Big data and population health

Hu, Howard; Galea, Sandro; Rosella, Laura; Henry, David

Published in:
Epidemiology

DOI:
10.1097/EDE.0000000000000711

Published: 04/07/2017

Document Version:
Peer reviewed version

Epidemiology Publish Ahead of Print

DOI: 10.1097/EDE.0000000000000711

Big Data and Population Health: Focusing on the Health Impacts of the Social, Physical, and Economic Environment

(Short title: "Big Data and Population Health")

Accepted (March, 2017) for publication in EPIDEMIOLOGY

Howard Hu* (1), Sandro Galea (2), Laura Rosella (3) (5), David Henry (3)(4)(5)

* Corresponding author

(1) Office of the Dean, Dalla Lana School of Public Health, University of Toronto; 155 College Street, Toronto, ON, M5T 3M7, Canada; (416) 978-1841; howard.hu@utoronto.ca

(2) Office of the Dean, Boston University School of Public Health, Boston, MA, USA; sgalea@bu.edu

(3) Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada; laura.rosella@utoronto.ca

(4) Institute for Health Policy Management & Evaluation and the Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada; david.henry@utoronto.ca

(5) Institute for Clinical Evaluative Sciences, Toronto, ON, Canada

Acknowledgments : A number of the ideas in this manuscript were originally conceived and discussed by HH in the John R. Goldsmith Lecture at the 2015 Annual Meeting of the International Society for Environmental Epidemiology (Sao Paolo, Brazil, August 2015). None of the authors are aware of any conflicts of interests with respect to this manuscript. This commentary was sponsored by the International Society of Environmental Epidemiology (ISEE). However, the contents are the sole responsibility of the author(s) and do not necessarily reflect the official views of the ISEE.

Introduction

The growing abundance of data on the factors that produce health, and the capacity to link these to data from individuals, hold extraordinary promise for improving population health. Although public health is built on a long history of creatively using health data going back to the pioneering work of John Snow and William Farr[1, 2], the full potential of contemporary rich data sources is not being fully realised. Instead, the impact of this revolution is being seen mainly in the burgeoning precision medicine agenda, globally, and through the Precision Medicine Initiative (PMI) in the U.S.A.[3] In that context, linkage of omics data to phenotypic information from health records is being exploited as new research platforms, with the hope that this will transform clinical practice. Similar momentum has not yet been achieved for the population health sciences.

We propose that there are two challenges for the population health agenda. First, ensuring that we make better use of existing data, and in some cases, enhance data linkage and methods to do this. Second, in extending our efforts from the individual to the population by exploiting new complex, and sometimes unstructured, data sources.

We recognize that there are major technological challenges when dealing with the massive and rapid flows of data that come from both traditional data sources, such as large administrative databases, and new sources, such as genomics, land-use, neighborhood and climate data, and unstructured social media feeds. However, we argue that the more important challenge is conceptual and perhaps ideological. How do we identify the data that matter most to improve health for the whole population? In some cases, these are existing data and we need to determine how to enrich, link, and analyze these routinely collected data to maximize their value. We also need to think about new sources of big data that can improve our understanding of health, and how we better integrate these in our approaches to studying population health while avoiding the temptation to analyze everything that may

present itself as a new opportunity. Distinct from personalized medicine, the 'what matters most' question can be addressed through an understanding of the pressing health challenges of our time and an understanding of how factors across economic, social, behavioral, and biological domains may interact.

The past two decades have shown how factors at multiple 'levels of influence' are associated with both individual and population health.[6] Multilevel causal frameworks suggest that policies, features of institutions, characteristics of neighborhoods and communities, living conditions, and social relationships all contribute, together with individual behaviors and individual factors such as genotype poverty and race, to the production of health.[7] For example, quality of the built environment has been shown to be associated with mental ill health and diabetes.[8, 9] Certain social network characteristics are associated with the risk of obesity.[10] It is a logical extension of the goals of big data collection to introduce measures that can capture potential risks at multiple levels of influence. The feature that turns these into "big data" is the coverage and complexity of this information, which enables the population health approach, across geographies, time, and the life course.

Examples of the value of population data linkage and the potential contributions of big data

Two recent papers illustrate the importance of linking health, ethnicity, and personal financial data. Chetty and colleagues linked personal taxation and social security data of the US population to study the independent effects of income and place on life expectancy.[11] This entailed analyses of 1.4 billion person–years of observations ("big data").. Being poor in Detroit is more hazardous to health than being similarly disadvantaged in New York or San Francisco. Case and Deaton documented alarming rises in death rates in middle-aged non-Hispanic white males and females after the global financial crisis.[12] Both studies show that

physical and social factors interact strongly to mitigate or enhance the effects of poverty. Even partial mitigation of these effects could be life-saving on a scale that would compare well with any impact of precision medicine interventions.

Consider how the value of such studies would be enhanced by more complex data that enable longitudinal analyses to determine the effects of migration (e.g moving from Detroit to New York), the role of behavioral factors, genomics and epigenomic modifications, and access to healthcare, etc. This would enable the study of gene/environment interactions that determine health, the biologic changes induced by stresses associated with poverty and marginalization, the role of health systems, and mediating factors that could potentially afford new opportunities for disease prevention. There are a growing number of centers where such linkages and studies will be possible. One notable example is the Big Data Institute at the University of Oxford, where data from the UK biobank can be linked to administrative data and electronic health records (www.bdi.ox.ac.uk).

These types of data and research can also inform health services planning and delivery. Just as genomic and phenotypic data can predict an individual's response to treatment, linked socioeconomic, environmental, and health services data can be used to predict population risk for certain diseases. Examples include a population diabetes risk prediction tool for policy makers, and evidence that the development of chronic disease in low socio-economic circumstances identifies individuals who will become very high cost users of the healthcare system.[14,15]

Linkage with data on education, crime, and occupations can also help elucidate the wider impacts of health status. Such linkage and longitudinal follow-up enabled researchers in Manitoba to document, among teen mothers, the full negative impact of their health status on the educational attainment of their off spring. [16] In British Columbia researchers used routinely collected data from the Early Development Instrument (EDI) to measure the

negative impact of early childhood vulnerability (a composite of physical health, social competence, emotional maturity, language and cognitive development, communication skills, and general knowledge in the majority language and culture) on school achievement, standardized test scores, and criminality.[17]

With regards to environmental health, the availability of repeated, increasingly detailed, and geospatially mapped measures of air pollution has allowed studies linking living near major roads and/or exposure to fine particulate matter and other air toxics with cardiovascular disease, diabetes, autism and, very recently, neurodegenerative diseases[18, 19] Linkage with social data, which currently occurs rarely, would help researchers determine, for example, the environmental contribution to the health risks of living in high-poverty areas.

Prioritizing data from marginalized populations

Finally, we highlight the current disconnect between big data for health agendas and the health of marginalized populations. Undocumented immigrants, migrant workers, the homeless, and indigenous populations have some of the worst health outcomes in societies.[20] This is in contrast to documented immigrants who typically have better health than resident populations.[21] Marginalized groups should benefit from 'big data' efforts that identify new targets for prevention and patterns of health care utilization and inform studies of the epigenomic and other biologic impacts of poverty and social exclusion. These populations are rarely included in large cohort studies; however, advances in electronic medical records, data linkage and data security, and socialization, make it possible to track and document the health experience of individuals who may be migratory or homeless while protecting their privacy. As an example with respect to indigenous populations, the administratively linked data on over 13.5 million Ontarians held at the Institute for Clinical Evaluative Sciences which is enabling an accelerated "big data for population health" initiative now includes data

on over 200,000 First Nations individuals. This was made possible by especially close collaboration with and the approval of indigenous communities and their associated governance bodies[22], a strategy and principle that must extend to research on other disadvantaged groups.

Recommendations

At a time when there is so much interest and activity in big data analytics and precision medicine, allowing such efforts to diverge from the interests of population health is unwise. Rather, we have an opportunity to use the momentum that has been created by this movement to do two things: first, to enhance our uses of large population data-sets to gain a better understanding of the determinants of health and health inequities to support policies that will optimize health, and second, to push for more complex data that more closely reflects a person's context from the social, demographic, economic, and biologic perspectives over time. In so doing we need to be strategic about how we prioritize data efforts and achieve consensus on where to start to build the momentum to a level that is at least as high as the precision medicine realm. Our challenge is not so much employing all types of data and methods, as avoiding the temptation to do everything.

We identify 5 priorities:

1) Data on socio-economic status must be improved and linked to health services and health outcomes data. It is possible to securely link income tax data if revenue agencies can be assured that client privacy will not be breached. In keeping with the big data agenda, as much as possible these data should be longitudinal, reflect the dynamic and contextual nature of socio-economic status and environmental factors.
2) We must strive to include data that identify the most vulnerable members of the population, including indigenous peoples, migrants, refugees, and the homeless. Governance of the use of such data must involve those affected.

3) We should accelerate the shift from ecological studies by securing linkage of data at the level of the individual, which enables analyses at different levels, including micro- and macro-environments, such as occupational and community-level exposures to pollution; neighborhood walkability, "food desert" measures, occupational data, and social isolation among others.
4) Collaborations are needed with: (a) researchers in the fields of education, environment, and social sciences to ensure the validity and accuracy of multi-level data and (b) decision-makers in the public health and healthcare sectors to ensure that research questions and findings are those likely to have maximal impact on health.
5) All reasonable efforts must be made to protect privacy. Analyses should use de-identified data in facilities that provide the necessary combination of policies, staff training, physical, and electronic security.

It is only by addressing these priorities that we can assure that "Big Data" initiatives that pertain to health are not limited to the relatively narrow allopathic goals of the precision medicine agenda. By pursuing, instead, a more expansive "Big Data for Population Health" vision, initiatives so aligned carry the promise of generating insights on the drivers of health that can lead to interventions and policies that promote health on a population scale: the very core mission of public health.

Reference List

1. Langmuir AD. Farr,W - Founder of Modern Concepts of Surveillance. International Journal of Epidemiology 1976;5(1):13-18.
2. Snow J. Cholera and the water supply in the south districts of London in 1854. Journal of Public Health and Sanitary Review 1856;2:239-257.
3. Collins FS, Varmus H. A new initiative on precision medicine. The New England journal of medicine 2015;372(9):793-795.
4. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. JAMA 2014;311(24):2479-2480.
5. Rubin R. Precision medicine: The future or simply politics? JAMA 2015;313(11):1089-1091.
6. Roos LL, Magoon J, Gupta S, Chateau D, Veugelers PJ. Socioeconomic determinants of mortality in two Canadian provinces: Multilevel modelling and neighborhood context. Social Science & Medicine 2004;59(7):1435-1447.
7. Siddiqi A, Nguyen QC. A cross-national comparative perspective on racial inequities in health: the USA versus Canada. Journal of Epidemiology and Community Health 2010;64(01):29-35.
8. Creatore MI, Glazier RH, Moineddin R. Association of neighborhood walkability with change in overweight, obesity, and diabetes. JAMA 2016;315(20):2211-2220.
9. Latkin CA, Curry AD. Stressful neighborhoods and depression: A prospective study of the impact of neighborhood disorder. Journal of Health and Social Behavior 2003;44(1):34-44.
10. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. New England Journal of Medicine 2007;357(4):370-379.

11. Chetty R, Stepner M, Abraham S. The association between income and life expectancy in the united states, 2001-2014. JAMA 2016;315(16):1750-1766.
12. Case A, Deaton A. Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century. Proceedings of the National Academy of Sciences 2015;112(49):15078-15083.
13. Mackenbach JP, Kulhanova I, Artnik B et al. Changes in mortality inequalities over two decades: register based study of European countries. Bmj-British Medical Journal 2016;353.
14. Rosella LC, Manuel DG, Burchill C, Stukel TA. A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). Journal of Epidemiology and Community Health 2011;65(7):613-620.
15. Fitzpatrick T, Rosella LC, Calzavara A et al. Looking Beyond Income and Education Socioeconomic Status Gradients Among Future High-Cost Users of Health Care. American Journal of Preventive Medicine 2015;49(2):161-171.
16. Jutte DP, Roos NP, Brownell MD, Briggs G, MacWilliam L, Roos LL. The Ripples of Adolescent Motherhood: Social, Educational, and Medical Outcomes for Children of Teen and Prior Teen Mothers. Academic Pediatrics 2010;10(5):293-301
17. Kershaw P, Warburton B, Anderson L, Hertzman C, Irwin LG, Forer B. The Economic Costs of Early Vulnerability in Canada. Canadian Journal of Public Health-Revue Canadienne de Sante Publique 2010;101:S8-S12.
18. Feng S, Gao D, Liao F, Zhou F, Wang X. The health effects of ambient PM2.5 and potential mechanisms. Ecotoxicology and Environmental Safety 2016;128:67-74.
19. Chen H, Kwong JC, Copes R, Tu K, Villeneuve PJ, van Donkelaar A, Hystad P, Martin RV, Murray BJ, Jessiman B, Wilton AS, Kopp A, Burnett RT. Living near

major roads and the incidence of dementia, Parkinson's disease, and multiple sclerosis: a population-based cohort study. Lancet. 2017 Jan 4. pii: S0140-6736(16)32399-6. doi: 10.1016/S0140-6736(16)32399-6. [Epub ahead of print] PubMed PMID: 28063597.

20. Hwang SW. Homelessness and health. Canadian Medical Association Journal 2001;164(2):229-233.
21. Kennedy S, Kidd MP, McDonald JT, Biddle N. The Healthy Immigrant Effect: Patterns and Evidence from Four Countries. Journal of International Migration and Integration 2015;16(2):317-332.
22. Walker J, Jones C, First Nations Data Sovereignty in action: Creation of the largest First Nations health research study cohort in Canada. Director's Seminar Series, Australian National University, Nov 3rd 2016. Available at: http://rsph.anu.edu.au/news-events/first-nations-data-sovereignty-action-creation-largest-first-nations-health-research (accessed feb 6 2017)

.