In all disciplines, research reproducibility can only be achieved if open research practices are not an afterthought (CHEN et al).

Achievement of research reproducibility requires investigators to shift focus from an afterthought to adoption of research practices and tools that automate the

Underlying the ongoing debate about a lack of reproducibility in biomedical and social science research is the necessity for investigators to shift their focus. Research reproducibility is largely unachievable when it is merely viewed as an afterthought. Altering the research lifecycle to include tools to automate reproducibility will improve its attainability.

**Abstract** Recent years has witnessed an increasing interest in open-science tools such as Linux, GitHub, Python, SQLite, and R. Organizations which offer short-format workshops and seminars – the Carpentries being a prominent player in this space – frequently teach the tools in isolation. A typical Carpentries workshop, for example, covers Git, SQL, Linux, and a programming language in four self-contained sessions over a two-day period. However, what is conspicuously missing is instruction on how to integrate these technologies into a research system which is greater than the sum of its parts. In other words, scientists rarely use tools in isolation from other tools. Thus the ability to envision and then build an integrated research system is just as important as mastering the technical details of individual technologies. Because little has been written about the design of integrated research systems, this article represents an initial effort to rectify this deficiency. It consists of four parts. An introduction to the topic and review of the literature is developed in part one, followed by the presentation of an integrated research system model in part two. An in-depth discussion of the interfaces between the system components is articulated in part three. And finally, the article concludes with some thoughts and best-practice recommendations for research system design.

**Introduction and Literature Review**

By the early 1960’s, there was a growing awareness of systems thinking and the systems perspective. The launch of the IBM 360 in 1964 probably contributed to this phenomenon, given the fact that it was a fully integrated system with broad commercial appeal. It is therefore not surprising that one of the first attempts to view the research process as an integrated system of interrelated parts occurred that same year. Garvey and Griffith’s (1964) article was appropriately titled, *The Structure, Objectives, and Findings of a Study of Scientific Information Exchange in Psychology*. Not only did the authors identify key processes but they identified the relationships between the components through the judicious use of lines. By today’s standards, their research system diagram appears primitive. Nevertheless, Garvey & Griffith deserve to be recognized for the innovative nature of this work. The y-axis of the diagram, for example, displays months prior and subsequent to article publication, thereby giving the reader an idea of what processes are active at a given point in time.

A large step forward in systems thinking occurred four years later, when Bertalanffy’s (1968) *General System Theory* was released. The systems perspective developed in this book has impacted virtually every scientific field, including but not limited to physics, biology, economics, psychology, and many other disciplines. Early work by Jay Forrester deserves to be mentioned as well. Forrester (1961) did foundational work in the area of system dynamics, in both industrial and operations management settings. The value of this early work lies in the emerging realization that systems ought to be understood as living and dynamic organisms, ever changing and adapting to an external environment.

This initial conceptual framework, however, had to wait a decade or so for adequate computing power to instantiate the models. It was not until the late 1970’s – at the start of the personal computer revolution – that scientists had access to affordable computational systems capable of modeling natural systems as well as manage the scientific documents, data, code, and workflows of modern research.

As it turned out, the 1990’s proved to be a pivotal decade, creating a technical platform capable of supporting 21st century research systems. The launch of the Netscape browser in the mid-1990’s moved science one step closer to a vision of global collaboration, but arguably the more important development was the emergence of the open-source movement in the latter part of the decade. Incidentally, Netscape played a formative role here as well, with the decision to release their Navigator source-code into the public domain in early 1998. This led, in part, to the creation of the Open Source Initiative.

Many of the open source tools used in scientific environments – Python, R, Git, SQLite, and Linux were available by the early 2000’s.

The big data revolution, 2010 forward… Systems analysis and design has been an important function in business I.T. departments for some time now.

Ideally, the configuration of an integrated research system will reflect both the research question as well as the research method method(s) used to answer it.

**A Basic Research System Design**

**The Research System Component Interfaces**

**Conclusion and Best-Practice Recommendations**

**Article Titles**

*The reproducibility machine: Designing an integrated big data research system to support precision health researchers*

*Anatomy of an environment wide association study: Making wise decisions at the strategic inflection points*

**Contextualizing the Big Data Crisis**

In the late 1980's, the software industry faced a programming crisis. From the earliest days of commercial programming in the early 1960's, code had been written in a linear fashion. This approach, however, led to increasing complexity, with programs littered with the infamous goto statement. The result was "spaghetti code" which jumped from point to point and was virtually impossible to maintain. The response to this crisis in the 1990's was Object-Oriented Programming, or OOP as it is called.

The object-oriented response to the crisis in programming involved the first two elements listed below, coupled with a disciplined approach to implementation.

* Systems Thinking
* Architecture
* Discipline

Often, a "crisis" is fueled by sales organizations and individuals with financial interests in solving the latest problem. Even so, real issues develop over time, laying the groundwork for a paradigm shift, a phrase made famous by Thomas Kuhn.

Mindset for Reproducibility Success

* Test, test, test
* Aesthetic sensibility
  + Linus Torvalds (Style)
  + 15 Minute Rue
  + Zen

Object-Oriented Programming

* Contract
* Black Box (code encapsulated within an interface)

Intuitive sense of what matters and what doesn't. Complexity raises the probability of model overfitting.

**Quotes**

Freman Dyson (2006) commented, “The great advances in science usually result from new tools rather than from new doctrines” (p. 33). I found this quote in Hunt (1997, p. 19).

**The 21st Century Public Health Research System**

The opportunity is for academic research libraries to design and develop integrated research systems for our clients. Most organizations in the scientific computing space offer training in specific tools such as GitHub, Python, or R -- the Carpentries being the most prominent player. However, few talk about integrating the pieces into a whole which is greater than the sum of its parts. In other words, primary investigators today must not only master the technical details of the tools they use but they must also act as system designers, ensuring that workflow processes are properly designed and maintained. It is rare that one uses a tool in isolation. Normally a variety of tools must be configured to work together. For example, RStudio supports the creation of projects but one also needs to track versions of project components in GitHub or a similar system.

A distinction ought to be made between a public health **research** systems and its counterparts -- the public health information systems which underpin the epidemiological decisions made by public health professionals. The two types of public health systems share some commonalities, but there are key differences as well. For example, the primary users (researchers vs. practicing professionals) differ somewhat in their information and data needs. For the public health professional, actionable information and data is of primary importance. While for the public health researcher, a deeper level of insight and understanding is sought. Where the focus for the practicing professional is tactical, the focus for the researcher is strategic.

A tactical focus - in the case of public health information systems - also entails a robust surveillance capability. This is not usually the case with public health research systems. This, then, is another important distinction between the two kinds of systems. Naturally, surveillance data that is both accurate and granular is critical when seeking to curb an outbreak or epidemic. Such data, however, has the potential of clogging up a research system as much of it is routine, nothing more than the white noise of normality.

Ideally, the research methodology ought to inform - specify even - the way in which data is modelled in a big data system. Public health researchers often wish to explore relationships between different levels. For example, a researcher might wish to understand how state policies impact the delivery of neonatal care in a metropolitan statistical area (MSA). In this case, the MSA is at a lower-level of analysis and most likely contained within a single state at the higher-level. The image of Russian dolls immediately comes to mind, with smaller dolls inside larger ones. The challenge, then, of public health data is creating structures which allow researchers to easily move up and down a data hierarchy, exploring effects at each level as well as interactional effects between levels.

**Big Data**

Public health professionals, as Ola & Sedig (2014) point out, are "overwhelmed with massive amounts of data on a regular basis" and data set overload has "forced many epidemiologists to become data managers" (p. 11 & 12). The PH informatics community has yet to address this issue in an appropriate way, especially in the data visualization space.

Public health data is characterized by high volume, great variety, high velocity, and low veracity (Ola & Sedig, 2014, p. 3). The same can be said of big data more generally, with the addition of velocity. These four attributes (volume, variety, veracity, and velocity) are also referred to as the four V's of big data.

In recommendation # 5 of their article, *Transforming Epidemiology for 21st Century Medicine and Public Health*, Khoury et al (2013) write, "The development of systematic approaches to robustly manage, integrate, analyze, and interpret large complex data sets is **crucial** (p. 14). Furthermore, the authors recognize that this challenge requires not only expertise but political acumen as well.

Technology makes it easier to search out and discover associations in big data environments. Yet this capability also enhances the probability of "discovering" spurious correlations. The same holds true for precision medicine. Khoury, Iademarco, and Riley (2016) write, "As with precision medicine, separating signal from noise will not be easy" (p. 401). In other words, big data is not a panacea because "big error can plague big data" (Khoury & Ioannidis, 2014, p. 1054). The problem lies in determining which signals are real and which are simply false positives. As of now, there is no method for quickly and cost-effectively separating the two, for deciding which leads ought to be followed and which ought to be abandoned. Public health researchers frequently report correlations but then fail to validate their findings. Admittedly, validation can - and often is - an expensive proposition. But without it, the underlying research foundation for public health lacks credibility.

**Integrated Knowledge Management**

Khoury et al (2013) favor a knowledge integration approach to "drive research, policy and practice" (p. 14). According to the authors, "knowledge management is a continuous process of identifying, selecting, storing, curating, and tracking relevant information across disciplines" (p. 15). This therefore entails the selection and curation of articles, systematic reviews, and meta-analysis.

The fragmentation of public health datasets creates data silos which make it practically impossible to achieve a complete picture of either the individual or the population (Dickerson & Yao, 2014, p. 90). Often, data silos are the natural byproduct of discipline-focused research, with each discipline using vocabulary specific to it.

“For data acquired from disparate sources, harmonization of definitions can be a challenge” (Khoury et al, p. 14).

In addition to curating datasets, a holistic big data research system ought to include information artifacts as well. These include full-text copies of research articles and collateral (grey) literature of specific interest to the lines of inquiry being pursued within a given university program. Indeed, pure research needs to be contextualized, especially in clinically-focused fields which rely so heavily on *actionable* information. Busy clinicians - whether physicians, nurses, or epidemiologists - need actionable insights, based on valid research findings, to share with patients and clients. Ideally, this information would be context-specific. A nurse practitioner in rural Appalachia, for example, will view the opioid crisis from a lens that differs significantly from a public health professional in an urban setting like Los Angeles.

**Precision Public Health Defined**

The rise of precision medicine has created an environment conducive for the development of its public health counterpart. Khoury, Iademarco, & Riley (2016) compare the aims of the two by writing, "If precision medicine is about providing the right treatment to the right patient at the right time, precision public health can be simply viewed as providing the right intervention to the right population at the right time" (p. 398). With precision medicine, the primary unit of analysis is the *person* whereas the *population* takes center stage in precision public health.

Precision public health - if it is to succeed - will need to embrace an interdisciplinary approach. That is, data to support research questions in this emerging field will need to come from a variety of fields, including statistics, geography, economics, sociology, health, biology, genetics, and ecology, to name just a few.

**Notes**

Need to highlight the role of validation in the public health research process and the concomitant design implications for the underlying research system. This is a feedback loop of sorts. Published research that has been validated ought to inform future lines of inquiry. Hence, the need to store full-text articles in the research system alongside the raw data.

**References**

Dickerson, D and Yao, P. (2014). In Magnuson, J.A. and Fu, P.C. (eds) Public health informatics and information systems (2nd ed.). New York, NY: Springer.

Freeman, D. (2006). The scientist as rebel. New York, NY: New York Review of Books.

Hunt, M. (1997). How science takes stock: The story of meta-analysis. New York, NY: Russell Sage Foundation.

Khoury, M. J., et al (2013). Transforming Epidemiology for 21st Century Medicine and Public Health. *Cancer Epidemiology, Biomarkers & Prevention, 22*(4), 508–516. doi: 10.1158/1055-9965.EPI-13-0146

Khoury, M. J., Iademarco, M. F., and Riley, W. T. (2016). Precision public health for the era of precision medicine. *American Journal of Preventive Medicine, 50*(3), 398-401. doi: 10.1016/j.amepre.2015.08.31

Khoury, M.J. and Ioannidis, J.P.A. (2014). Big data meets public health. *Science, 346*(6213), 1054-1055.

Ola, O. and Sedig, K. (2014). The challenge of big data in public health: An opportunity for visual analytics. *Online Journal of Public Health Informatics 5*(3):e223.