Big Data in Precision Medicine

Ryan L. Irey, M.A.
Indiana University
107 S Indiana Ave
Bloomington, Indiana 47405
rlirey@iu.edu

ABSTRACT

Precision medicine (PM) refers to the nationwide initiative to incorporate individual differences into disease treatment and prevention. The initiative, supported by the National Institutes of Health (NIH), leverages the sophistication of big data analytics and cloud computing infrastructures toward increasing the efficacy of individualized care.

KEYWORDS

precision medicine, health care, big data, hid318, i523

1 INTRODUCTION

Precision medicine is the term used to describe an individualized approach to the treatment and prevention of disease, leveraging genetic, geophysical, and patient lifestyle data toward the formulation of patient-optimized intervention plans. The concept of precision medicine is a diversion from the more classic treatment approach in which a patient is subject to a treatment plan that typically works for the average person. In 2015, President Barack Obama announced the Precision Medicine Initiative (PMI), which seeks to facilitate a successful precision medicine framework by investing in the technical infrastructure needed to collect, process, analyze, and apply patient-specific data. The PMI is also dedicated to establishing the massive cohort of participants necessary for implementing precision medicine.

Big data methodology has an established history in the field of medicine, with particular regard to electronic medical records and clinical decision support systems, among others. The convergence of big data on precision medicine follows an increasing trend of applying data science methods to better inform decision making in a variety of health care domains. In this article, current big data applications in medicine are reviewed, followed by a discussion of some challenges that precision medicine is likely to impose on current big data methodology.

2 OVERVIEW OF PRECISION MEDICINE AND THE PRECISION MEDICINE INITIATIVE

President Barack Obama announced the PMI "to bring us closer to curing diseases like cancer and diabetes, and to give all of us access to the personalized information we need to keep ourselves and our families healthier" [6][7]. By incorporating individual variability in genomic, environmental, and lifestyle factors, precision medicine aims to identify more effective disease treatments and establish practical disease prevention strategies. Some of the specific scientific opportunities identified in the initiative include:

- Methodology for measuring disease risk based on environmental and genetic factors, and their interaction
- Identify biological markers that signal increased or decreased risk of disease development
- Identify new disease classifications and relationships
- Creation of a data collection platform to research targeted therapies

The PMI is primarily led by the National Institutes of Health (NIH), but draws on the effort of other federal agencies, each having explicitly defined, data-centric roles toward implementing the PMI. For example, the Food and Drug Administration (FDA) is to "fiadvance the development of high quality, curated databases to support the regulatory needs to advance innovation in precision medicine" [4], while the Office of the National Coordinator (ONC) is to "...address privacy and enable secure exchange of data across systems" [4]. Of particular relevance to big data is the formation of the PMI Cohort Program (PMI-CP) - a voluntary research cohort with more than one million participants. The aim of the PMI-CP is not only to engage the public and recruit voluntary participants, but to collect biological samples making up a large biobank. The vision of the PMI-CP describes a "central biorepository, which will support collection, processing, storage, retrieval, analysis, and shipment processes" [4].

3 BIG DATA APPLICATIONS

Applications of big data in healthcare, or in any domain, are generally characterized by the "volume, velocity, variety, veracity, and value" of a dataset[3]. The field of healthcare has already capitalized on several such opportunities in the form of electronic patient health records (EHRs), clinical decision support systems (CDSSs), and social media mining. Insofar as precision medicine is concerned, He, Ge, & He (2017) had the following to say:

*"The **volume** comes from large amounts of records that can be derived from the EHRs for patients... The **velocity** occurs when data is accumulated at high speeds, which can be seen when monitoring a patientfis real-time conditions through medical sensors... The **variety** refers to data sets with a large amount of varying types of independent attributes, such as data sets that are gathered from many different*

resources. ***Veracity*** *is a concern when working with possibly noisy, incomplete, or erroneous data where such data need to be properly evaluated using other relevant true evidence.* ***Value*** *portrays the usefulness for improving healthcare outcome."*[2]

3.1 Research and Development

The research and development underlying the precision medicine framework is vast, as researchers are trying to navigate a variety of complexities such as natural language processing approaches for unstructured data, computational efficiency surrounding the collection and access of EHR data, and optimization of clinical decision support systems (CDSSs). Moreover, the scalability, accessibility, and sustainability inherent to the ideal data infrastructure is currently being articulated for future deployment[2].

3.2 Prediction and Prevention

Disease management, prediction, and prevention are some of the most lucrative area in which precision medicine and its big data applications will thrive. In the area of epidemiology, big data approaches to epidemic solutions have been useful in answering questions related to asthma and environmental conditions, as well as more serious outbreaks such as Ebola; such analysis approaches have integrated social media data to track outbreak patterns and deploy intervention campaigns[5].

Big data analysis approaches have also been used to answer questions related to emergency vehicle demand and weather. Particularly in Hong Kong, it is hypothesized that women, low-income groups, and the elderly are most susceptible to the effects of extreme weather, and with an aging population, long-term predictions of ambulance demand will fall short of actual needs[8].

To compliment these big data applications to prediction and prevention in healthcare, the intersections of big data analytic methods and the scope of precision medicine have yielded some interesting additional applications. These include predictive models for hospital readmission, incorporating EHR's and genomic data toward earlier diagnoses and reduced treatment complications, and the aggregation of comprehensive patient health profiles[5]. Finally, the surge of wearable sensor-equipped fitness devices is generating an immense volume of data for further incorporation into predictive modeling in precision medicine[1].

4 ANTICIPATED CHALLENGES

The general novelty of precision medicine has given rise to a number of realized and anticipated challenges[5][2], such as interoperability between genomic data and electronic medical records, data sharing between institutions, and soliciting participation from patients[1]. Perhaps even more dire than the aforementioned are issues surrounding the treatment of unstructured data and metadata, data acquisition and storage, training, and data privacy[1][2].

4.1 Unstructured Data and Metadata

For precision medicine to work as intended, healthcare professionals need the complete picture. While structured data in healthcare is plentiful, it is the unstructured data and metadata that are currently without tools for mining their information. In a similar vein, the prospect of sharing data among precision medicine organizations means that the value of metadata, and the means to mine it, will certainly come into focus. Querying metadata would allow for a more clear understanding of the setting/context under which certain data were collected, potentially contributing to an increase in statistical power within research efforts[1].

4.2 Data Acquisition and Storage

Data acquisition requirements are, at present, lacking in an explicit definition[1]. In addition to acquisition, the question of how best to store and access large genomic datasets has yet to be addressed. Since many precision medicine organizations are currently storing data in-house, there exists some potential for a uniform system to be adopted, thus making the potential for data sharing much more viable (save for the privacy concerns). With this in mind, the establishment of uniform data definitions ought to be undertaken also.

4.3 Training

Issues surrounding training are essentially two-fold: (1) since the data are essentially gathered by medical professionals, the training of such employees on how to record, store, and access genomic data as it relates to precision medicine is no small undertaking. The success of precision medicine as a philosophy depends on a workforce that can access the information obtained in real-time, and contribute new information just as easily. Moreover, the analytic side of big data in precision medicine pines for data scientists with strong backgrounds in medical fields (or conversely, medical professionals with strong data science and statistics skillsets)[2].

4.4 Data Privacy

The notion of data privacy as it relates to precision medicine also encapsulates issues pertaining to security of data collected and the risk inherited by organizations maintaining such data. If warehousing the data locally, as many organizations involved in precision medicine are doing, it is tempting to ignore these issues and simply apply the same data security and privacy tactics the organization has implemented for itself[1][2]. However, as cloud-based solutions with remote databases are explored as an option for the ever-expanding volumes of genomic data, extra precautions and specialized security measures must be taken to not only protect the privacy of patient information, but to also maintain its integrity with utmost confidence.

5 CONCLUSION

Precision medicine seeks to define a new era of healthcare such that the diversity in genomic makeup, geophysical surroundings, and personal lifestyle is taken into account. Big Datafis influence has already been observed in some domains of healthcare and related fields, such as the systems designed to manage electronic health records, clinical decision systems, and genomic sequencing. Due to their inherent entanglement, the proliferation of precision medicine depends, in some sense, on the maturation of healthcare policy, big data analytics, and the expansion of computational infrastructures supporting the collection, processing, analysis, and long-term management of patient data. The extraordinary effort required to recruit a massive cohort of participants notwithstanding, Big Data is certainly poised to propel precision medicine into the mainstream of health care practice.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and all teaching assistants of INFO-I523 for their thoughtful comments and guidance on previous versions of this manuscript.

REFERENCES

[1] Foundation for eHealth Initiative. 2017. Data Management for Precision Medicine and Genomics in 2017. *White paper* (2017), 25. https://www.ehidc.org/sites/default/files/resources/files/Data_Management_for_Precision_Medicine_and_Genomics_in_2017_3.15.17_0.pdf

[2] Karen He, Dongliang Ge, and Max He. 2017. Big Data Analytics for Genomic Medicine. *International Journal of Molecular Sciences* 18, 2 (feb 2017), 412. https://doi.org/10.3390/ijms18020412

[3] H. V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, and Cyrus Shahabi. 2014. Big data and its technical challenges. *Commun. ACM* 57, 7 (jul 2014), 86–94. https://doi.org/10.1145/2611567

[4] Bray Patrick-Lake Kathy Hudson, Rick Lifton. 2015. *The Precision Medicine Initiative Cohort Program - Building a Research Foundation for 21st Century Medicine.* resreport. National Institutes of Health. https://www.nih.gov/sites/default/files/research-training/initiatives/pmi/pmi-working-group-report-20150917-2.pdf

[5] Daniel Richard Leff and Guang-Zhong Yang. 2015. Big Data for Precision Medicine. *Engineering* 1, 3 (Sept. 2015), 277–279. https://doi.org/10.15302/j-eng-2015075

[6] Barack H. Obama. 2015. President Obama Speaks on the Precision Medicine Initiative. YouTube. (Jan. 2015). https://www.youtube.com/watch?v=MKiw7yAqqsU

[7] Sharon F. Terry. 2015. Obama's Precision Medicine Initiative. *Genetic Testing and Molecular Biomarkers* 19, 3 (mar 2015), 113–114. https://doi.org/10.1089/gtmb.2015.1563

[8] Ho Ting Wong, Qian Yin, Ying Qi Guo, Kristen Murray, Dong Hau Zhou, and Diana Slade. 2015. Big data as a new approach in emergency medicine research. *Journal of Acute Disease* 4, 3 (aug 2015), 178–179. https://doi.org/10.1016/j.joad.2015.04.003