Published in final edited form as:
Public Health Genomics. 2018 ; 21(5-6): 244–250. doi:10.1159/000501465.

Beyond Public Health Genomics: Can Big Data and Predictive Analytics Deliver Precision Public Health?

Muin J. Khoury[1], **Michael Engelgau**[2], **David A. Chambers**[3], **George A. Mensah**[2]

[1]Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, Georgia [2]Center for Translation Research and Implementation Science, National Heart, Lung, and Blood, Institute, Bethesda, Maryland. [3]Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Maryland

Abstract

The field of public health genomics has matured in the past two decades and is beginning to deliver genomic-based interventions for health and health care. In the past few years, the terms precision medicine and precision public health have been used to include information from multiple fields measuring biomarkers, environmental and other variables to provide tailored interventions. In the context of public health, precision implies delivering the right intervention to the right population at the right time, with the goal of improving health for all. In addition to genomics, precision public health can be driven by "big data" as identified by volume, variety, and variability in biomedical, sociodemographic, environmental, geographic, and other information. Most current big data applications in health are in elucidating pathobiology and tailored drug discovery. We explore how big data and predictive analytics can contribute to precision public health by improving public health surveillance and assessment, and efforts to promote uptake of evidence-based interventions, by including more extensive information related to place, person and time. We use selected examples drawn from child health, cardiovascular disease and cancer to illustrate the promise and current methodologic and analytic challenges of big data to fulfill the promise of precision public health.

Keywords

big data; genomics; implementation science; medicine; predictive analytics; public health

The term "precision" is increasingly used in medicine (1) and public health (2). Precision medicine is often used synonymously with genomic medicine, and precision public health has been equated with applications of precision medicine in populations (3). Applications of precision medicine (e.g. cancer genomics) are unlikely to lead to improved population

Please address correspondence and reprint requests to Muin J. Khoury MD, PhD, Office of Public Health Genomics, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, Georgia 30329. Ph (404)-498-001, muk1@cdc.gov.

Disclosure: The opinions expressed in this paper are those of the authors and do not reflect the official positions of the Centers for Disease Control and Prevention or the National Institutes for Health.

Conflicts of interest: None

health, as targeted interventions benefit only a small subset of the population. Nevertheless, as we and others have discussed, there is a bigger role for precision in public health beyond genomics (4). Precision public health can be viewed as the delivery of the right intervention to the right population at the right time, and includes consideration of social and environmental determinants of health (4). As recently stated by Richard Horton, “precision public health offers a compelling opportunity to reinvigorate a discipline that has never been more important for advancing the health of our most vulnerable and excluded communities. Precision public health is about using the best available data to target more effectively and efficiently interventions of all kinds to those most in need.” (5)

Increasingly, a large volume of health and non-health related data from multiple sources is becoming available that has the potential to drive precision implementation. The term “big data” is often used as a buzzword to refer to large data sets that require new approaches to manipulation, analysis, interpretation, and integration. (6) Such data include genomic and other biomarkers, sociodemographic, environmental, geographic, and other information. Our ability to improve population health depends toa large extent on collecting and analyzing the best available population level data on burden and causes of disease distribution, as well as level of uptake of evidence-based interventions that can improve health for all (6). In this commentary, we ask how the emerging abundance of data and its associated predictive analytics can contribute to precision public health by including more extensive information in public health assessment of disease burden, facilitators and barriers to evidence-based intervention implementation and outcome measures, as related to person, place and time. (see Table 1 for definitions and examples).

Can big data help better characterize population health outcomes, implementation needs and disparities in health and health care?

Public health and implementation scientists explore strategies for improving uptake of evidence-based health interventions that target multiple levels (from patient/person level to provider, system, community and policy interventions) (7). Implementation gaps and health outcomes are usually measured with limited data by place, persons and time.

More Precision Assessment by Place:

The use of big data sources could allow a more in-depth analysis of disease burden and implementation gaps and disparities in healthcare systems and population subgroups. For example, using small area analysis, we might be able to uncover pockets of disparities in implementation of health interventions that are often masked in analysis performed on areas such as counties or states. A case in point is the recent local burden of disease analysis on child mortality under age 5 across 46 African countries (8). When mortality is analyzed at a high spatial resolution, new maps showed major disparities in child mortality even though progress in implementation of evidence-based interventions was reported at the country level. Similarly, neighborhood deprivation metrics can assess disparities and implementation gaps within regions (e.g. wealthy towns may have micro pockets of deprivation (9)). More “precision” in geographic, community and health system analysis can pinpoint how best to target interventions to reduce morbidity in difficult to reach subpopulations and help reduce disparities.

More Precision Assessment by Person:

Similarly, in characterizing gaps and disparities in implementation and outcomes, personal characteristics of patients, providers and policy makers can be further refined beyond the use of traditional indicators such as age, gender, race/ethnicity. Genomic and other biomarkers can stratify disease outcomes and susceptibility into subgroups that reflect the underlying disease heterogeneity and potential response to different types of interventions (1). For example, while implementing cholesterol education campaigns at the population level, individuals with familial hypercholesterolemia, a genetic disorder affecting 1 in 250 people, remain largely undiagnosed as they require more intense identification, high intensity LDL lowering drugs and cascade screening in families. Failure to ascertain this high-risk population subgroup of a million or more individuals in the United States, will lead to undertreatment and worse health outcomes (10), particularly as we miss the opportunity to implement evidence-based care for this subgroup. Another example is colorectal cancer screening. While evidence-based guidelines recommend colorectal cancer screening after age 50 in the “average” population, such guidelines will miss the 1 in 280 individuals with Lynch syndrome who are at increased risk of colorectal cancer and will require screening much earlier, requiring a more targeted approach to find these genetically high-risk individuals in the population and implement effective screening, follow-up and referral to care. (11)

More Precision Assessment by Time:

Big data may also improve precision through analysis of repeated measurements of the same variables over time. The use of personal devices such as sensors, smart phones and other digital devices (12) can provide measurement of variability over time, for various health indicators such as nutrition, physical activity, and blood pressure. Most surveys rely on infrequent or cross-sectional measurements of these and other health indicators. Smartphones are used increasingly to deliver evidence-based interventions (e.g. diet/ nutrition programs, psychotherapy, exercise). The data collected through digital devices gives a picture of how the interventions have been implemented and the outcomes generated with much greater precision.

Can big data inform next generation implementation studies?

Big data-driven public health assessment studies provide directions about how to enhance implementation in sub populations and can drive implementation studies that tailor interventions by place, persons and time.

More Precision Implementation by Place:

Implementation studies evaluate delivery of interventions in real-world contexts of health care delivery systems and communities, with the goal of delivering interventions optimally across populations. The use of machine learning and decision support tools (13) adapted to specific health care delivery systems could enhance implementation of evidence-based guidelines. As discussed by Engelgau et al (14), tools of predictive analytics and big data can help identify major challenges for implementation including the identification of key barriers and facilitators within the socioecological context, various health and community policies, delivery strategies within health systems (e.g., physical infrastructure, availability of interventions), and community contexts (e.g., community resources, social deprivation, and economic issues). Predictive analytics based on big data offer new approaches to pinpoint key barriers and facilitators across the community context and lend insights into promising implementation strategies. (14)

More Precision Implementation by Person:

In order to reach subpopulations with unique health conditions, targeted intervention strategies will be needed. For example, the FH Foundation (https://thefhfoundation.org/) has developed a decision support tool using a machine-learning algorithm based on structured and unstructured data to help identify individuals with probable FH within electronic health records, large-scale laboratories and claims databases (15). The idea behind this and other similar tool is to aid clinicians to find more FH patients in selected populations, most of whom are currently undiagnosed and undertreated. Finding FH patients will also allow better implementation of recommended cascade screening in families.

More Precision Implementation by Time:

Smartphone apps can use big data to allow real-world collection and analysis over time for many evidence-based interventions (e.g., testing of adherence to medication use and longer-term measuring of outcomes over time). Apps could serve as a microcosm of a learning system that collects data on person, place and time and use the patterns detected to adjust an intervention based on its overall pattern of use and effectiveness. For example, in a recent paper (16), a randomized clinical trial of 411 adults with poorly controlled hypertension showed that patients receiving a smartphone app used repeated measures to show a small improvement in reported adherence to medication use. This pilot study points to the need to evaluate more broadly the effects of mobile health interventions on implementation processes and clinical outcomes in order to understand the context of why some interventions successfully implement while others do not. Big data drawn from social media platforms are also increasingly used in the context of public health emergencies to enhance public health surveillance, detection of disease clusters, and facilitation of communication and behavior change to accelerate disease prevention and control. Dunn et al recently reviewed progress and promise of social media interventions that could enhance precision public health. They also cautioned about little we know about the health impact of such interventions and the risks of unintended consequences. (16)

There are immediate opportunities to explore how big data can influence the next generation of implementation science. For example, NCI recently put out a funding announcement to create a set of Implementation Science Centers for Cancer Control (17). This program will support rapid development and testing of innovative approaches to implement evidence-based cancer control interventions; establish implementation laboratories in clinical and community sites; advance methods in studying implementation; and develop reliable implementation measures. (18) Implementation science centers will be an ideal forum for using big data to accelerate implementation science in cancer control. These centers are expected to use multiple sources of data and develop implementation and outcome measures using a multi-level approach, including individual, family, school, workplace, social network, community, as well as natural, built, economic, policy, institutional and health care environments. Data will be integrated and captured across observational, self-report, objective and/or real time data to identify synergistic, complementary and interacting implementation research strategies.

Similarly, NHLBI is supporting implementation science to help turn discoveries into improved population health. This will require the collaboration of diverse stakeholders to leverage “expertise from biomedical, behavioral, and social science research, as well as experts from engineering, bioinformatics, behavior economics, and the emerging field of big data science (19).”

The crucial role and current challenges of machine learning and predictive analytics

To maximize the benefits of big data in precision public health, robust analytic methods are needed for individual studies and to synthesize information across studies (20). Machine learning and predictive analytic tools are increasingly used in healthcare and population health settings to make sense of the large amount of data, both for assessment and implementation purposes (21). In principle, predictive analytics can provide novel approaches to analyze disease prediction and forecasting models and to pinpoint key barriers and facilitators to delivery of proven effective interventions. The field of oncology provides a salient application of the big data analytics is how to combine data from multiple sources (DNA germ line and tumor sequencing, gene expression, epigenetics, proteomics, etc.) along with individual and population level variables in arriving at optimal and individualized intervention strategies both for treatment and prevention. Similar complex prediction models for heart disease prediction have been developed. For the most part, however, big data predictive analytics have not provided better quantitative risk prediction models when compared to classical statistical methods such as logistic regression analysis. (22)

There are numerous gaps and methodologic limitations that need to be overcome before big data can fulfill the promise of precision public health (21). Issues involving data inaccuracy, missing data, and selective measurement are substantial concerns that can potentially affect predictive modeling results and decision-making. In addition, deficiencies in model calibration can interfere with inferences. For example, patients may see multiple healthcare providers who use different health records in different health care delivery systems. Often, data are not shared across platforms, or are incomplete. Coding for health care billing also can vary from one system to another, and health records’ completeness can vary. This can potentially create biases in effect estimation and prediction modeling. Prediction models derived in one population may not be generalizable to other settings. Furthermore, methodologic deficiencies such as systematic bias in prediction models and non-representative studies sets along with limited or differential access in subpopulations can contribute to widening of health disparities, especially for racial and ethnic minority populations. For example, recent studies have consistently shown that the accuracy of genetic risk prediction models based on genome-wide association studies (GWAS) is less among non-European populations compared to European populations (23). This is due to the fact that most GWAS have been conducted in populations of European descent. The ultimate goal of big data analytics is to improve decision making both in clinical and population setting. Providing outcome probabilities may or may not change physician, patient or health system behavior. We need to evaluate the balance of benefits versus harms of specific implementation interventions (clinical utility).

As discussed in a special recent issue (24) on machine learning, predictive analytics need to have a clear purpose. Ideally, the development and validation of prediction models should strive for external validation. The current literature contains studies on machine learning approaches that have undergone retrospective testing but not prospective evaluation. Descriptive studies using existing data can identify barriers and facilitators to implementation. Interventions to address barriers are then integrated into prospective studies. As a result, the current applications of machine learning in health care systems remain severely limited (25). These limits also apply to public health activities that are concerned with implementation challenges and health outcomes in whole populations and subgroups that are outside the health care delivery system. This issue is especially relevant in dealing with global health data, and a large effort goes into knitting together disparate and noisy information to describe health outcomes and implementation challenges globally (25). Severe constraints on resources dictate the need to develop and evaluate alternative solutions. Issues of fairness, accountability, transparency, and privacy-preserving methods, and the anticipation of deleterious effects are essential considerations for ensuring that the promise of big data to improve health and reduce health disparities does not lead to unintended and opposite effects. (25)

Conclusions

In the age of genomics and big data, more extensive information by place, person and time are becoming available to measure public health impact and implementation needs. Using a few examples, we have shown how such data may provide more information derived from public health assessment studies and next generation of implementation studies. In principle, big data could point to implementation gaps and disparities and accelerate the evaluation of implementation strategies to reach population groups in most need for interventions. However, major challenges need to be overcome. For precision public health to succeed, further advances in predictive analytics, and practical tools for data integration and visualization need to be made. As most public health and implementation scientists are not well versed in big data science, it will be crucial to offer robust training and career development at the intersection of big data and public health. This research/training agenda will help turn the promise of big data into effective precision implementation strategies to maximize the benefits of evidence-based interventions to improve population health.

Acknowledgements:

This paper was supported by the authors' salaries from the Centers for Disease Control and Prevention, and the National Institutes for Health.

References

1. Collins FS, Varmus H A new initiative on precision medicine. N Engl J Med. 2015; 372: 793–795 [PubMed: 25635347]
2. Khoury MJ, Iademarco MF, Riley WT Precision public health for the era of precision medicine. Am J Prev Med. 2016; 50(3):398–401. [PubMed: 26547538]
3. Chowkwanyun M, Bayer R, Galea S. “Precision” public health- between novelty and hype. N Engl J Med. 2018;379(15):1398–1400. [PubMed: 30184442]
4. Dowell SF, Blazes D, Desmond-Hellman S. Four steps to precision public health. Nature 2016; 540:189–191.
5. Horton R Offline: In defense of precision public health. Lancet 2018;392(10157):1504 [PubMed: 30496048]
6. Dolley S Big data’s role in precision public health. Front Publ Health 2018;6:68
7. Chambers DA. Increasing connectivity between implementation science and public health. Advancing methodology, evidence integration, and sustainability. Annu Rev Public Health. 2018;39:1–4 [PubMed: 29272164]
8. Golding N, Burstein R, Longbottom J, et al. Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals. Lancet. 2017;390(10108): 2171–2182. [PubMed: 28958464]
9. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible- The neighborhood atlas. New Engl J Med 2018;378:2456–2458. [PubMed: 29949490]
10. Knowles JW, Rader DJ, Khoury MJ. Cascade screening for familial hypercholesterolemia and the use of genetic testing. JAMA. 2017;318(4):381–382. [PubMed: 28742895]
11. Yurgelun MB, Hampel H. Recent advances in Lynch syndrome: diagnosis, treatment and cancer prevention. Am Soc Clin Onc Edcu Book 2018;38:101–109.
12. Topol EJ, Steinhubl SR, Torkamani A. Digital medical tools and sensors. JAMA.2015;313(4):353–4. [PubMed: 25626031]
13. Steinhubl SR, Muse ED, Topol EJ. The emerging field of mobile health. Sci Transl Med. 2015;7(283):283rv3
14. Engelgau MM, Khoury MJ, Roper RA et al., Predictive analytics: Helping guide the implementation research agenda at the National Heart, Lung and Blood Institute, Glob Heart 2019l14(1):75–79.
15. Banda JM, Sarraju A, Abbasi F et al. Finding missed cases of familial hypercholesterolemia in health systems using machine learning. Nature NPJ Digital Medicine 2019; 2: article 23.
16. Dunn AG, Mandl KD, Coiera EW. Social media interventions for precision public health: promises and risks. NPJ Digital Medicine, 2018; 1: 47. [PubMed: 30854472]
17. Morawski K, Ghazinouri R, Krumme A, et al. Association of smartphone application with medication adherence and blood pressure control. The MedI-SafeBP randomized clinical trial. JAMA Intern Med. 2018 6 1;178(6):802–809 [PubMed: 29710289]
18. National Cancer Institute: Implementation Science for Cancer Control Advanced Centers. Accessed online December 3, 2018 https://grants.nih.gov/grants/guide/rfa-files/RFA-CA-19-006.html
19. National Heart, Lung and Blood Institute. Stimulating T4 Implementation Research to Optimize Integration of Proven-effective Interventions for Heart, Lung, and Blood Diseases and Sleep Disorders into Practice (STIMULATE). Accessed online December 3, 2018 https://grants.nih.gov/grants/guide/rfa-files/rfa-hl-19-014.html
20. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. JAMA 2018;320(1):27–28. [PubMed: 29813156]
21. Parikh RB, Kakad M, Bates DW. Integrating Predictive Analytics Into High-Value Care: The Dawn of Precision Delivery. JAMA 2016; 315(7):651–2 [PubMed: 26881365]
22. Christodoulouy E, Ma J, Collins GS, Steyerberg EW, Verbakel EW, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;2 11, pii: S0895-4356(18):31081–3.

23. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genetics. 2019;51(4):584–591. [PubMed: 30926966]

24. Nevin L, on behalf of PLOS Medicine editors. Advancing the beneficial use of machine learning in health care and medicine: Toward a community understanding. PLOS Medicine 2018; 15(11): e1002708. [PubMed: 30500811]

25. Flaxman AD, Vos T. Machine learning in population health: Opportunities and threats. PLOS Medicine 2018; 15(11): e1002702. [PubMed: 30481173]

Table 1

Examples of the Potential Applications of Big Data in Precision Public Health, by Person, Place and Time

Source of Big Data	**Public Health Assessment** **(Characterizing population health outcomes and implementation needs and disparities in populations; natural experiments)**	**Implementation Studies** **(Conducting multilevel intervention studies to improve implementation and health outcomes)**
Place (zip codes, census-tract data, countv-level data, neighborhood characteristics, health systems, and linkages with environmental and socio-economic characteristics)	• Use refined geographic area analysis in surveillance (e.g. small area analysis) • Link geographic data with other sources of information such as environmental exposures	• Target implementation strategies across geographic locations or health systems (e.g. randomizing use of decision support tools to prompt providers/patients in different healthcare systems)
Person (demographic characteristics, genetics, biomarkers, electronic health records, personal devices, social media use)	• Use molecular subtypes in cancer surveillance, • Use of family history and genetic factors to stratify the population by risk level	• Target implementation studies based on characteristics of patients, providers, and policy makers
Time (longitudinal data of individual personal characteristics and environmental/ spatial information)	• Use longitudinal data in addition to cross sectional information in assessment (e.g. repeated measures in public health surveys)	• Collect longitudinal data in implementation studies (e.g. using smart phone apps to provide reminders to medication use, and to measure adherence to hypertension treatment)