Paper 840-2017

HEALTHCARE DATA SHARING AND INNOVATIVE ANALYTIC DEVELOPMENT: LESSONS LEARNED FROM DISTRIBUTED DATA NETWORKS

Jennifer R. Popovic, Harvard Medical School and Harvard Pilgrim Health Care Institute

Boston, MA

ABSTRACT

Secondary use of administrative claims data, electronic health and medical record (EHR/EMR) data, registry data, and other sources within the health data ecosystem provide rich opportunity and potential to study topics ranging from public health surveillance to comparative effectiveness research to identifying patients who may be eligible for inclusion in a specific study or for a specific intervention. Data sourced from individual sites can be limited in their scope, coverage, and statistical power. Sharing and pooling data from multiple sites and sources, however, present administrative, governance, analytic, and patient-privacy challenges. Distributed data networks represent a paradigm shift in health-care data sharing. They have evolved at a critical time when big data and patient privacy often have competing priorities.

A distributed data network is one for which no central repository of data exists. Data reside behind the firewalls of each data-contributing partner in a network. Each partner transforms its source data in accordance with a common data model and allows indirect access to those data through a standard query approach using flexibly designed informatics tools. This paper discusses how distributed data networks have matured to make important contributions to the health-care data ecosystem and the evolving Learning Healthcare System.

INTRODUCTION

This paper introduces the concept of distributed data networks, including their purpose, guiding principles and benefits. This paper also discusses the concept of analytic infrastructure and provides examples of three important elements of distributed data network analytic infrastructure, as well as six foundational characteristics of analytic infrastructure. The three elements of distributed data network analytic infrastructure discussed in this paper include:

- Standardized data structure (common data model)
- Standardized assessment of data quality
- Analytic tool design and development

The six foundational characteristics of analytic infrastructure discussed are:

- consistency
- flexibility
- reusability
- scalability
- transparency
- reproducibility

The U.S. Food and Drug Administration (FDA)-sponsored Sentinel Initiative is referred to throughout this paper as an example use-case of one successful large-scale, multi-site, healthcare-related distributed data network utilizing electronic healthcare data from multiple sources.

DISTRIBUTED DATA NETWORKS

A distributed data network is one for which no central repository of data exists. Rather, data are maintained by and reside behind the firewall of each data partner, which allow indirect analytic access to their patient-level data via programming code that is securely distributed to them and intended to execute on their side of the firewall. The data are therefore 'distributed' due to the lack of centrality.

Distributed data networks exist by a set of guiding principles [1]:

- Data partner sites maintain control over their data,
- Data partner sites have standardized their data to a common data model,
- Data partner sites' ongoing involvement is needed in order to interpret data and findings; they know their data the best, so are true ***partners*** in the network,
- Analytic programming code gets securely distributed to data partners for them to execute locally and in a manner that makes it easy for them to execute,
- Following execution of analytic programming code, data partners return results that were produced by the executed code, to the requestor. Typically, data returned are aggregated rather than patient-level.

PURPOSE AND BENEFITS OF A DISTRIBUTED DATA NETWORK

Distributed data networks often allow for access to more data than what a single or centralized site might be able to offer. By pooling resources (data) across several sites, with security and governance in place such that each site maintains ownership over its own data, these networks provide several key benefits [2]:

- Offering alternative ways to study occurrences of rare outcomes, uptake or usage of new drugs or therapies, and diverse populations of individuals,
- Achieving greater statistical power due to larger numbers of observations,
- Encouraging the development of novel analytic and statistical methods that do not rely solely on the use of patient-level data,
- Addressing and alleviating data partners' concerns over data security, patient privacy and proprietary interests,
- Challenging analytic programmers to approach projects with the intention of building reusable, flexible and scalable programs for infrastructure purposes, rather than a series of one-off programs.

SENTINEL INITIATIVE: AN EXAMPLE OF A DISTRIBUTED DATA NETWORK

The Sentinel Initiative is a program sponsored by the U.S. Food and Drug Administration (FDA) to create an active surveillance system to monitor the safety of FDA-regulated medical products. Section 905 of the Food and Drug Administration Amendments Act (FDAAA) of 2007 mandated the FDA to enhance their ability to monitor the post-market safety of the medical products it regulates [3]. This system, named Mini-Sentinel in its pilot phase and known as Sentinel since 2016 [4], is intended to augment, not replace, FDA's existing post-market safety monitoring systems [5]. Adverse event reporting systems in existence prior to Sentinel relied on external sources (e.g., product manufacturers, consumers, patients, healthcare professionals) to report suspected adverse events that may be associated with FDA-regulated products to the agency. This is often referred to as "passive surveillance." In contrast, Sentinel is an "active surveillance" system, enabling FDA to initiate its own medical product safety evaluations, using data curated for and maintained within the Sentinel Distributed Database (SDD) [3].

The SDD currently consists of quality-checked data held by 18 partner organizations, which are either health insurers, integrated delivery systems or provider networks. Data partners transform and standardize data from their source systems in accordance with the schema outlined in the Sentinel Common Data Model (SCDM), and they store these SCDM-formatted datasets as SAS® datasets behind their firewalls. Each site maintains physical control and ownership of their data, controls all uses of their data and controls all transfer of their data. Data partners refresh their source data into SCDM-formatted data quarterly to annually, depending on the site.

As of March 2017 [6], the SDD contained data on:

- 223 million individuals (43 million of whom are accruing new data),
- 425 million person-years of observation time,
- 5.9 billion outpatient dispensings,
- 7.2 billion unique medical encounters.

Sentinel leveraged some technical architecture, infrastructure and governance aspects from other distributed data networks but is unique in several ways. It was the first distributed data network for which dedicated funding was allocated to a central Coordinating Center specifically for the purpose of designing, building, maintaining and expanding systems and analytic infrastructure, in order to create a nationally-representative distributed database. Sentinel is also the only distributed data network that is an integral part of a Federal regulatory agency's regulatory activities [7]. Although the FDA remains the sole primary user of Sentinel and the only entity that can initiate queries of the SDD, they have encouraged others to leverage Sentinel's data and analytic capabilities as a potential national evidence generation system [8-10].

OTHER DISTRIBUTED DATA NETWORKS IN EXISTENCE

In addition to Sentinel, there are several other healthcare-related distributed data networks in existence, and some of these networks have a particular focus. For example, PCORnet focuses on conducting comparative effectiveness and patient-centered outcomes research, the NIH Health Care Systems Research Collaboratory's mission is to improve the way clinical trials are conducted, and the Biologics and Biosimilars Collective Intelligence Consortium's (BBCIC) focus is on post-market evidence generation for biologics and their corresponding biosimilars. Below are examples of other healthcare-related distributed data networks:

- PCORnet: The National Patient-Centered Clinical Research Network
- NIH Health Care Systems Research Collaboratory
- Biologics and Biosimilars Collective Intelligence Consortium (BBCIC)
- Health Care Systems Research Network (HCSRN)
- Innovation in Medical Evidence Development and Surveillance System (IMEDS)
- Observational Health Data Sciences and Informatics (OHDSI) program
- Cancer and Cardiovascular Research Networks
- Vaccine Safety Datalink (VSD)

Several of these networks share the same data partners. Some may also share or leverage the same common data model, analytic tools and/or other infrastructure as the backbone to support the manner in which their network operates and analyzes data.

ANALYTIC INFRASTRUCTURE

The term analytic infrastructure has been discussed in the analytics community and literature. In this paper, analytic infrastructure includes the systems, processes, governance, data, software, tools and people that facilitate the analytic process.

This paper does not discuss all aspects of analytic infrastructure, but rather focuses on three distinct elements: standardized data structure, standardized data quality assessment, and analytic tool development, and highlights how those three elements can be developed to encompass the six foundational characteristics of analytic infrastructure, as shown in Figure 1 [7].

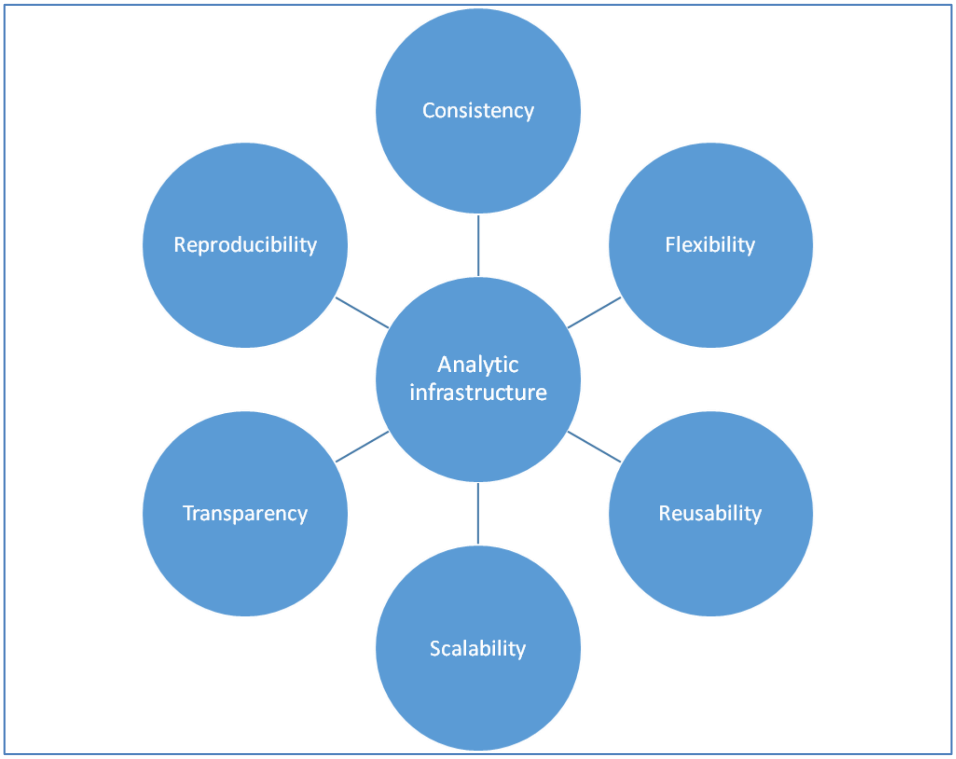


Figure 1: Six foundational characteristics of analytic infrastructure.

There is a lot of synergy across these six characteristics. Consistency is adherence to some common principles or conditions. Analytic consistency refers not only to consistency in the structure and flow of how analytic programs are designed and developed but also in analytic approach, such as keeping algorithms consistent and stable across analytic tools.

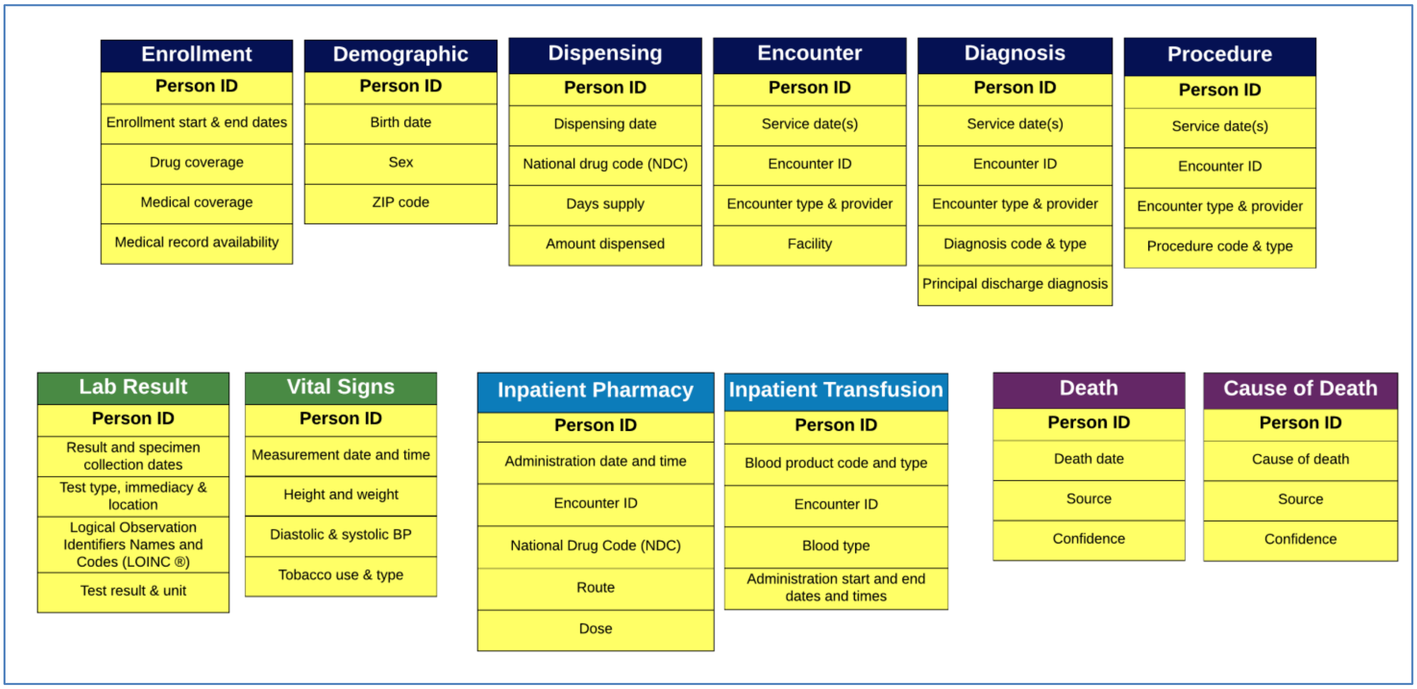
Flexibility is the power to adapt, such as to new or changing study design criteria. Analytic tools that are flexibly designed and developed are parameter-, data- and/or table-driven, for both study reusability and scalability purposes. Flexibly designed and developed tools are intended to be reusable across studies with similar types of analytic study designs, but also flexible to make maximal use of available hardware/software resources, which gets at scalability. Scalability is the idea of being easily expandable (or retractable) based on needs and resources. Analytic programs written with scalability in mind are equipped to make optimal use of computing resources that are appropriate for the analytic need and/or volume of data that are being analyzed.

Transparency and reproducibility are hallmark characteristics of analytic infrastructure, as well, and are often realized by making models, tools, and other infrastructure components open-source and freely available, as well as by making any products of those models and tools (e.g., study protocols and reports) readily available to the public.

STANDARDIZED DATA STRUCTURE: THE COMMON DATA MODEL

The purpose of any common data model is to “standardize the structure, format and content of data, such that standardized applications, tools and methods can be applied to them” [11]. There are several healthcare-related common data models in existence and some were even born out of each other. For example, the SCDM is in-part based on the HCSRN Virtual Data Warehouse (VDW) CDM, and the PCORNet CDM is in-part based on the Sentinel CDM [12-14]. Although common data models are similar in their goal of standardizing the capture and storage of data elements from various source systems, the design philosophies and implementation can be quite different. The SCDM, for example, is designed to capture and structure data from source systems, using native coding systems (e.g., ICD, HCPCS, CPT) with minimal need to transform or map original values to other values or systems. The Observational Medical Outcomes Partnership (OMOP) CDM, by contrast, employs more use of derived fields and its own Standard Vocabulary to which original source system values are to be mapped [11, 15]. Neither of these design philosophies or implementations are necessarily superior to the other; rather they represent varied approaches to achieving analytic goals.

The SCDM tables and their key data elements are depicted in Figure 2. The SCDM is a detailed, patient-level schema that consists of a suite of several tables. Six of the tables (those in the first row of Figure 2) are considered ‘core’ and are present and populated across all Sentinel data partner sites that contribute data to the SDD. There are additional tables that are considered ancillary (those in the second row of Figure 2), as they are not present or populated at all Sentinel data partner sites.



Enrollment
Person ID
Enrollment start & end dates
Drug coverage
Medical coverage
Medical record availability

Demographic
Person ID
Birth date
Sex
ZIP code

Dispensing
Person ID
Dispensing date
National drug code (NDC)
Days supply
Amount dispensed

Encounter
Person ID
Service date(s)
Encounter ID
Encounter type & provider
Facility

Diagnosis
Person ID
Service date(s)
Encounter ID
Encounter type & provider
Diagnosis code & type
Principal discharge diagnosis

Procedure
Person ID
Service date(s)
Encounter ID
Encounter type & provider
Procedure code & type

Lab Result
Person ID
Result and specimen collection dates
Test type, immediacy & location
Logical Observation Identifiers Names and Codes (LOINC ®)
Test result & unit

Vital Signs
Person ID
Measurement date and time
Height and weight
Diastolic & systolic BP
Tobacco use & type

Inpatient Pharmacy
Person ID
Administration date and time
Encounter ID
National Drug Code (NDC)
Route
Dose

Inpatient Transfusion
Person ID
Blood product code and type
Encounter ID
Blood type
Administration start and end dates and times

Death
Person ID
Death date
Source
Confidence

Cause of Death
Person ID
Cause of death
Source
Confidence

Figure 2: Sentinel Common Data Model structure.

STANDARDIZED DATA QUALITY ASSESSMENT

The purpose of any data quality assessment initiative is to ensure that data intended to be used for a particular purpose are fit to be used for that purpose. Data in a distributed data network, by design, serve multiple studies and projects, and thus are intended to be multi-purpose. Data quality assessment endeavors within distributed data networks therefore tend to be more expansive, general and standardized, though likely less specific and directed, than data quality assessment endeavors for one-off or singularly-focused research projects or studies.

The Sentinel suite of SAS programs that assess the quality of the data in the SDD consists of approximately 1,200 individual data checks [16]. Multiple stakeholders (e.g., epidemiologists, data scientists, biostatisticians, programmers) work together to establish and define the required checks and resultant metrics. Each data check has its own unique code and description, and checks are categorized into four different levels of complexity.

- Level 1 checks include single-variable, basic SCDM compliance checks. These ensure that the tables and fields populated by a data partner contain data that conform to the formats specified in the SCDM schema (e.g., data types, variable lengths, acceptable values, etc.), and that no missing data are encountered where none should exist.
- Level 2 checks assess multiple and/or cross-variable compliance, to ensure the integrity of data values within a variable or between two or more variables, within and between tables (e.g., ensuring that some fields are populated only if other fields have certain values).
- Level 3 checks examine distributions and trends over time, both within a data partner's database (by examining output by year and year/month) and across a data partner's databases (by comparing updated SCDM tables to previous versions of the tables).
- Level 4 checks include data logic checks that examine the occurrence of nonsensical diagnoses or care practices (e.g., the proportion of prostate cancer diagnoses among women).

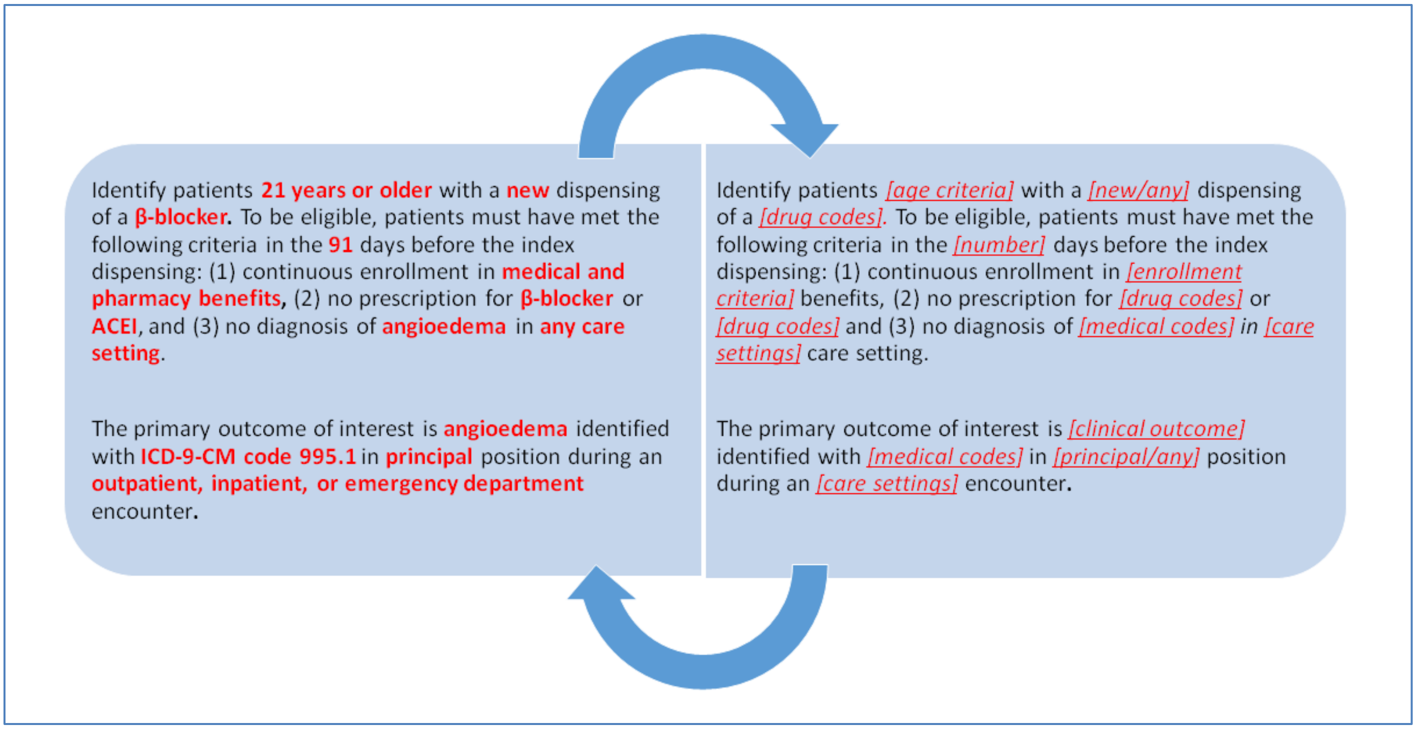
These checks are performed at every data partner refresh, in order to assess whether the data in the SDD meet some minimum and reasonable quality standards to support Sentinel studies. Some Sentinel studies may perform additional quality checks, above and beyond these standard checks, that are more specific to the goals of their particular study.

Performing standardized data quality assessments of all of the data in the SDD is important because these data support studies that inform regulatory decision-making processes of a Federal agency. The data therefore are checked in accordance with established standard operating procedures and best-practices, and in a way that is transparent to both the agency as well as the public.

ANALYTIC TOOL DEVELOPMENT

Fundamental approaches to building analytic tools for infrastructure purposes in a distributed data network include recognizing analytic- and programming-approach patterns where they exist, routinizing analytic programming projects and tasks whenever possible, and approaching all programming tasks from the perspective of the six fundamental characteristics of analytic infrastructure (see Figure 2).

Figure 3 is a graphical representation of approaching analytic tool development from an infrastructure perspective, rather than from a one-off study perspective. This method take a specific study question, such as the one on the left in Figure 3, and converts it to its most fundamental analytic *components*, in accordance with the design on the right.



Identify patients **21 years or older** with a **new** dispensing of a **β-blocker.** To be eligible, patients must have met the following criteria in the **91** days before the index dispensing: (1) continuous enrollment in **medical and pharmacy benefits**, (2) no prescription for **β-blocker** or **ACEI**, and (3) no diagnosis of **angioedema** in **any care setting**.

The primary outcome of interest is **angioedema** identified with **ICD-9-CM code 995.1** in **principal** position during an **outpatient, inpatient, or emergency department** encounter.

Identify patients *[age criteria]* with a *[new/any]* dispensing of a *[drug codes]*. To be eligible, patients must have met the following criteria in the *[number]* days before the index dispensing: (1) continuous enrollment in *[enrollment criteria]* benefits, (2) no prescription for *[drug codes]* or *[drug codes]* and (3) no diagnosis of *[medical codes]* in *[care settings]* care setting.

The primary outcome of interest is *[clinical outcome]* identified with *[medical codes]* in *[principal/any]* position during an *[care settings]* encounter.

Figure 3: Graphical representation of approaching analytic tool development from an infrastructure perspective.

Differences in implementation of these approaches in programming code can be explained as the difference between programming code that contains hard-coded study-design values embedded in the code, versus programming code that is entirely parameter-, data- and/or table-driven.

Executing analytic programs at multiple sites within a distributed data network is done with a “one program, multiple sites” approach. That is, each site has transformed their data and had them quality-checked to be in compliance with a CDM, therefore each site can execute the same programming code on their data.

The ”one program, multiple sites” approach ensures that the same analysis is performed at all participating sites, which means that the analytic output from each site is created from the same programming code, is standardized in its structure, and is therefore easily aggregated. In a distributed data network, each participating site in a study is one piece of the whole. In order to create that whole, the data aggregated from each of the pieces must be standardized and structurally identical.

The primary analytic tool that answers the majority of Sentinel study questions is the Cohort Identification and Descriptive Analysis (CIDA) system. CIDA is designed as a suite of about 35 highly-parameterized SAS macros, each performing a distinct function, making it a consistent, flexible, reusable, scalable and easily maintainable system [17]. The CIDA system has the capability to flexibly identify and extract cohorts of patients from SCDM-formatted data based on several dozen user-defined options and parameters (e.g., study dates, exposure definition, outcome definition, incidence criteria, inclusion/exclusion criteria, continuous enrollment requirements, relevant age groups, and so forth). Nearly all aspects relating to the identification of a study cohort have been built into the system as user-specified macro parameters or as user-supplied input files. CIDA is capable of producing not only descriptive analytics but also includes two confounder-adjustment approaches, one by design (self-controlled risk interval design) and one by analysis (propensity score-based methods), to produce point estimates based on inferential methods.

All programming code, as well as results from most studies/queries utilizing CIDA, are posted to the public Sentinel website; contributing to transparency and reproducibility.

CONCLUSION

Healthcare-related distributed data networks based on claims and EHR/EMR data sources have matured to the point of being able to serve as solid blueprints for data sharing and analytic development initiatives that involve other healthcare-related data streams, including but not limited to clinical trial, genomics or social media sources. Distributed data network design and architecture embody six foundational characteristics of analytic infrastructure. Using the Sentinel Initiative as an example, this paper discussed the enormous potential that distributed data networks hold for population health analytic infrastructure to allow access to greater volumes of data, without sacrificing privacy or analytic capabilities.

REFERENCES

[1] “Mini-Sentinel Common Data Model: Guiding Principles, Version 1.0.” 2010. https://www.sentinelinitiative.org/sites/default/files/data/DistributedDatabase/Mini-Sentinel_CommonDataModel_GuidingPrinciples_v1.0_0.pdf. Accessed Mar 5 2017.

[2] Popovic, J.R. 2015. "Distributed data networks: A paradigm shift in data sharing and healthcare analytics." Proceedings of the 2015 Pharmaceutical Industry SAS Users Group Conference, Orlando, FL. http://www.pharmasug.org/proceedings/2015/HA/PharmaSUG-2015-HA07.pdf. Accessed Nov 30 2015.

[3] Food and Drug Administration Amendments Act of 2007, Pub. L. no. 110-85, Page 121 Stat. 944 (2007). http://www.gpo.gov/fdsys/pkg/PLAW-110publ85/html/PLAW-110publ85.htm. Accessed Oct 20 2013.

[4] Mehzar M. 2016. “Woodcock: Drug Safety Surveillance System Ready for Full Operation.” http://raps.org/Regulatory-Focus/News/2016/02/03/24248/Woodcock-Drug-Safety-Surveillance-System-Ready-for-Full-Operation/. Accessed Oct 2 2016.

[5] Behrman, R.E., J.S Benner, J.S. Brown, M. McClellan, J. Woodcock & R. Platt. 2011. "Developing the Sentinel System - A National Resource for Evidence Development." N Engl J Med; 364:498-499.

[6] “Sentinel Data”. 2017. https://www.sentinelinitiative.org/sentinel/data. Accessed Mar 5 2017.

[7] Popovic, J.R. 2017. “Distributed data networks: a blueprint for Big Data sharing and healthcare analytics.” Ann. N.Y. Acad. Sci., 1387: 105–111. doi:10.1111/nyas.13287.

[8] Sherman, R.E, and R.M. Califf. 2016. “What We Mean When We Talk About EvGen Part I: Laying the Foundation for a National System for Evidence Generation.” FDA Voice. https://blogs.fda.gov/fdavoice/index.php/2016/04/what-we-mean-when-we-talk-about-evgen-part-i-laying-the-foundation-for-a-national-system-for-evidence-generation/. Accessed Mar 5 2017.

[9] Sherman, R.E, and R.M. Califf. 2016. “What We Mean When We Talk About EvGen Part II: Building Out a National System for Evidence Generation.” FDA Voice. https://blogs.fda.gov/fdavoice/index.php/2016/05/what-we-mean-when-we-talk-about-evgen-part-ii-building-out-a-national-system-for-evidence-generation/. Accessed Mar 5 2017.

[10] Califf, R.M. 2017. “Introducing IMEDS, a Public-Private Resource for Evidence Generation.” FDA Voice. https://blogs.fda.gov/fdavoice/index.php/2017/01/introducing-imeds-a-public-private-resource-for-evidence-generation/. Accessed Mar 5 2017.

[11] Observational Medical Outcomes Partnership: Common Data Model v5.0. 2014. http://omop.org/CDM. Accessed Mar 5 2017.

[12] "Sentinel Common Data Model v6.0." 2016. Sentinel Distributed Database and Common Data Model. https://www.sentinelinitiative.org/sites/default/files/data/DistributedDatabase/Sentinel_Common-Data-Model.xlsx. Accessed Mar 5 2017.

[13] “Health Care Systems Research Network Virtual Data Warehouse Data Model.” http://www.hcsrn.org/en/Tools%20&%20Materials/VDW/VDWDataModel/VDWSpecifications.pdf. Accessed Mar 5 2017.

[14] “PCORnet Common Data Model (CDM)”. 2015. http://www.pcornet.org/wp-content/uploads/2014/07/2015-07-29-PCORnet-Common-Data-Model-v3dot0-RELEASE.pdf. Accessed Mar 5 2017.

[15] Xu, Y., Zhou, X., Suehs, B.T. et al. 2015. “A Comparative Assessment of Observational Medical Outcomes Partnership and Mini-Sentinel Common Data Models and Analytics: Implications for Active Drug Safety Surveillance” Drug Saf 38: 749. doi:10.1007/s40264-015-0297-5.

[16] “Sentinel Data Quality Review and Characterization Programs v3.3.4”. 2017. https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/112. Accessed Mar 5 2017.

[17] “Routine Querying Tools (Modular Programs)”. 2016. Routine Querying System. https://www.sentinelinitiative.org/sentinel/surveillance-tools/routine-querying-tools/routine-querying-system. Accessed Mar 5 2017.

ACKNOWLEDGMENTS

The author would like to extend a special thank you to all staff at the Sentinel Operations Center (SOC) at Harvard Medical School/Harvard Pilgrim Health Care Institute (HMS/HPHCI). The expertise and ideas of my colleagues were invaluable to the development of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jennifer R. Popovic, DVM, MA
Harvard Medical School and Harvard Pilgrim Health Care Institute
617.867.4811
Jennifer_Popovic@harvardpilgrim.org