Paper 3270-2019

Multi-site Public and Population Health Research: Analytic Lessons From Distributed Data Networks

Jennifer R. Popovic, RTI International

ABSTRACT

Real-world data—such as health insurance claims, data from providers' electronic health record systems, and disease registries—are rich sources of potential evidence that are used to inform topics from public health surveillance to comparative effectiveness research. Data sourced from individual sites can be limited in their scope, coverage and statistical power. Pooling data from multiple sites and sources, however, present provenance, governance, analytic and patient privacy challenges.

Distributed data networks resolve many of these challenges. A distributed data network is a system for which no central repository of data exists. Data are instead maintained by and reside behind the firewall of each data-contributing partner in a network, who transform their data into a common data model and permit indirect access to those data via the use of a standard query approach.

This paper discusses the contributions of several national-level distributed data networks, including the Sentinel Initiative, the National Patient-Centered Clinical Research Network, and the Observational Health Data Sciences and Informatics program. Focus is placed on the analytic infrastructure—the common data models and reusable analytic tools—each network has developed to support their scientific aims.

This paper also considers how organizations that are not members of these networks can adopt or adapt what has been achieved at the national-level for the development of their own analytic infrastructure.

INTRODUCTION

Public health research and surveillance efforts have been making use of data from multiple, and often disparate, sources for many years. There are myriad challenges involved with multi-site, multi-source studies.

From a governance perspective, there are questions and challenges around where the data come from, who owns them, who can use them, and for what purpose(s). From an analytic perspective, there are challenges around data standardization, harmonization, meaning, and interpretation. From a privacy perspective, there are concerns around preserving patient identity while still maintaining analytic integrity. Distributed data networks, and the guiding principles by which they operate, resolve many of these challenges.

DISTRIBUTED DATA NETWORKS

A distributed data network is a system 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 (bring the analytics to the data, rather than the data to the analytics),
- 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 leveraging 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]. The system is intended to augment, not replace, FDA's existing post-market safety monitoring systems [4, 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) [4].

The SDD currently consists of quality-checked data held over a dozen partner organizations, which are either health insurers, integrated delivery systems or provider networks. Data partners map 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 2019 [6], the SDD contained data on:

- 66.9 million members with medical and drug coverage currently accruing new data
- 292.5 million cumulative patient identifiers between 2000 and 2017
- 14.4 billion pharmacy dispensings
- 13.3 billion unique medical encounters
- 45.6 million members with at least one laboratory test result

FDA initiates queries of the data in the SDD using a suite of reusable, flexible, fully parameterized SAS-based analytic tools. Results from these queries are used to support regulatory decision-making [7].

OTHER DISTRIBUTED DATA NETWORKS IN EXISTENCE

Several other healthcare-related distributed data networks exist, and some of these networks have a particular focus. 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. These 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)
- Observational Health Data Sciences and Informatics (OHDSI) program

All of these networks are built on the philosophy and architecture of a distributed/federated database; that is, none have a central repository of patient-level data and all employ a common data model. 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 WITHIN DISTRIBUTED DATA NETWORKS

Analytic infrastructure is defined as the systems, processes, governance, data, software, tools and people that facilitate the analytic process [8].

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 six foundational characteristics of analytic infrastructure, as shown in Figure 1.

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. There are several healthcare-related common data models in existence, including:

- U.S. Food and Drug Administration's (FDA) Sentinel Common Data Model (SCDM)[1]
- National Patient-Centered Clinical Research Network (PCORNet) Common Data Model[2]
- Health Care Systems Research Network Virtual Data Warehouse (HCSRN VDW) Common Data Model[3]
- Observational Health Data Sciences and Informatics' Observational Medical Outcomes Partnership (OHDSI OMOP) Common Data Model[4]
- Clinical Data Interchange Standards Consortium's (CDISC) multiple common data models and standards[5]

Although CDMs are similar in their goal of standardizing the capture and storage of data elements from various source systems, the design philosophies and implementations can be quite different. Additionally, each of these CDMs was developed for a different purpose, to capture data from different sources. For example, FDA's SCDM is structured to prioritize capture of data elements that are primarily health insurance claims-based; PCORNet's CDM is more clinical and patient-reported outcomes focused; the OMOP CDM is more clinical/EHR-focused; CDISC's models and standards are most attuned to capturing data generated from clinical trials.

In general, CDM philosophies, designs, strengths and weaknesses can be identified and summarized by understanding why each of those models came to exist in the first place. For example, the FDA's SCDM was designed to capture and structure data from health insurance 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 OMOP CDM, by contrast, employs more use of derived fields and its own Standard Vocabulary to which original EHR-based source system values are to be mapped [9]. Neither of these design philosophies or implementations are necessarily superior to the other; rather they represent varied approaches to achieving analytic goals. Many of these models also leveraged each other. For example, the SCDM is in-part based on the HCSRN VDW CDM, and the PCORNet CDM is in-part based on the Sentinel CDM [10].

---

[1] https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model

[2] https://pcornet.org/pcornet-common-data-model/

[3] http://www.hcsrn.org/en/Tools%20&%20Materials/VDW/

[4] https://www.ohdsi.org/data-standardization/

[5] https://www.cdisc.org/standards

Other open-source, common data model initiatives have evolved more recently. These are focused more on facilitating collection and standardization of data elements to support clinical decision-making and care, as opposed to supporting secondary-use research objectives. These initiatives include:

- American Medical Association's (AMA) Integrated Health Model Initiative (IHMI)[6]
- MITRE Corporation's Standard Health Record Collaborative (SHRC)[7]

It is beneficial for organizations seeking to develop data and analytic infrastructure to be aware of these existing CDM and standardization efforts, including their varying philosophies, goals, strengths and weaknesses. Organizations seeking to structure data from disparate sources into a CDM have the option to adopt an existing structure or adapt one for their own unique needs and uses [11].

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 standard data quality assessment approach in Sentinel, for example, incorporates various checks across four categories. Each data check has its own unique code and description, and checks are categorized into four different levels of complexity [12].

- Level 1 checks are basic, single-variable SCDM compliance checks.
- 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).

Performing standardized data quality assessments of all of the data in a distributed database is important because the data, by definition, are used to support numerous studies for a variety of purposes (e.g., regulatory decision-making, comparative effectiveness research).

ANALYTIC TOOL DEVELOPMENT

[6] https://www.ama-assn.org/amaone/integrated-health-model-initiative-ihmi

[7] http://standardhealthrecord.org/#sitehomepage

Tools are analytic programs designed to answer types of questions, as opposed to specific questions and can be thought of as “off-the-shelf” or “canned” solutions that are designed and developed for use/reuse across multiple projects or initiatives.

Fundamental approaches to building analytic tools for infrastructure purposes 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 six fundamental characteristics of analytic infrastructure (see Figure 1).

Analytic tool development efforts differ from custom, one-off programming efforts in that, when we build tools, we dissect specific study questions/analyses into their component pieces, to transform them into more generalized types of questions/analyses. Differences in coding implementation between these approaches may 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 [10, 13, 14]. Tool-development approaches use the latter philosophy and are typically designed and developed to be easily reusable across projects or initiatives.

Figure 2 is a graphical representation of approaching analytic tool development from an infrastructure perspective, rather than from a one-off perspective. This method takes a specific study question, such as the one on the left, and converts it to its most fundamental analytic components, in accordance with the design on the right [10].

The approach to developing code that is reusable, flexible and scalable typically requires a design phase before any programming code is developed. During a design phase, analytic developers devote time and resources to brain-storming what they need for their program(s) to do, what key features and options their program will need to include, what parameters are needed to achieve those features/options, who are the intended end-users of their tools, how to develop their tools for easy maintenance, expansion and up-versioning, and so forth. Only after a well thought-through design is completed and drafted should any programming commence. This is akin to an architectural phase that includes drawing up plans before any actual construction can begin.

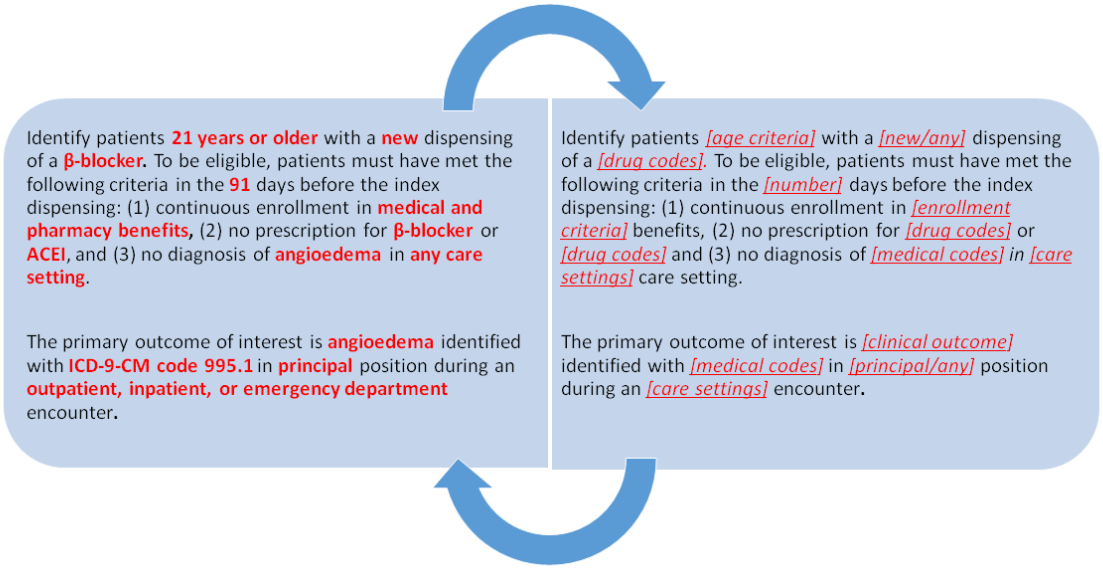

Figure 2: Graphical representation of approaching analytic tool development from an infrastructure perspective.

ANALYTIC INFRASTRUCTURE WITHIN RESEARCH ORGANIZATIONS

Organizations that are members of one or more of these distributed data networks participate in studies and queries that initiate within those networks, and they are also permitted to use their data that they have mapped to a network common data model structure for their own internal research purposes [8]. Thus, being part of a network provides organizations with analytic infrastructure to support their own non-network-related research agendas.

Organizations that do not participate in a distributed data network can also take advantage of the infrastructure that the networks have developed, because much of that infrastructure—in particular the common data model schemas and the analytic tools that are designed to be executed against data structured in that manner—are freely and publicly available.

There are many reasons why a research or healthcare organization may want to adopt or adapt what has been designed at the national level, for their own internal research purposes.

From an internal organizational perspective, an organization that invests resources in developing infrastructure is investing in itself. Approaching data and analytic work from an infrastructure-building perspective leaves an organization and its staff with capabilities and capacity that continue to contribute to that organization's mission and institutional knowledge. Conversely, approaching data and analytic development work from a one-off perspective often leads to inefficient development and re-work. One-off approaches seldom contribute to building institutional knowledge, capabilities and capacity.

From an external perspective, an organization that invests resources in developing infrastructure is developing a strong scientific foundation on which their research will reside. Approaching data and analytic work from an infrastructure-building perspective often results in work that is transparent, reproducible and reusable on future project work. Approaching data and analytic development work from a one-off perspective, without regard for what other like-projects within an organization may be doing, can lead to a lack of scientific defensibility and transparency to clients, particularly if divergent analytic approaches, algorithms, and methods are being used for the same client across separate but similar projects.

BENEFITS TO DIFFERENT ORGANIZATIONAL ROLES

The benefits of designing, building and maintaining analytic infrastructure are multifaceted and valuable for multiple roles within an organization, including individual contributor analytic staff, individual project investigators or directors, as well as management and the organization as a whole. These benefits are discussed below and summarized in Figure 3 [15].

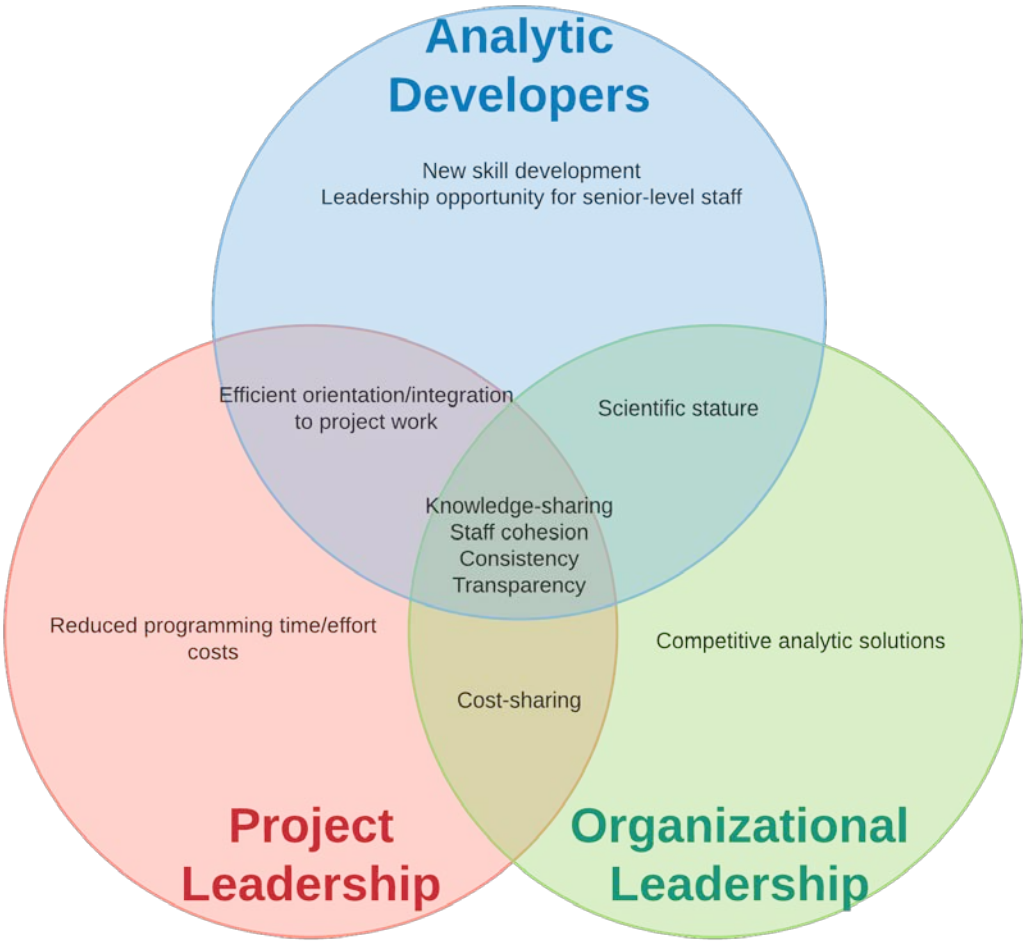

Figure 3: Benefits of building analytic infrastructure

For individual contributor analytic staff, there are many benefits to approaching data and analytic development work from an infrastructure-building perspective. First, this perspective creates opportunities for analytic programmers and researchers to develop new skills by challenging them to develop flexibly-designed and reusable tools, rather than solely one-off programming efforts that have limited capability for reuse. Second, designing and developing infrastructure supports a culture of sharing ideas, analytic approaches and programming code, providing opportunity for cohesion across analytic staff as they contribute to common initiatives and goals. Third, infrastructure-building may create opportunity for junior-level staff to become involved with scientific projects more quickly via the use of already-developed data and analytic tools and resources. Simultaneously, this may create opportunity for senior- and lead-level developers to orient and train junior-level staff on common analytic methods and approaches, while allowing senior- and lead-level developers more time to focus on infrastructure development projects and initiatives. Lastly, infrastructure-building initiatives create opportunity for individual analytic development staff to contribute to the public body of knowledge and scientific stature of the organization via conference presentations, journal publications or other professional communication modalities.

For project investigators or directors, there are similarly several benefits to approaching work from an infrastructure-building perspective. These approaches have the potential to reduce project staff time, effort and funds spent on planning, developing, testing and validating analytic programing code that is developed for commonly used methods and approaches (e.g., not reinventing the wheel). Infrastructure-building approaches also promote analytic consistency and transparency within and across projects by building and maintaining reusable, well-documented solutions for common analytic tasks and approaches.

For management and the larger organization, the benefits of building analytic infrastructure are similar to and harmonious with the benefits for other staff roles. These approaches promote and invest in a model of data and software development techniques and implementations that results in a library of off-the-shelf analytic solutions for commonly-used methods and analytic approaches. These approaches therefore have the potential to build the resume of organizational analytic capabilities and to position the organization to be competitive on submissions for future analytic and data-centric work. Infrastructure-building efforts also contribute to organizational scientific stature by show-casing data and analytic approaches and tools in a public forum (e.g., conferences, proceedings, journals).

CONCLUSION

Distributed data network design and architecture embody six foundational characteristics of analytic infrastructure. This paper discussed the enormous contributions that distributed data networks have made to support multi-site/multi- public and population health research, to allow access to greater volumes of data without sacrificing privacy or analytic capabilities.

Organizations that are not members of one of these networks can adopt or adapt what has been achieved at the national-level for the development of their own analytic infrastructure. Many staff roles within an organization, from individual contributors to senior management, can contribute to and benefit from investing time and effort to develop organizational analytic infrastructure.

REFERENCES

[1] “Mini-Sentinel Common Data Model: Guiding Principles, Version 1.0.” 2010. https://www.sentinelinitiative.org/sites/default/files/data/distributed-database/Mini-Sentinel_CommonDataModel_GuidingPrinciples_v1.0_0.pdf. Accessed Apr 1 2019.

[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 Apr 1 2019.

[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 Apr 1 2019.

[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 Apr 1 2019.

[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”. 2019. https://www.sentinelinitiative.org/sentinel/data/snapshot-database-statistics. Accessed Apr 1 2019.

[7] “Sentinel Assessments”. 2019. https://www.sentinelinitiative.org/drugs/assessments. Accessed Apr 1 2019.

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

[9] 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.

[10] Popovic, J. R. 2017. “Healthcare data sharing and innovative analytic development: Lessons learned from distributed data networks.” Proceedings of the 2017 SAS Global Forum (Paper 840-2017). http://support.sas.com/resources/papers/proceedings17/0840-2017.pdf . Accessed Apr 1 2019.

[11] Garza, M., Del Fiol, G., Tenenbaum, J., Walden, A., Zozus, M. 2016. “Evaluating common data models for use with a longitudinal community registry.” Journal of Biomedical Informatics, Volume 64, Pages 333-341.

[12] “Sentinel Data Quality Assurance Practices”. 2017. https://www.sentinelinitiative.org/sites/default/files/data/distributed-database/Sentinel_DataQAPractices_Memo.pdf. Accessed Apr 1 2019.

[13] Hughes, T. M. 2016. SAS Data Analytic Development: Dimensions of Software Quality. Cary, NC: SAS Institute.

[14] Nelson, G. S. and Zhou, J. 2012. “Good Programming Practices in Healthcare: Creating Robust Programs.” Proceedings of the 2012 SAS Global Forum (Paper 412-2012). http://support.sas.com/resources/papers/proceedings12/417-2012.pdf. Accessed Apr 1 2019.

[15] Popovic, J.R. 2018. "Infrastructure for healthcare analytics: A foundational approach to development." Proceedings of the 2018 Pharmaceutical Industry SAS Users Group Conference, Orlando, FL. https://www.pharmasug.org/proceedings/2018/LD/PharmaSUG-2018-LD20.pdf. Accessed Apr 1 2019.

CONTACT INFORMATION

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

Jennifer R. Popovic, DVM, MA
RTI International
781.434.1767
jpopovic@rti.org