How the National Institute of Health Can Lead
Evidence-based Medicine into the Age of Data Science?

by Ted Pham

Imagine a scenario where you had tried to exercise and followed a strict diet but your blood test still came back positive for high cholesterol. When your doctor prescribed CRESTOR to aid the condition, you asked your doctor for the evidence and he/she cited that CRESTOR had been shown to be effective in lowering cholesterol in randomized controlled trials. Several weeks past but your blood test showed no signs of improvement. The doctor reluctantly informed you that while he had followed the treatment guideline you might be a non-responder, analogous to the outliers in the controlled studies, to the drug. You were advised to try with a different drug and wait to see if this new drug would be effective.

The hypothetical scenario emphasizes a limitation of the current evidence-based approach to treatment of diseases. Clinical trials based on average results are insufficient to inform decisions about real patients who inevitably differ from the mean estimates in the trials [1,2]. Moreover, vested interests such as drug manufactures can influence research agenda and produce skewed conclusion without any significant statistical analysis[3,4]. Thus, the healthcare industry based on the established evidence-based practice provide patients and clinicians with "one size fits all" drugs that can help some but are prescribed to many. Figure 1 shows how imprecise the four highest-grossing drugs in the United States are in helping patients.

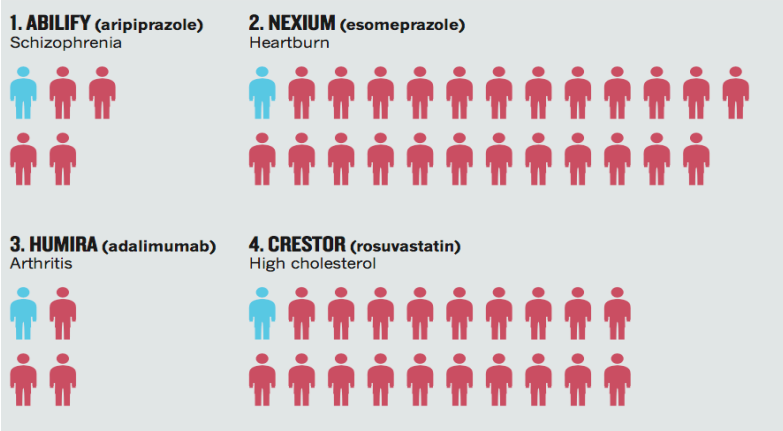

Figure 1: For every person the drug do help (blue), the number of people the drug fail to improve the conditions are shown in red Adapted from[5]

Does better evidence exist for clinicians and patients to make better, more informed clinical decisions? The answer is yes and the evidence can be much more precise but the **data** sources are too **heterogeneous**. The advent of high-throughput, high resolution technologies have generated complex and **evolving datasets on molecular omics** (e.g. genomics, proteomics, metabolomics), imaging and clinical data[6]. These data are stored in a complex web of public and private repositories which might not be accessible or even findable. In addition, patients' **records, written physician notes, consultant notes, radiology notes, discharge notes**, consultant notes etc. contain valuable information on patients' unique characteristics and unique responses to treatment[7]. However, these data are unstructured and might require

tremendous effort to digitize them. Electronic health records exist in a variety of standards, thus posing challenges for data integration and investigation. **Wearables** and even implanted continuous health sensors are also another data source at the individual levels. The concept of the "q**uantified self**"—behavioral, lifestyle, physiological and other clinically relevant data—would transform an individual into an active participant in their health management but only when the data is aggregated into a common system should there be actionable insights for individualized medicine[6]. Collectively, these data provide tremendous potential for individualized or precision medicine but at their present forms, they might be incomprehensible to clinicians and patients who might not have the expertise, time, or effort to decipher.

Considering the availability and existence of biomedical and health data and the heterogeneity associated with them, the need for a collaborative ecosystem for the data to live is apparent. The National Institute of Health (NIH) and its constituent institutes and centers are uniquely positioned to create such an ecosystem for biomedical data and guidelines on how to utilize such a system. As a public and independent institution, the NIH faces with a decision of how to build an effective ecosystem that will consolidate biomedical data and bring together all the stakeholders – physicians, scientists, pharmaceutical companies, medical devices companies, start-up data mining companies, non-profit organizations and the public— and capitalize on the transformative power of big data. In other words, *how can the NIH lead the effort to bring evidence-based medicine into the age of data science where the evidence is much more precise and at a much larger scale.*

To accomplish such an ambitious goal, the NIH directors first need to hire competent leaders[8] specializing in interdisciplinary research, computer sciences, clinical bioinformatics, and databases. Taking a pragmatic approach combining both qualitative and quantitative research methods will ensure the provision of data science tools and infrastructure to stakeholders on a continual learning basis. More specifically, using performance data of what tools and infrastructure are being used can help allocate resources for better development while hosting consortiums to seek feedbacks and inputs from the stakeholders to adequately improve and avoid common pitfalls of data science such as analysis without a question/plan or a development without any clients or obsolete. In addition of developing tools and computational infrastructure, the NIH can also focus on workforce training, fostering open science, and providing guidelines for data design. Such a guideline is **the FAIR principle** in which data is **Findable, Accessible, Interoperable, and Reusable.**

Building blocks of an NIH-sponsored ecosystem for the massive biomedical data utilization have been established such as the oncology genes repository by the National Cancer Institute and the genome sequence database by the National Human Genome Research Institute etc. The process of democratizing access to NIH-generated and related data would benefit from one-key authentication for cross-platform databases from different centers, institutes, and agencies. A promising is to develop a relational data infrastructure analogous to relational database so that one repository can cross talk to another. Getting other organizations e.g. hospitals, private practices, data mining companies, pharmaceuticals and individuals to

share will be a challenge but the applications of a more comprehensive data lake lead to better wellness and lower the cost of healthcare[9].

Building the ecosystem for massive biomedical data does not mean to forgo traditional evidence-based medicine where researches are funded to assess the risk and efficacy of new treatments through carefully designed randomized clinical trials. The NIH must be wary in allocating resources properly considering the appeal of novelty. In contrast, organizational bias towards clinical trials as the gold and only standard in biomedical research must be overcome through proper workforce continual education of analytics. The new paradigm of mining big data to better healthcare cannot and should not replace clinical trials but it can be complementary to the trials. The concept of **"N-of-one" trials**—studies focusing on a single person— is a consequence of the new paradigm[5].

Nevertheless, one crucial task that would lead to unexpected consequence the NIH must consider is to keep the public informed with the development of the data ecosystem and the research that uses their data to ensure transparency and confirm data security. If the NIH failed to keep the public abreast of the development, the idea of government misusing personal health data could be lit and spread contagiously, leading to voters to demand for defunding of the NIH, a situation should be avoided at all cost.

Another consequence of progressing evidence-based medicine into the age of data science is the transformation of medical education to incorporate statistics, combinatorics, machine learning, and computing. Medical professionals will be better equipped and trained to form expert judgements assessing the massively available data regarding patients' care[10].

Let's go back to our original hypothetical scenario, your blood cholesterol is still high but now with big data and the tools to deal with them, your doctor pulls data from the NIH-ecosystem and shows you infographics, visualizations, option grids and other decision aids to help you better understand your options. He/she could show you the statistics of treatments on individuals sharing similar characteristics with you or start a "N-of-one" trial for you. However, this time you can know if the treatment is working because of the data gathered from wearable and implanted devices. Such a scenario might seem futuristic but because of the collective talents available, the future might arrive sooner than we think.

References:

1. Collins, F. & Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **372,** 793–795 (2015).
2. Greenhalgh, T., Howick, J. & Maskrey, N. Evidence based medicine: a movement in crisis? *BMJ* **348,** g3725–g3725 (2014).
3. Ioannidis, J. P. A. Evidence-based medicine has been hijacked: A report to David Sackett. *J. Clin. Epidemiol.* **73,** 82–86 (2016).
4. DuBiner, H. *et al.* Efficacy and safety of bimatoprost in patients with Elevated intraocular pressure: A 30-day comparison with latanoprost. *Surv. Ophthalmol.* **45,** 353–360 (2001).

5. Schork, N. J. Personalized medicine: Time for one-person trials. *Nature* **520,** 609–611 (2015).
6. Beckmann, J. S. & Lew, D. Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities. *Genome Med.* **8,** 134 (2016).
7. Marr, B. How Big Data Is Transforming Medicine. (2016). Available at: https://www.forbes.com/sites/bernardmarr/2016/02/16/how-big-data-is-transforming-medicine/#a64c4927ddc6. (Accessed: 3rd February 2017)
8. NIH. Chief, Cancer Data Science Laboratory. (2017).
9. Schadt, E. & Chilukuri, S. The Role of Big Data in Medicine. Available at: http://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/the-role-of-big-data-in-medicine. (Accessed: 3rd February 2017)
10. Chen, R. C., Gabriel, P. E., Kavanagh, B. D. & McNutt, T. R. How Will Big Data Impact Clinical Decision Making and Precision Medicine in Radiation Therapy? *Int. J. Radiat. Oncol.* **95,** 880–884 (2016).