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

**Precision Public Health**

In many ways, precision public health is but a natural extension of epidemiology combined with medical geography. 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). This quote also highlights a key difference between the two fields. With precision medicine, the focus is on the **individual**. While with precision public health, the focus shifts to the **population**. Additionally, precision public health concerns itself with prevention, intervention before the expression of disease. This is less the case with precision medicine where the primary concern is the identification of an appropriate treatment, after the fact.

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.

**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 more 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 nothing more than 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).

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.

**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.

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.

