Achieving Value in Population Health Big Data

Daniel D. Bu[1], Shelley H. Liu[1], Bian Liu[1], and Yan Li[1,2]

[1]Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [2]Department of Obstetrics, Gynecology, and Reproductive Science, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

Several population health big data projects have been initiated in the USA recently. These include the County Health Rankings & Roadmaps (CHR) initiated in 2010, the 500 Cities Project initiated in 2016, and the City Health Dashboard project initiated in 2017. Such projects provide data on a range of factors that determine health—such as socioeconomic factors, behavioral factors, health care access, and environmental factors—either at the county or city level. They provided state-of-the-art data visualization and interaction tools so that clinicians, public health practitioners, and policymakers can easily understand population health data at the local level. However, these recent initiatives were all built from data collected using long-standing and extant public health surveillance systems from organizations such as the Centers for Disease Control and Prevention and the U.S. Census Bureau. This resulted in a large extent of similarity among different datasets and a potential waste of resources. This perspective article aims to elaborate on the diminishing returns of creating more population health datasets and propose potential ways to integrate with clinical care and research, driving insights bidirectionally, and utilizing advanced analytical tools to improve value in population health big data.

KEY WORDS: population health; big data; social determinants of health; integration of clinical and population health data.

J Gen Intern Med 35(11):3342–5
DOI: 10.1007/s11606-020-05869-0

Health is being increasingly recognized as a function of social determinants and environmental factors[1]—common metrics that are aggregated and tracked through population health big data. Similarly, health care is integral to promoting and maintaining health and wellness. We propose that the vast troves of available public health big data can be better positioned, analyzed, and integrated with clinical data to improve health, both at the population and the individual patient levels. In recent years, various large, well-funded data projects have been initiated in the USA. These include the County Health Rankings & Roadmaps (CHR), which ranked counties based on various health factors in 2010, the 500 Cities Project, which examined the burden of unhealthy behaviors in 500 USA cities in 2016, and the City Health Dashboard, another visualization of behavioral and health outcomes data for different US cities in 2017. These recent initiatives were built from data collected using long-standing and extant public health surveillance systems from the Centers for Disease Control and Prevention (CDC) and U.S. Census. However, with the advent of major advances in analytics, computational power, and insight-generating methods, it is also important to evaluate how to best attain value with these large data sets.

Are We Getting Diminished Returns?. Considerable resources have been expended in building data visualizations like interactive maps and zoomable regional statistics. The putative goal of these projects is to communicate complicated and often messy data into clear signals, thereby offering researchers, policymakers, community workers, and physicians a better understanding of complex problems and health resource allocation. However, these visualization tools are built using the same base data. Thus, the redundancy between projects may be contributing to diminishing returns. Often the biggest difference lies in the geographic area of analysis: city, country, or state.[2] Relying on the same or similar data sources, the result is a visually different, but substantively similar representation, of the same data. As an illustration, for cities ranging from Mobile, AL, to Jonesboro, AR, to Philadelphia, PA, the City Health Dashboard, CHR, and 500 Cities initiatives all visualize prevalence of cardiovascular disease, medication adherence, and binge drinking levels among many other factors identical across the three databases. Yet the expenditure of human capital and resources remains high. For example, data sets need to be manually integrated across thousands of geographic tracts. Such endeavors require many person-hours for potentially diminishing returns. It is important to recognize these efforts have opportunity costs.

How Do We Take Advantage of the Currently Available Big Data?. In the context of limited public health resources, we suggest a shift away from merely data aggregation and visualization and toward data hypothesis and insight. Such initiatives can help design, implement, and evaluate effective practices and policies at the local, state, and federal levels to improve outcomes. For example, we should mine data to derive insights to guide interventions by clinicians on the

Received December 11, 2019
Revised March 13, 2020
Accepted April 22, 2020
Published online May 11, 2020

frontlines—e.g., focusing clinical care to specifically underserved patients, or partnering with local-level welfare workers and other professionals to change some of the underlying determinants of health.[3–5] To date, efforts on knowledge discovery from data mining, including the use of dynamic activity paths, have yielded novel findings but have been limited in scope and applicability to the clinic.[6] New data science tools can reveal powerful and hidden associations between social determinants across large datasets. For example, machine learning is one method that enables researchers to harness data repositories, such as those available from Behavioral Risk Factor Surveillance System (BRFSS) and American Community Survey (ACS), to forecast future trends.[7] Using machine learning techniques, researchers have found relationships between the built environment and obesity, and reordered predictors of cardiovascular disease across US cities.[8, 9] With convolutional neural networks, a type of deep learning, Maharana et al. showed that information of the built environment extracted from high-resolution satellite images could predict adult obesity prevalence in multiple cities. Although these investigators used the same base data from the aforementioned CDC datasets, they focus on providing new insights instead of generating additional visualizations of existing data. We should increase similar efforts to capitalize on the availability of population health big data.

Big data, in conjunction with both novel and traditional techniques, can also be used to reexamine and reweight existing associations—e.g., cardiovascular (CVD) risk factors and outcomes).[9–11] Using machine learning, recent analyses combine neighborhood-level factors like binge drinking and obesity, with individual-level clinical indicators of blood pressure and medication adherence to add nuance to patient risk profiles for CVD and heart failure.[9, 10] The rise of population-level big data also allows for new uses of traditional statistical analysis tools. Even as machine learning allows us to drive insight with novel risk scores for prediction, traditional methods using linear and logistic regression models allow for the use of more hypothesis-driven analysis.[7, 12] This in turn may be advantageous in evaluating specific effects of individual predictors. Taken together, these different tools for data analysis demonstrate multiple paths exist to achieving further value with big data in population health.

Nevertheless, the integration of population health big data in clinical practice is still in its infancy. Current clinical practice could help bridge existing individual-level clinical data with community data. For instance, the American Medical Association and United Healthcare have recently proposed to develop a set of ICD-10 codes to capture patients at risk of non-medical issues such as food and housing. The development and utilization of such codes to understand social determinants of health could be used as connectors between the clinical data available from individual patient visits and the macro-level trends in public health big data. Further adoption of this approach will also allow physicians to focus on patient-centered care, including the needs of the whole person, and not medical care alone. Clinicians can better track the social needs of patients, thereby delivering more personalized care. Public health practitioners can better aggregate data and build a strategy based on social determinants. Additionally, such coding efforts may ultimately afford clinicians the chance to build data from CHR, 500 Cities, and City Dashboard into their practice. With these linkages, the patient-physician relationship can be strengthened and informed by important, patient-centered, contextual clues.

How Do We Maximize the Value of Big Data?. In order to further optimize the value of public health data, we should look beyond overlaying public health survey and census data with each other and move toward vertical integration of area-level data with clinical and individual-level data. Public health research could identify sociodemographic risk factors and combine them with data extracted from patient-level clinical courses to form richer data streams. At present, health measures have already been collected at each census level. For example, we could improve policy by integrating data vertically, combining electronic health record data with existing public health data into a data repository. Functionally, this allows us to fuse behavioral factors, prevention measures, health care, and environmental factors. In contrast to large cardiometabolic and cancer surveillance data, which often already span multiple health systems, few initiatives attempt to combine broader clinical data across separate health systems.[13–15] One such venture, the INSIGHT clinical research network (formerly NYC-clinical data research network) funded by the Patient Centered Outcomes Research Institute (PCORI), has integrated not only 2.5 million longitudinal individual-level clinical and EHR data across health systems, but also building in almost 200,000 different signals pertaining to social determinants and public health signals.[16] This has allowed for patient-centered population health big data research, enabling clinicians and researchers alike to parse the effect of public health interventions on patient-level factors for diabetes care disparities, rare diseases, and cardiovascular comparative effectiveness studies.[13, 16] Future initiatives will need to look at health care data as an integrated whole.

The momentum behind population health big data also requires careful attention toward potential pitfalls. Models fitted to big data carry some inherent assumptions. Even as data repositories include more granularity and cover more distinct unit areas like census tracts, methods, like small area estimation, used to generate this level of detail have inherent limitations. Subjecting these new streams of data to trend interpolations can lead to logically incorrect conclusions arrived at by circular logic. For instance, in some of the smaller census tracts, where data (e.g., cardiovascular disease prevalence) is readily available, it is imputed based on linear

regression models. If we use machine learning algorithms to then extract a relationship, we may be simply using circular logic to derive what we assume are new insights. Such concerns can be mitigated with close inspection of data sources. Combining data streams will also be challenging, with thoughtful consideration required of how best to link public health data with individualized data so that the result is functional, accurate, and easy to use while bearing the responsibility for data privacy and patient autonomy in mind.[17–19] Additionally, data science approaches such as machine learning are certainly not a panacea. Although these models are built based on new modes of analysis, they are strongly limited by the quality of the data inputs, and the degree to which these inputs accurately encompass the breadth and the depth of the problem.[17, 20] Finally, there is the all-important issue of data privacy. Interlinked systems of big data may increase the chances of triangulating the identities of patients from deidentified datasets. To deal with such concerns, we need strengthened guardrails that emphasize patient consent and data security. Existing regulations, such as HIPAA, predate not only the use of electronic health records, but also widespread penetrance of the internet. New measures are needed, and likely require a combination of newer regulation, data use committees, and more transparent authorization disclosure forms.[21, 22]

This pursuit of value with population health big data will ultimately come to benefit major stakeholders. For clinicians, knowing the community and environment in which the patient lives will contextualize the clinical relevance of how manage patients. For public health stakeholders, big data platforms will serve as a launching pad for new lines of investigation, allowing for analysis of patient-level factors in the aggregate. For policymakers, data insights can facilitate comprehensive evaluations of the effects of policy from the census tract to the patient bedside.

Operationally, better directed population health big data will (1) drive insights that allow physicians to better track the social needs most pressing for their patients, (2) aggregate patient-level data to form a coherent strategy formed by social determinants, and (3) help emphasize areas where productive partnerships can be built within the broader health care community. To achieve these goals, we should encourage coordination between funders and researchers, allowing them to move from data aggregation to data integration of clinical and public health measures. Finally, as these efforts progress, it will be crucial not only to measure the impact on, but also to seek the opinions of, patients themselves. Preliminary evidence indicates that patients recognize both the urgent need for and the importance of such data.[13] Similarly, it is important for researchers to recognize both the responsibility and the insight that can be gleaned from big data, along with the necessity of maintaining data privacy. Only then, as a community, can researchers utilize this patient-centered approach to achieve the most value from population health big data.

Corresponding Author: *Yan Li, Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, USA (e-mail: yan.li1@mountsinai.org).*

Funding Information *This study is supported, in part, by a grant from the National Heart, Lung, And Blood Institute of the National Institutes of Health under Award Number R01HL141427.*

Compliance with Ethical Standards:

The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Conflict of Interest: *The authors declare that they have no conflict of interest.*

REFERENCES

1. **Wilkinson RG**, **Marmot M**. Social determinants of health: the solid facts. World Health Organization; 2003.
2. **Gourevitch MN**, **Athens JK**, **Levine SE**, **Kleiman N**, **Thorpe LE**. City-Level Measures of Health, Health Determinants, and Equity to Foster Population Health Improvement: The City Health Dashboard. Am J Public Health. 2019;109(4):585-592.
3. **Cheng JK**. Confronting the Social Determinants of Health — Obesity, Neglect, and Inequity. N Engl J Med. 2012;367(21):1976-1977.
4. **Ludwig J**, **Sanbonmatsu L**, **Gennetian L**, **et al.** Neighborhoods, Obesity, and Diabetes — A Randomized Social Experiment. N Engl J Med. 2011;365(16):1509-1519.
5. **Tsai J**, **Gelberg L**, **Rosenheck RA**. Changes in Physical Health After Supported Housing: Results from the Collaborative Initiative to End Chronic Homelessness. J Gen Intern Med. 2019;34(9):1703-1708.
6. **Wiebe DJ**, **Richmond TS**, **Guo W**, **et al.** Mapping Activity Patterns to Quantify Risk of Violent Assault in Urban Environments. Epidemiology. 2016;27(1):32-41.
7. **Obermeyer Z**, **Emanuel EJ**. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375(13):1216-1219.
8. **Maharana A**, **Nsoesie EO**. Use of Deep Learning to Examine the Association of the Built Environment With Prevalence of Neighborhood Adult Obesity Prevalence. JAMA Network Open. 2018;1(4):e181535.
9. **Li Y**, **Liu SH**, **Niu L**, **Liu B**. Unhealthy Behaviors, Prevention Measures, and Neighborhood Cardiovascular Health: A Machine Learning Approach. J Public Health Manag Pract. 2019;25(1):E25-e28.
10. **Angraal S**, **Mortazavi BJ**, **Gupta A**, **et al.** Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction. J Am Coll Cardiol. 2020;8(1):12-21.
11. **Fox ER**, **Samdarshi TE**, **Musani SK**, **et al.** Development and Validation of Risk Prediction Models for Cardiovascular Events in Black Adults: The Jackson Heart Study Cohort. JAMA Cardiol. 2016;1(1):15-25.
12. **van den Heuvel ER**, **Vasan RS**. Statistics in cardiovascular medicine: there is still gold in the old. Heart. 2018;104(14):1227-1227.
13. **Goytia CN**, **Kastenbaum I**, **Shelley D**, **Horowitz CR**, **Kaushal R**. A Tale of 2 Constituencies: Exploring Patient and Clinician Perspectives in the Age of Big Data. Med Care. 2018;56(Suppl 10 Suppl 1):S64-s69.
14. Investigators A. The atherosclerosis risk in communities (ARIC) study: Design and objectwes. Am J Epidemiol. 1989;129(4):687-702.
15. **Warren JL**, **Klabunde CN**, **Schrag D**, **Bach PB**, **Riley GF**. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Med Care. 2002;40(8 Suppl):Iv-3-18.
16. **Kaushal R**, **Hripcsak G**, **Ascheim DD**, **et al.** Changing the research landscape: the New York City Clinical Data Research Network. J Am Med Inform Assoc. 2014;21(4):587-590.
17. **Char DS**, **Shah NH**, **Magnus D**. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. N Engl J Med. 2018;378(11):981-983.
18. **Rajkomar A**, **Dean J**, **Kohane I**. Machine Learning in Medicine. N Engl J Med. 2019;380(14):1347-1358.
19. **Rosati RA**, **McNeer JF**, **Starmer CF**, **Mittler BS**, **Morris JJ Jr.**, **Wallace AG**. A New Information System for Medical Practice. Arch Intern Med. 1975;135(8):1017-1024.

20. **Beaulieu-Jones B**, **Finlayson SG**, **Chivers C**, **et al.** Trends and Focus of Machine Learning Applications for Health Research. JAMA Network Open. 2019;2(10):e1914051.
21. **Vayena E**, **Blasimme A**. Health research with big data: Time for systemic oversight. J Law Med Ethics. 2018;46(1):119-129.
22. **Parasidis E**, **Pike E**, **McGraw D**. A Belmont report for health data. N Engl J Med. 2019;380(16):1493.