are added to Neptune that includes data from existing patients and approximately 21 000 new patients (see Figure 2). Neptune also receives health insurance claims data from the UPMC Health Plan and from large institutional research projects like the Magee Obstetric Maternal & Infant (MOMI) Database and Biobank (>300 perinatal variables from mother and infant, ~200K deliveries since 1995),^{32,33} and genomic data from the Pitt+Me Discovery Biobank.

Neptune provides data to data marts for several national projects that include the Accrual to Clinical Trials (ACT) network,^{8,34} which is based on i2b2²² and Shared Health Research Information Network (SHRINE)³⁵; the All of Us Research Program which is based on the OMOP data model^{36,37}; the PCORnet, which is based on PCORnet's Common Data Model (CDM) and PopMedNet;^{38,39} and the Genomic Information Commons (GIC) which is based on tranSMART.^{40,41} Neptune also provides data to data marts for local projects such as an Alzheimer's disease project and an antibiotic us-

age project. Typically, data are automatically updated in both the national and local data marts following the monthly data updates in Neptune (see Figure 2).

In addition, Neptune serves as a source of EHR and other patient data for local research in the institution. The RIO provisions data to hundreds of individual research projects per year. Finally, RIO responds to approximately 1000 requests per year, including preparatory to research requests and letters of support for research grants.

DISCUSSION

We described the design and implementation of Neptune, a new RDW, at Pitt and UPMC. Neptune is designed to integrate data from several EHR systems with replicated patient records as well as non-EHR data, support both identified and deidentified data needs, and service efficiently commonly used analytics-oriented data models and data needs of individual investigators. The rich partnership between Pitt and UPMC supported the rapid technical development and implementation of Neptune. This warehouse is an increasingly rich repository of EHR and other patient data and is progressively benefitting the dynamic research environment at Pitt and UPMC.

Distinctive features of Neptune

This section describes distinctive aspects of Neptune's technical infrastructure. Desiderata for the successful implementation and operation of an RDW has been described by Huser and Cimino.⁴² These include a single patient identifier, protected health information (PHI) management, an extensible information storage model, semantic integration with standard terminologies, metadata and documentation, and documentation of historical evolution of data sources. Several of the features of Neptune described below align with these desiderata.

Atomic data warehouse

Neptune is architected as a canonical model for ingestion and storage of patient data derived from multiple sources and from which data are subsequently transformed into analytics-oriented data models. The canonical model in Neptune uses a normalized atomic design. Normalization is a key database principle that enables the efficient correction of data errors and optimization of storage space. The atomic design enables rapid ingestion of data in bulk, tracking of data provenance, isolation, separate processing of changing data, and provides a single place for data cleaning and transformation rather than duplicating these processes for each data source. The advantage of an atomic warehouse is it can both provide answers to queries at a very detailed level and summarize data rapidly that may be needed for analytics-oriented data models. Neptune enables us to avoid converting data from one data model to another, for example, from OMOP to PCORnet's Common Data Model or vice-versa, which is typically more complex to implement than an ETL process from Neptune to an individual data model. Furthermore, due to information loss, it is not possible to inter-convert between data models with complete fidelity. The atomic design also enables the stepwise addition of new data domains without the need to redesign or implement a comprehensive set of all possible data domains that will eventually be needed.

Single patient identifier and management of PHI

A key feature in Neptune is the management of patient identifiers such that all data related to a patient originating from different EHR systems, health insurance, and research studies are linked to a

single enterprise research identifier. While UPMC maintains a single enterprise clinical patient identifier that links all clinical identifiers of a patient, the enterprise clinical patient identifier was not usable in the RDW for several reasons. We needed an identifier—the enterprise research identifier—that is not PHI and can be linked to both clinical patient identifiers and patient identifiers used in research studies. The identity management layer is used to integrate clinical and nonclinical identifiers and assign a unique enterprise research identifier to each patient. This layer resides on the health system side since it contains PHI; thus, the architecture of Neptune helps ensure that PHI does not leave the confines of the health system network. The enterprise research identifier accomplishes several goals. It is used to resolve historical patient identifiers and link historical data to the patient. It also provides a framework for data extractors (honest brokers) to work with different tools for the different types of data, such as structured data, clinical document data, and clinical imaging data, on the Pitt side of Neptune without using PHI. The majority of our honest brokers accomplish their work entirely on the Pitt side of Neptune, thus reducing the number of personnel with access to PHI to the minimum. Furthermore, the enterprise research identifier is not provided to investigators; it is transformed to a study-specific identifier to avoid the risk of investigators linking a patient's data in a study with that patient's data in another study, which is not generally allowed under HIPAA.

Privacy and study patient identifiers

The enterprise research patient identifier is restricted for use within Neptune and is not used to identify patients when data are delivered to data marts and for research projects. Unique study patient identifiers are created and assigned for patients in each data mart and research data set that are derived from Neptune. A function is used to systematically transform the enterprise research patient identifier to a study patient identifier for each data set, and function details associated with each data set are archived in Neptune. The function allows warehouse personnel to link study identifiers to enterprise research patient identifiers for future updates to study data, but investigators cannot link data by patient across different data sets that were provisioned under different IRB protocols that may have common patients.

Extensible information storage model

An important consideration for Neptune was rapid implementation, starting with key data domains so that the warehouse would be functional within months rather than years of development. The key data domains were identified based on the clinical domains required to populate the data marts for national projects such as the All of Us Research Program. Neptune initially contained only structured EHR data in the domains of demographics, diagnoses, procedures, laboratory test results, and medications. This enabled implementation in under 6 months. The addition of a new domain includes a selection of sources, identification of deduplication strategies if necessary, aligning patient identifiers, a bulk backload of the data going back to 2004, and implementation of a monthly ETL process. The extensible information storage model implemented in Neptune has enabled the stepwise addition of new data domains without the need to rearchitect existing data domains.

Dereplication of data from multiple sources

Multiple EHR systems and an archival system are in use in UPMC. Assembling a longitudinal health record from these multiple sources

is another key requirement for Neptune. In addition to the multitude of patient identifiers, another challenge associated with the use of multiple EHR systems is the replication of patient data across systems. We achieved dereplication in several ways. One approach is the selective extraction of data from a single source if a particular domain is systematically replicated across the EHR systems. For example, since 2015, in UPMC, laboratory test results from all care settings are available in EpicCare; thus, laboratory test results after 2015 are extracted only from EpicCare. Another approach compares timestamps and metadata of suspected replicated data to identify replication. For example, laboratory test results before 2015 were obtained from several sources, and replications were identified and systematically eliminated.

Binding at query

Binding is the process of mapping data to standard terminologies (eg, translation of a local code for a laboratory test to the appropriate LOINC code) and application of definitions (eg, application of a standard definition of an outpatient visit and calculation of the length of stay). Binding standardizes the data and makes it usable for research. In some warehouse designs and analytics-oriented models, mappings and definitions are applied early during data ingestion; such early binding has the disadvantage that changes to the mappings and definitions will need data to be corrected and updated continually. Since Neptune uses binding at query time, changes in mappings and definitions affect data only at the time data are delivered from Neptune, and new data sources are rapidly integrated into Neptune without making decisions about mappings upfront.

Limitations

Neptune has several limitations. One limitation is that the data in the warehouse lag the source systems by a month. While this delay is acceptable for most research that uses retrospective data, it limits research in clinical decision support and biosurveillance applications that typically require current or near current EHR data. But, as mentioned previously for COVID-19 data, extending the capabilities to support requirements of more frequent data updates is possible with additional development. A second limitation is that there is no efficient mechanism to query clinical document data, while structured data can be queried by the warehouse personnel by directly querying Neptune or via the i2b2. We have separately implemented Elasticsearch technology for efficient query and analysis of clinical documents. A third limitation arises from the duplication of patient data in the source systems, especially across the inpatient and outpatient EHR systems. This results in complex and time-consuming analyses to design ETL processes to ensure that data in Neptune is deduplicated, which is further complicated by the dynamic nature of duplication. A fourth limitation arises from the need for data quality checks at every level of warehouse function in order to provide meaningful data for research. The multiple EHR source systems and the duplication of data in them have required more than the usual volume of data checks which are still likely incomplete.

CONCLUSION

The Neptune RDW implemented at Pitt is increasingly enabling extensive reuse of patient data for a wide range and high volume of clinical and translational research. Neptune is designed as a normalized atomic warehouse. The atomic design enabled the warehouse to be built “better, faster, cheaper” because there is no need to exten-

sively model or standardize the data. Neptune integrates patient data from multiple EHR systems as well as from other sources, maintains a robust patient identity management system for research, and enables efficient delivery of data to both large data marts based on analytics-oriented data models and to individual investigators. Creating a dedicated RDW at Pitt has enabled us to better serve the investigators at Pitt, participate national data networks, and advance informatics research.

FUNDING

The research reported in this article was supported by awards from the National Center for Advancing Translational Sciences of the National Institutes of Health (NIH) under award numbers UL1 TR001857, UL1 TR001857-01S1, and U01 TR002623, the Office of the Director of the NIH under award number OT2 OD026554, the National Library of Medicine of the NIH under award number R01 LM012095, and the PCORI award RI-CRN-2020-006. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or PCORI.

AUTHOR CONTRIBUTIONS

SV conceived and designed the study, participated in data analysis and interpretation, drafted the manuscript, and approved the final version for submission. BM conceived and designed the study, participated in data collection, analysis and interpretation, drafted the manuscript, and approved the final version for submission. NC participated in data analysis and interpretation, made critical revisions to the manuscript, and approved the final version for submission. MM participated in data analysis and interpretation, made critical revisions to the manuscript, and approved the final version for submission. JTM and SER made critical revisions to the manuscript and approved the final version for submission. JCS participated in data analysis and interpretation, made critical revisions to the manuscript, and approved the final version for submission. MJB conceived and designed the study, made critical revisions to the manuscript, and approved the final version for submission.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

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

REFERENCES

- Blumenthal D. Launching HITECH. *N Engl J Med* 2010; 362 (5): 382–5.
- Evans RS, Lloyd JF, Pierce LA. Clinical use of an enterprise data warehouse. *AMIA Annu Symp Proc* 2012; 2012: 189–98.
- U.S. Dept. of Health & 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 2020. <https://www.hhs.gov/guidance/document/de-identification-guidance> Accessed September 24, 2020.
- Obeid JS, Beskow LM, Rape M, *et al.* A survey of practices for the use of electronic health records to support research recruitment. *J Clin Transl Sci* 2017; 1 (4): 246–52.
- Hripscak G, Ryan PB, Duke JD, *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A* 2016; 113 (27): 7329–36.
- Longhurst CA, Harrington RA, Shah NH. A ‘green button’ for using aggregate patient data at the point of care. *Health Aff (Millwood)* 2014; 33 (7): 1229–35.
- Haerian K, Varn D, Vaidya S, Ena L, Chase H, Friedman C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clin Pharmacol Ther* 2012; 92 (2): 228–34.
- 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.
- Denny JC, Bastarache L, Roden DM. Phenome-wide association studies as a tool to advance precision medicine. *Annu Rev Genomics Hum Genet* 2016; 17: 353–73.
- Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE. Medical data mining: knowledge discovery in a clinical data warehouse. *AMIA Annu Symp Proc* 1997; 1997: 101–5.
- Chen JH, Altman RB. Automated physician order recommendations and outcome predictions by data-mining electronic medical records. In: *AMIA Jt Summits Transl Sci Proc* 2014; 2014: 206–10.
- King AJ, Cooper GF, Clermont G, *et al.* Using machine learning to selectively highlight patient information. *J Biomed Inform* 2019; 100: 103327.
- Hauskrecht M, Batal I, Hong C, *et al.* Outlier-based detection of unusual patient-management actions: an ICU study. *J Biomed Inform* 2016; 64: 211–21.
- Sitapati A, Kim H, Berkovich B, *et al.* Integrated precision medicine: the role of electronic health records in delivering personalized treatment. *Wiley Interdiscip Rev Syst Biol Med* 2017; 9 (3): 10.1002/wsbm.1378.
- Shin S-Y, Kim WS, Lee J-H. Characteristics desired in clinical data warehouse for biomedical research. *Health Inform Res* 2014; 20 (2): 109–16.
- Northwestern University Clinical and Translational Sciences Institute Enterprise Data Warehouse; 2020. <https://www.nucats.northwestern.edu/resources/data-science-and-informatics/nmedw/index.html> Accessed September 24, 2020.
- Starren JB, Winter AQ, Lloyd-Jones DM. Enabling a learning health system through a unified enterprise data warehouse: the experience of the Northwestern University Clinical and Translational Sciences (NUCATS) Institute. *Clin Transl Sci* 2015; 8 (4): 269–71.
- Horvath MM, Ruscovitch SA, Brinson S, Shang HC, Evans S, Ferranti JM. Modular design, application architecture, and usage of a self-service model for enterprise data delivery: the Duke Enterprise Data Unified Content Explorer (DEDUCE). *J Biomed Inform* 2014; 52: 231–42.
- Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE—an integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc* 2009; 2009: 391–5.
- Datta S, Posada J, Olson G, *et al.* A new paradigm for accelerating clinical data science at Stanford Medicine. arXiv preprint arXiv:2003.10534; 2020.
- Danciu I, Cowan JD, Basford M, *et al.* Secondary use of clinical data: the Vanderbilt approach. *J Biomed Inform* 2014; 52: 28–35.
- Murphy SN, Weber G, Mendis M, *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* 2010; 17 (2): 124–30.
- Hripscak 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): 574–8.
- 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.
- Haertzen D. *The Analytical Puzzle: Data Warehousing, Business Intelligence and Analytics*. Technics Publications, LLC; 2012: 80–1.
- Yount RJ, Vries JK, Councill CD. The Medical Archival System: an information retrieval system based on distributed parallel processing. *Inf Process Manag* 1991; 27 (4): 379–89.
- Center for Clinical Research Informatics (CCRI). 2020. <http://www.ccri.thevislab.com/> Accessed September 24, 2020.
- Research Informatics Office (RIO). 2020. <http://rio.pitt.edu/> Accessed September 24, 2020.
- Kayaalp M, Browne AC, Dodd ZA, Sagan P, McDonald CJ. An Easy-to-Use Clinical Text De-Identification Tool for Clinical Scientists: NLM Scrubber. *AMIA Annu Symp Proc* 2015; 2015: 1522.

30. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32 (Database issue): D267–70.
31. Bodenreider O, Nguyen D, Chiang P, *et al.* The NLM value set authority center. *Stud Health Technol Inform* 2013; 192: 1224.
32. Naimi AI, Platt RW, Larkin JC. Machine learning for fetal growth prediction. *Epidemiology* 2018; 29 (2): 290.
33. Magee Obstetric Maternal & Infant (MOMI) Database and Biobank. 2020 <https://mageewomens.org/for-researchers/core-facilities/momi> Accessed September 24, 2020.
34. The ACT Network. 2020. <https://www.actnetwork.us/> Accessed September 24, 2020.
35. Weber GM, Murphy SN, McMurry AJ, *et al.* The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc* 2009; 16 (5): 624–30.
36. All of Us Research Program. The “All of Us” research program. *N Engl J Med* 2019; 381 (7): 668–76.
37. All of Us Research Program. 2020. <https://allofus.nih.gov/> Accessed September 24, 2020.
38. Davies M, Erickson K, Wyner Z, Malenfant J, Rosen R, Brown J. Software-enabled distributed network governance: the PopMedNet experience. *EGEMS (Wash DC)* 2016; 4 (2): 1213.
39. PopMedNet. 2020. <https://www.popmednet.org/> Accessed September 24, 2020.
40. Mandl KD, Glauser T, Krantz ID, *et al.*; Genomics Research and Innovation Network. The Genomics Research and Innovation Network: creating an interoperable, federated, genomics learning system. *Genet Med* 2020; 22 (2): 371–80.
41. Scheufele E, Aronzon D, Coopersmith R, *et al.* tranSMART: an open source knowledge management and high content data analytics platform. *AMIA Jt Summits Transl Sci Proc* 2014; 2014: 96–101.
42. Huser V, Cimino JJ. Desiderata for healthcare integrated data repositories based on architectural comparison of three public repositories. *AMIA Annu Symp Proc* 2013; 2013: 648–56.