Paper 1969-2018

Infrastructure for healthcare analytics: A foundational approach to development

Jennifer R. Popovic, RTI International

ABSTRACT

Healthcare-related organizations and the data they generate are notoriously siloed, with varied systems, structures, data standards, and governance processes and policies. Research teams and organizations analyzing these data can be similarly siloed and challenged in their ability to create robust analytic infrastructure that would enable the sharing of their most valuable analytic assets: data and the analytic expertise and methods, approaches and tools to analyze them. These challenges often lead to inefficient analytic development based on one-off approaches that do not prioritize or promote the ideas of reusability, flexibility, scalability or reproducibility.

This paper discusses the concept of analytic infrastructure and will discuss strategies that research teams and organizations can use in their approaches to develop infrastructure. We discuss the unique expertise and contributions that various staff engaged in healthcare research or health services research can provide to the development of organizational-level analytic infrastructure, including management, research investigators and analytic developers. We discuss strategies and provide examples of various artifacts that can be foundational to organizational analytic infrastructure, including the roles of common data models, flexibly-designed analytic tools, and knowledgebases.

INTRODUCTION

The development of organizational analytic infrastructure, particularly within research organizations engaged in healthcare analytics, can be a worthy investment of time and resources.

This paper presents a definition of analytic infrastructure, examines the benefits of analytic infrastructure to multiple staff roles across an organization, and discusses specific ways that staff within research organizations can contribute to designing and developing analytic infrastructure. This paper also provides readers with an explanation of the differences between “one-off” data and analytic development approaches versus infrastructure-building approaches, as well as when and why the latter may be considered in favor of the former.

ANALYTIC INFRASTRUCTURE

Analytic infrastructure is defined as the systems, processes, governance, data, software, tools and people that facilitate the analytic process (Popovic 2017b). This paper does not discuss all aspects of analytic infrastructure, but rather focuses on six characteristics considered to be foundational to developing, maintaining and growing analytic infrastructure (Hughes 2016, Popovic 2017a, Popovic 2017b). These six characteristics are shown in Figure 1.

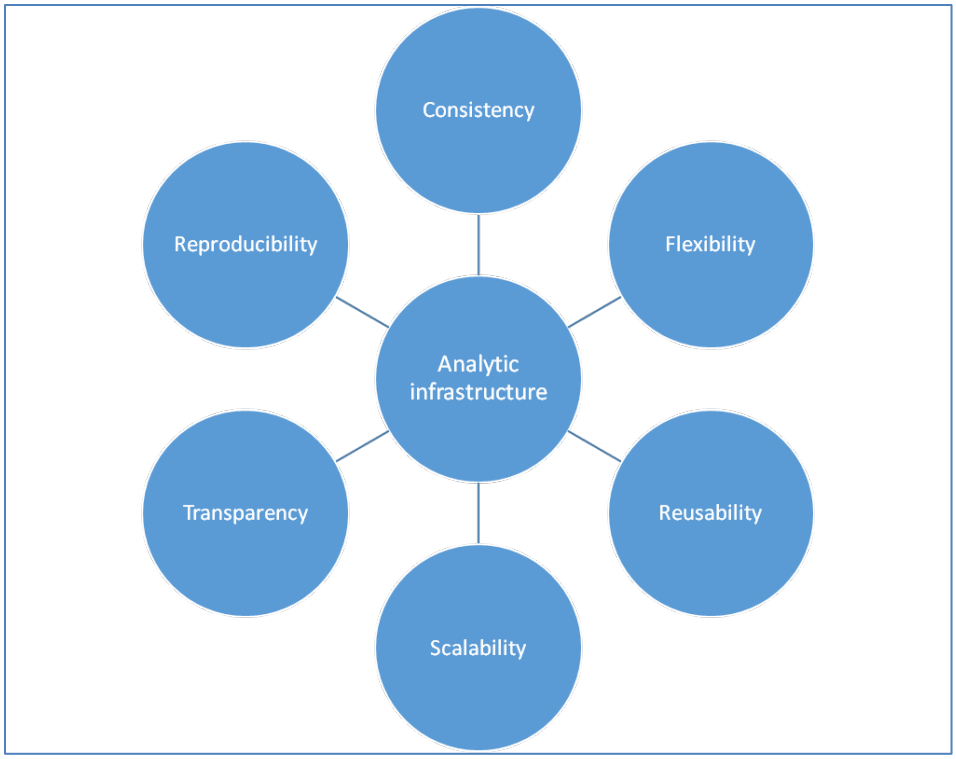

Figure 1: Six foundational characteristics of analytic infrastructure

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 for review and use.

RATIONALE FOR DEVELOPING ANALYTIC INFRASTRUCTURE

There are internal and external organizational reasons for developing analytic infrastructure. These are discussed here and should be considered synergistic rather than mutually exclusive.

From an internal 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 your 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 for these different roles and summarized in Figure 2.

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).

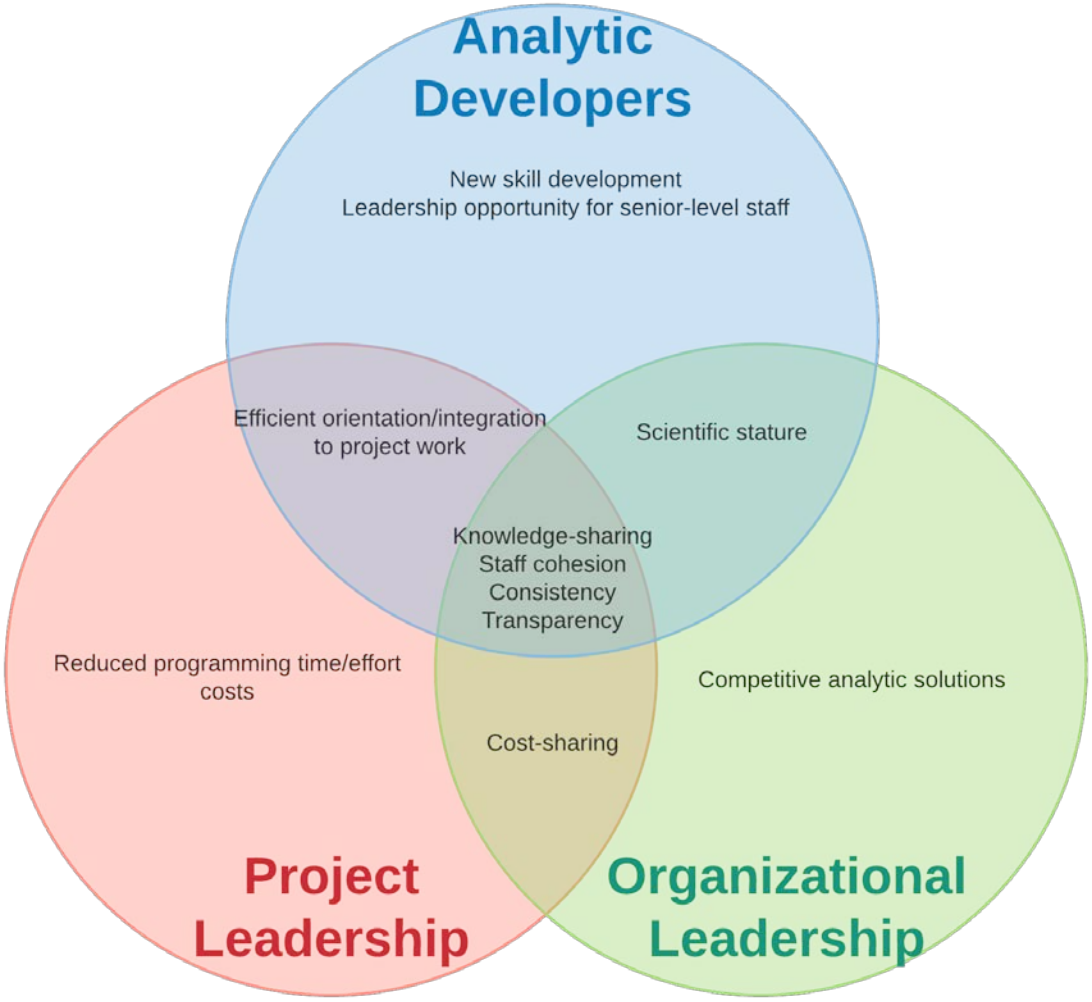

Figure 2: Benefits of building analytic infrastructure

FORCES THAT INHIBIT INFRASTRUCTURE DEVELOPMENT

Despite the cross-organizational benefits of investing time and resources in developing analytic infrastructure, several forces exist that may inhibit an organization's willingness or ability to do so.

van Panhuis, et al (2014) conducted a systematic literature review to identify potential barriers to data sharing between and across public health organizations. The authors identified and grouped these barriers into a taxonomy consisting of twenty specific barriers within six categories. The six categories included: technical, motivational, economic, political, legal and ethical. Though the authors focused on barriers to data sharing between and across public health organizations, their framework and findings are relevant and applicable to the lack of resource-sharing, and subsequent lack of infrastructure- and institutional knowledge-building initiatives, that healthcare research organizations experience.

Four of the six categories are particularly relevant to the focus of this section. It is important to note that there are interactions across the four categories; they are not necessarily mutually exclusive. These categories are:

- Technical

- Motivational
- Economic
- Legal

TECHNICAL

Technical barriers to developing analytic infrastructure within healthcare research organizations can take on several forms, and the most relevant for this paper is the lack of data standards and structures across like-datasets.

One example of this is in the various data structures found across health insurance claims-based datasets, from state all-payer claims databases (APCDs) to commercially-available claims databases to the myriad databases and datasets available from the Centers for Medicare and Medicaid Services (CMS). State APCDs capture similar data elements—those found on standard claim forms, as well as information about patient enrollment, provider characteristics, and so forth—but no common data model standards exist across states to allow researchers to approach each in a consistent analytic manner. On the federal side, even within just CMS, there are many different systems and datasets, and by extension data structures, that capture similar claims-based data elements with inconsistent data structure standards in terms of dataset names, structures (e.g., units of analysis, primary keys), field names, and so forth.

This lack of standards presents challenges to research organizations that seek to use multiple data sources and wish to approach them with standardized analytic tools. Standardized tools can only be designed and developed to run on multiple data sources when those sources are structurally identical.

One solution to this technical barrier is organizations' adoption or adaptation of a common data model structure to apply to similar data from disparate sources, such as the health insurance claims-based example above. Common data models and their potential utility are discussed in the next section of this paper.

MOTIVATIONAL

Several motivational forces can create disincentives to building analytic infrastructure. For example, the development of common data model structures and standardized analytic tools to apply to them may require project investigators, analytic developers and management teams to agree, and perhaps compromise, on a set of standards. Forces that may interfere with consensus on standards include questions about scientific credit and issues of scientific discord as to the most appropriate standards to adopt.

The right balance of incentives across multiple staffing layers needs to be created for organizations and their staff to realize the advantages of creating analytic infrastructure that benefits many across an organization as opposed to only a few.

ECONOMIC

Project timelines and budgets often compel project teams and organizations to select one-off analytic methods to "get the job done" for the immediate task as opposed to looking for ways to build reusable and flexible solutions and capabilities to fulfill both present and future analytic needs.

Designing and developing data standards and standardized analytic approaches and tools for infrastructure purposes may cost more, initially, than one-off methods. Apprehension about the economics of building analytic infrastructure can be summed up with a question: Who pays for an analytic solution for which there are multiple beneficiaries and for which there may be a delay in return-on-investment?

Similar to the importance of creating the right balance of incentives to counter-balance motivational forces that may work against the pursuit of infrastructure-building initiatives, it is also critical for organizations to identify and create the right balance of economic incentives to be supportive of developing analytic infrastructure.

LEGAL

Acquisition of data for secondary-use research purposes often comes with data use agreements (DUAs). Generally, a DUA is a legal agreement between a data owner and an external party that stipulates who can use the data, for what purpose(s), for how long, and the storage and security conditions that must be adhered to during that period of data use. DUAs can introduce governance challenges and legal barriers to usage of data due to their sometimes-narrow scope.

For example, if a DUA specifies that data can only be used to answer a narrow set of research questions, that could complicate efforts to use those data to develop reusable, flexible, scalable tools, whose options and capabilities may extend outside of the scope outlined in the DUA.

ARTIFACTS THAT CAN BE FOUNDATIONAL TO BUILDING ANALYTIC INFRASTRUCTURE

The design, development and implementation of analytic infrastructure can take on several forms. Some valuable artifacts that organizations can invest in developing and building include: common data models, flexible-designed analytic tools that are intended for reuse, and knowledgebases within which institutional knowledge can be codified and shared. These three artifacts are discussed in more detail below.

COMMON DATA MODELS

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]

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

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

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

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

- 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 implementation 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 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 likely be identified and summarized by knowing 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 (Xu, Zhou, et al 2015). 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 (Popovic 2017b).

There are also additional open-source, common data model initiatives that 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 in a CDM have the option to adopt an existing structure or adapt one for their own unique needs and uses (Garza, Del Fiol, et al 2016).

FLEXIBLY-DESIGNED ANALYTIC TOOLS

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.

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 the 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

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

[6] https://ama-ihmi.org/

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

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 (Hughes 2016; Nelson and Zhou 2012; Popovic 2017b). Tool-development approaches use the latter philosophy and are typically designed and developed to be easily reusable across projects.

Figure 3 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 (Popovic 2017a).

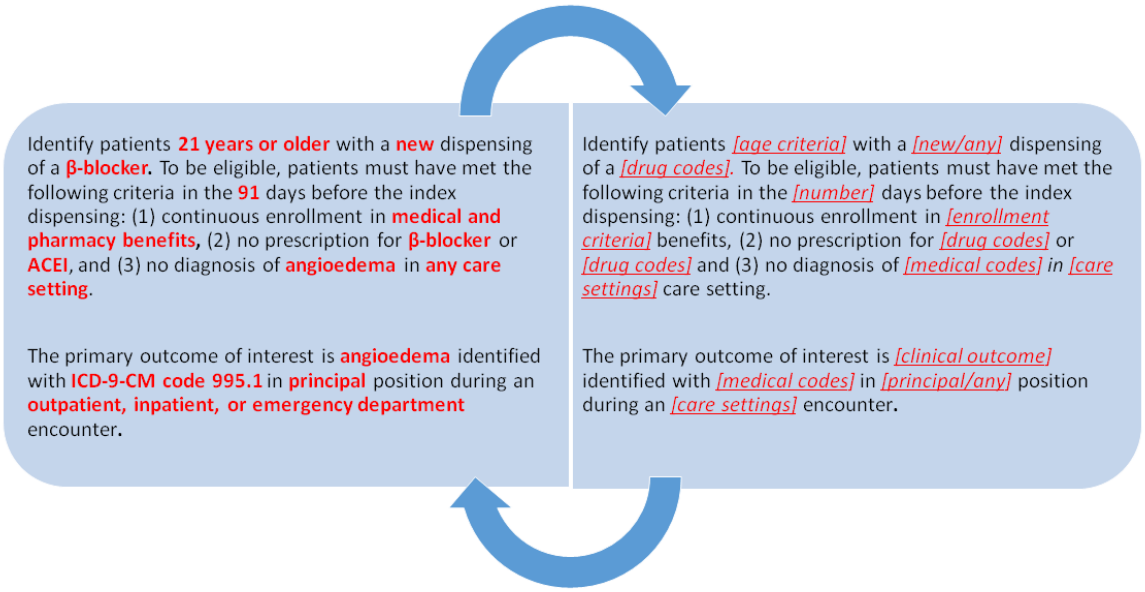

Figure 3: Graphical representation of approaching analytic tool development from an infrastructure perspective.

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.

KNOWLEDGEBASES

Knowledgebases are tools that can be used to formally codify institutional knowledge that is bound within individuals in an organization so that the organization as an entire entity can benefit. It is the act of moving information from individual staff and functional experts into a centralized structure accessible to all who may need it (Ashkenas 2013).

There are many reasons why institutional knowledge is not systematically captured and as many reasons why it may be critical to do so. These reasons include, but may not be limited to:

- Higher staff turn-over rates than in past (e.g., people do not stay at their jobs as long anymore)

- Higher numbers of near-retirement or retirement aged staff in the workforce

While knowledgebases are the tools that can help retain institutional knowledge, they must be adopted with intentional strategy (Hansen, Nohria and Tierney 1999). For example:

- What information will be codified?
- On what schedule will the information be kept up-to-date?
- Who will perform those tasks?
- To whom will the information be made visible/accessible?

Depending on the answers to these questions, suitable knowledgebase software must be identified and selected.

Knowledgebases are not only valuable to current employees but also can be useful on-boarding tools for use with new staff.

ORGANIZATIONAL ROLES ESSENTIAL FOR DEVELOPING ANALYTIC INFRASTRUCTURE

Earlier, we discussed the benefits that building analytic infrastructure could have on various roles within an organization. Here, we discuss the ways in which those same roles can contribute to developing analytic infrastructure within their organizations.

LEADERSHIP/MANAGEMENT

Planning, designing and developing common data models and analytic tools for purposes of building analytic infrastructure may seem counter-intuitive from a return on investment point-of-view when compared with one-off, on-demand custom development. Investments in infrastructure-based initiatives may introduce a flat or negative return on investment realization, at least initially, as the initial time and effort investment required to design, develop and grow infrastructure may outpace any immediate return. A return on investment may only be realized in the intermediate- and longer-term, as the infrastructure initiatives mature and are adopted by projects and teams over time.

In some cases, management may not be aware of the potential return on investment involved with designing and developing analytic infrastructure. Management's understanding of, and support for, the initial costs of building analytic infrastructure is crucial. This may translate to management teams finding strategic investment funds or other cost-sharing strategies with project teams to augment project funds to meet certain analytic goals. This may also translate to management teams being supportive of the extra time it may take to design and develop data structures and/or analytic tools that can be used across projects, as opposed to within just one project.

PROJECT INVESTIGATORS

Project investigators can contribute to the development of analytic infrastructure by agreeing on standardized analytic approaches and algorithms to common computational tasks. For example, if there is not a compelling reason to compute a readmission rate differently across projects, project leadership can come to consensus on a standardized way to implement that analysis. Alternatively, if some computational differences must be retained, project leadership can work with analytic developers to communicate what those are so they can be designed and developed as tool options as opposed to as separate tools or one-off approaches.

ANALYTIC DEVELOPERS

Analytic developers can contribute to developing organizational analytic infrastructure by developing their skills in parameter-, data- and/or table-driven coding strategies, rather than in one-off methods that often rely on hard-coding techniques and do not lend themselves well to reuse or flexibility. Analytic developers can also add other good programming practices and techniques to their individual technical toolboxes, such as defensive coding methods, to make their programs more robust and viable across a wider range of uses (Brucken and Levy 2016, Nelson and Zhou 2012, Pharmaceutical Users Software Exchange 2015).

CONCLUSION

Investing resources in the design and development of analytic infrastructure can be critical to expanding healthcare research organizations' technical capabilities, capacity and scientific stature. These initiatives can also result in analytic cost-saving measures in the long-run.

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

Ashkenas, R. 2013. "How to Preserve Institutional Knowledge." Harvard Business Review. https://hbr.org/2013/03/how-to-preserve-institutional . Accessed March 9, 2018.

Brucken, N. and Levy, D. E. 2016. "Defensive Coding by Example: Kick the Tires, Pump the Brakes, Check Your Blind Spots, and Merge Ahead!". Proceedings of the 2016 Mid-West SAS Users Group (Paper TT12). https://www.mwsug.org/proceedings/2016/TT/MWSUG-2016-TT12.pdf . Accessed March 9, 2018.

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.

Hansen, M. T., Nohria, N. and Tierney, T. J. 1999. "What's Your Strategy for Managing Knowledge?". Harvard Business Review. https://hbr.org/1999/03/whats-your-strategy-for-managing-knowledge? Accessed March 9, 2018.

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

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 March 9, 2018.

Pharmaceutical Users Software Exchange (PhUSE). 2015. "Good Programming Practice Guidance." http://phusewiki.org/wiki/index.php?title=Good_Programming_Practice_Guidance . Accessed March 9, 2018.

Popovic, J. R. 2017a. "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.

Popovic, J. R. 2017b. "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 March 9, 2018.

van Panhuis et al. 2014. "A systematic review of barriers to data sharing in public health." BMC Public Health, 4:1144.

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.

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