5 Big data: A new dawn for public health?

Cristina Gall and Elina Suzuki

Changing disease patterns and escalating costs of care make prevention, health promotion and public health pressing concerns and key parts of addressing the challenges facing health systems. Already, the technical capacity exists to pursue a new type of ‘precision’ public health – applying the principles and technology of precision medicine to disease prevention and public health policy. As the need for evidence-based policies grows, big data seems to hold the key to dramatic, rapid improvements to help promote health and prevent disease. At the same time, health systems have been slow to adopt new technologies, and must consider how these new approaches will affect privacy. In the face of these developments, public health policy makers need to discern the most effective ways in which they can leverage big data, as well as how to best address the challenges associated with these novel technologies.

5.1. Introduction

Big data keep getting bigger. It is estimated that 2.5 quintillion bytes of data are now created daily – that is, 2.5 billion billion or $2.5x10^{18}$ bytes (IBM, 2017[1]). 90% of the data in the world today were created in the last few years alone (DOMO, 2017[2]). By 2020, it is estimated that 1.7 megabytes of data will be created every second for every person in the world (DOMO, 2018[3]). As the capacity to generate and analyse large amounts of data continues to increase worldwide, the term "Big Data" has become ubiquitous. Big data are increasingly relevant to many different kinds of research and knowledge creation activities, across a variety of domains (Box 5.1).

The chapter explores the possibilities, risks and challenges of deploying Big Data in in the sphere of public health. As the need for evidence-based policies becomes increasingly pressing, big data and their associated analytic tools hold the promise of vastly improved strategies to promote health and prevent disease. In the face of these developments, public health policy makers need to discern the most effective ways in which they can leverage big data, as well as how they can best address the challenges associated with these novel technologies.

Big data, and big data analytics, can be used at all three levels of health promotion and disease prevention – research, surveillance, and intervention – by:

- **Allowing a more precise identification of at-risk populations**, through a more comprehensive understanding of human health and disease, including the interaction between genetic, lifestyle, and environmental determinants of health;
- **Enabling better surveillance** of both communicable and non-communicable diseases; and
- **Facilitating better targeted strategies and interventions** to improve health promotion and disease prevention.

The public health sector, however, has been a relatively slow adopter of big data analytics. While efforts to leverage big data in public health policy making are starting to gain momentum, such as the European Union's "Big Data supporting Public Health policies" programme, there is a need for a more systematic focus on and resource allocation for such initiatives. In a recent OECD survey, few respondents reported using any kind of non-traditional data sources for public health, and most of these initiatives are still in exploratory stages. Even when big data sources are used, such initiatives are typically limited to disease surveillance and the identification of isolated risk factors.

Using big data primarily for knowledge accumulation, rather than effective interventions, can be problematic: while harnessing big data can help answer many questions, it can also exacerbate the *"A lot is known, but little is put into practice"* policy dilemma. In other words, a risk also exists in deploying scarce resources to accumulate more health knowledge that then remains unused. Leveraging big data to help distil knowledge into clear public health interventions remains a major challenge. Though limited in scope, existing examples – such as Geisinger's GenomeFIRST initiative, which identifies patients at a high risk of treatable conditions, through whole exome sequencing – underscore the potential of harnessing new data for prevention and care. But such programmes need to be validated and applied at scale.

Box 5.1. Big data – a primer

Defining Big Data

The term "Big Data" is often poorly defined. In practice, it is "often described 'implicitly' through success stories or anecdotes, characteristics, technological features, emerging trends or its impact to society, organizations and business processes" and in reference to "a variety of different entities including social phenomenon, information assets, data sets, analytical techniques, storage technologies, processes and infrastructures" (De Mauro, Greco and Grimaldi, 2015[4]). The term "big data" is also frequently intertwined with the concept of "big data analytics" (Box 5.2). Various formal definitions have been proposed, which share a set of concepts and characteristics (De Mauro, Greco and Grimaldi, 2015[4]; Sivarajah et al., 2017[5]).

- **High volume**: the large scale of data sets.
- **High velocity**: the high rate of data inflow, as well as the speed with which it needs to be processed and analysed.
- **High variety**: the heterogeneity of data (i.e., diverse and dissimilar data formats).
- **Specific data extraction and analysis methods**, collectively known as "big data analytics".
- **Value creating**: producing valuable insights, which cannot be obtained from traditional data sources.

"Big Data" can be structured, semi-structured, or unstructured, and can come from "sensors, devices, video/audio, networks, log files, transactional applications, web, and social media – much of it generated in real time and in a very large scale" (IBM, n.d.[6]). In addition to Volume, Velocity, Variety, and Value, other qualifiers have been proposed – such as veracity (unbiased truthfulness), validity (accuracy), and volatility (whether the data is still valid) – to reflect issues such as data accuracy and utility (Bansal et al., 2016[7]). The concepts of "veracity" and "validity" refer to the need to use analytical methods that can account for the unreliability of big data, such as various biases (IBM, n.d.[8]).

Types and sources of data: an overview

Big data sources that can be used for public health include:

- *Structured* data, e.g. data from electronic medical records (EMR) and electronic health records (EHR), participatory surveillance systems (e.g., crowdsourcing, crowdmapping).
- *Semi-structured* data, e.g., data from health monitoring devices.
- *Unstructured* data, which presents the greatest potential to use non-health data to enhance public health. Sources of unstructured data include, among others:
  - Social media and online data, i.e. "virtual digital trails".
  - Consumption data, i.e. "real-life digital trails".
  - Spatial/geographic data.
  - Physical environment data.

Real-time data, such as sensor data from wearable technologies and environmental sensors that measure variables like air pollution and airborne allergens, can also enable the delivery of more personalised prevention strategies. The large amounts of environmental data (e.g. weather patterns, pollution levels, water quality) collected in non-health sectors can similarly be used to inform public health policies. In the context of climate change, these types of data will become increasingly important, particularly for infectious disease surveillance.

5.2. OECD countries are using new analytical tools to better link electronic health databases and draw policy insights for public health purposes

Results from the 2018 OECD Survey on Uses of Data and Digital Technology indicate that many OECD countries have begun to harness big data for public health purposes, though many of these remain at an early stage of development. These efforts have largely focused on using new improvements in computing and analytical power to take advantage of existing health databases in ways that would previously been too time- or resource-intensive.

In Australia, for example, the Data Integration partnership for Australia (DIPA) has supported the Department of Health to identify adverse events associated with medicines through analysing data from the Pharmaceutical Benefits Scheme. Through analysing data from the Pharmaceutical Benefits Scheme, as well as datasets such as the Medicare Benefits Schedule and hospital discharge data, the Australian government is working towards identifying and acting on medicine safety issues earlier, with the goal of increasing patient safety and reducing hospitalisation and treatment costs.

In the Czech Republic, the National Registry of Reimbursed Health Services, containing comprehensive reimbursement data from health care administrative records, was launched in 2018. The National Registry offers new possibilities to evaluate public health interventions, including screenings. In Norway, digital health information – including health registries and national surveys – have been used to develop municipal and regional public health profiles, which are used actively by municipalities to improve public health and by the media to compare local populations to performance across Norway.

5.2.1. New analytical techniques can enhance public health policy making

Traditional public health is data-poor and has traditionally lacked the key big data characteristics: high volume, high variety, and high velocity. Epidemiologic research generally relies on long-term, longitudinal, relatively small- or medium-scale cohort studies, in which data are gathered through participant questionnaires, physical examinations, and, for outcome data, health records. These data sources are often difficult to obtain and work with. Similarly, public health surveillance – defined as the ongoing systematic collection, analysis, and interpretation of data, as well as the dissemination of these data to public health practitioners, clinicians, and policy makers (Richards et al., 2017[9]) – and public health interventions have traditionally been performed through time-consuming, error-prone methods. These methods pose significant timeliness and efficiency limitations and suffer from time lags and lack of spatial resolution. Table 5.1 summarises the implications of big data for public health.

Table 5.1. Summary of the three Vs of big data and their implications for public health

Name	Meaning	Examples	Opportunities and Challenges	Implications for public health
Volume	Data sets with more observations	National electronic health record databases, social media datasets	Power to precisely measure unexpected associations, though potentially without substantive relevance	Evolutionary/incremental
Variety	Datasets with variables from different sources; more variables per observation	Neighbourhood data added to a phone survey	Capacity to assess complex interactions, but more complicated variable selection	Evolutionary/incremental
Velocity	Data collected and analysed in real-time	Medication adherence intervention messaging adapted to subject response pattern	Potential to design dynamic interventions	Potentially revolutionary

Source: Mooney, Westreich and El-Sayed (2015[10]), "Commentary: Epidemiology in the era of big data", http://dx.doi.org/10.1097/ede.0000000000000274.

However, the development of new analytical techniques has enabled policy makers to harness existing health datasets in ways that can transform existing disparate databases into a larger set of data with key big data characteristics.

In *non-communicable disease prevention*, big data can help better understand and address modifiable behavioural risk factors that contribute to a large fraction of the non-communicable disease burden (e.g. diet, physical activity, tobacco use). Big data analytics (Box 5.2) can enable policy makers to more effectively assess these risk factors at the population, subpopulation, and individual levels, as well as to design better targeted interventions aimed at mitigating them. In particular, big data analytics can help understand, at a causal level, how hereditary risk factors – as well as combinations of risk factors – interact with behaviour and the physical and social environment. In *communicable disease prevention*, hybrid tools that combine traditional methods and big data analytics can enhance communicable disease surveillance by harnessing novel data streams to complement – rather than replace – traditional methods.

Box 5.2. What are big data analytics?

The specialised tools and analytical methods needed to extract useful insights from big data sources are transforming the use of big data for public health. These specialised technologies are collectively known as "big data analytics". Using big data analytics can enhance public health at the research, surveillance, and intervention levels. It can thus enable the design and implementation of more effective, evidence-based public health policies.

Predictive analytics are the most common type of big data analytics. They represent one of the three main types of analytics, the other two being descriptive and prescriptive analytics (Sivarajah et al., 2017[5]). While descriptive analytics are backward looking and are used to measure facts and summarize data, predictive analytics are forward looking and use past or current data to make predictions about the future, through tools such as machine learning, data mining, and statistical models.

- *Machine learning* is an important tool for predictive analytics and refers to the design of algorithms that allow computers to "learn", i.e., progressively improve performance on a specific data-related task by adapting to patterns in data, with the aim of knowledge discovery and automatic decision making (Chen and Zhang, 2014[11]). Machine learning can be supervised, unsupervised, or semi-supervised (Fuller, Buote and Stanley, 2017[12]). While often conflated, machine learning and predictive analytics represent distinct concepts.
- *Data mining*, also known as knowledge discovery, refers to the extraction of potentially useful information from data, often using similar techniques as machine learning (Fuller, Buote and Stanley, 2017[12]).

Prescriptive analytics refer to optimization and randomized testing and address the "So what?" types of questions that arise after the data have been analysed through either descriptive analytics, predictive analytics, or a combination of both (Sivarajah et al., 2017[5]).

5.2.2. Big data can improve the identification of population- and person-level risk factors

The potential of big data to allow for a more precise identification of risk factors is attracting more and more interest from researchers and policy makers alike. An increasing number of large-scale population studies aim to leverage big data to pinpoint specific risk factors more precisely, using more data than traditional epidemiologic studies are able to. There is a particularly enthusiastic focus on identifying hereditary risk factors through genetic testing.

This approach, however, can backfire: the temptation to use big data to find new, exciting risk factors or more precisely measure the effects of known ones could come at the expense of engaging with "the broader causal architecture that produces population health"; in other words, the "proliferation of causal effects – typically identified through an approach that aims to isolate risk factors for particular outcomes – presents a conundrum for scientists, let alone the lay public, to synthesize and form evidence-based recommendations that can promote health." (Keyes and Galea, 2015[13])

The capacity to leverage bigger and better data to measure the effects of precise risk factors therefore needs to be carefully weighed against *what matters most for population health* (Keyes and Galea, 2015[13]). Indeed, the potential of big data to improve public health lies not in better measurements of various isolated risk factors, but in the ability to analyse the complex, dynamic interactions between human behaviour/lifestyle ("behavioural phenotypes"), genetics, and the physical and social environment to determine what matters most for public health policy.

There is therefore a need to move from clinical validity (confirming robust relationships between risk factors and disease) to clinical utility: in other words, when it comes to the public health impact of big-data driven research, researchers and policy makers should address the *"Who cares?"* and *"So what?"* questions.

Furthermore, discussions of big data often go beyond the technological and analytical aspects and suggest a "mythological" dimension: "the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy" (Boyd and Crawford, 2012[14])

Bigger is not necessarily better

When it comes to data, however, bigger is not always better. Big data sets, regardless of their size, are subject to biases that need to be well understood and accounted for to avoid misinterpretation and incorrect conclusions. The temptation to over-rely on big data without having robust methodologies in place for interpreting it can lead to *apophenia*: "seeing patterns where none actually exist, simply because enormous quantities of data can offer connections that radiate in all directions" (Boyd and Crawford, 2012[14]). For example, one analysis used data mining techniques to show that a strong, but spurious association could be found between the evolution of the S&P 500 stock market index in the United States and butter production in Bangladesh (Leinweber, 2007[15]).

In addition, big data is not necessarily "whole data": the large size of a data set does not necessarily mean it is a representative sample of a certain population (Boyd and Crawford, 2012[14]). Overlooking this issue when analysing big data sets can lead to biased results. Strong methodological standards for interpreting large data sets and accounting for biases must therefore be at the forefront of big data analysis.

5.3. Harnessing novel data sources represents a new frontier for public health

Integrating and harmonising data sources from the traditional medical model (e.g. health organisation databases) with novel big data sources, such as social media and internet query data, and wearable electronic devices, holds the potential to deliver gains in public health. Linking data from a wide variety of public sources can yield particularly powerful public health insights. For example, data from the traditional medical model can be combined with data on social determinants of health, offering new targets for personalised care and intervention.

While the majority of responding countries to a recent survey (9 of 15 OECD countries) outlined uses of electronic health information to inform public health initiatives, just three countries (Canada, Estonia, and the Netherlands) are using new, non-traditional sources of data to improve public health.

In Estonia, a large-scale clinical pilot on personalised medicine is using genomic data to target preventive care services at patients who are at risk for cardiovascular disease or breast cancer. The genetic data from more than 150 000 Estonians has been used to develop algorithms at the Estonian Genomic Centre, resulting in a clinical decision support software to help GPs detect patients at risk for cardiovascular disease or breast cancer. Over the long-term, the program intents to expand to include other preventable diseases.

In Canada, the Public Health Agency's Centre for Surveillance and Applied Research looked at how information from wearable technologies and physical activity applications ("apps") promoting healthy living can help to supplement or replace survey-based health indicators. The Centre also worked to use social media postings to inform surveillance for self-harm. While most initiatives are in the exploratory stage, a number of programs were launched to the public. Carrot Rewards, a healthy living mobile app and platform, offered participants private-sector loyalty points in exchange for healthy living behaviours, including physical activity, healthy eating, vaccinations, mental health, and reducing alcohol and tobacco consumption. More than 750 000 users participated in Carrot Rewards before the program was shut down in the summer of 2019 after filing for bankruptcy (Marotta, 2019[16]). In addition to the services provided to participants, Carrot Rewards collects user information that is shared with Canada's Public Health Agency, allowing the Agency to better target interventions to specific populations and geographic areas.

In the United States, some health insurers and care providers use comprehensive "health scores", developed by combining publicly available socioeconomic data through the *LexisNexis Socioeconomic Health Attributes Model*, to develop tailored care plans based on individual patient need. The attributes used in the model were clinically validated against claims data to confirm their predictive power and are clustered in categories such as address stability, education, and income (LexisNexis, 2019[17]). Linking big data sources in this way facilitates a better understanding of individual health risk and can thus enable improved health care personalisation, but poses data privacy concerns that must be balanced against the benefits of such an approach.

5.3.1. Novel data sources are uncommon as big data sources for public health

Recent technological advancements have led to new approaches to disease prediction and monitoring, with mixed results. Efforts have begun to systematically mine "virtual digital trails", such as social media and internet query data, to assess health-related behaviours and attitudes in near-real time. Current initiatives are promising, but most are limited to one-off projects. Moreover, where used on a wider scale this approach has delivered mixed results, with a number of high-profile failures, including Google Flu Trends, pointing to the challenges implicit in using these new data sources. Scaling up these approaches in a way that also protects data privacy and security is needed.

Infodemiology refers to systematically mining, aggregating, and analysing unstructured, textual, openly accessible online information to inform public health policy (Eysenbach, 2011[18]). *Infoveillance* refers to using infodemiology metrics for surveillance and trend analysis (Eysenbach, 2011[18]). Crowdsourcing can also be used as an alternative to infoveillance to collect public health data using big data approaches.

Social media platforms, in particular, represent sources of rich observational data for infodemiology and infoveillance (Kim, Huang and Emery, 2016[19]). For instance, in the area of non-communicable disease prevention, the dynamics of social networks can be studied to discern patterns of how social factors influence unhealthy behaviours, such as smoking (Andreu-Perez et al., 2015[20]). Furthermore, combining insights from social media data with geolocation data can enable a better understanding of patient behaviours and social demographics; it has been used to study, for instance, influenza outbreaks and antibiotic misuse (Andreu-Perez et al., 2015[20]).

Methods for collecting, filtering, and analysing social media data need to be clearly and transparently reported. As discussed further on in this chapter, data transparency is a key driver of many successful big

data initiatives, such as the “smart cities” and “smart health” initiatives. Infoveillance challenges and limitations include: privacy concerns, difficulties in establishing causal links due to lack of context, isolating signal from noise, and lack of generalisability of Internet or social media users, due to the overrepresentation of certain population groups.

Mining “virtual digital trails” – such as social media and Internet query data – offers the opportunity to assess self-reported health- related attitudes and behaviours, such as those pertaining to non-communicable diseases and their risk factors, in near real time; it can thus complement more traditional non-communicable disease surveillance data. For example, research has linked anger and stress expressed on Twitter to an increased risk of heart disease (Andreu-Perez et al., 2015[20]; Eichstaedt et al., 2015[21]). Another study was able to accurately predict the likelihood of smoking and alcohol consumption based on the user’s behaviour on Facebook, including how many and what they “Liked” on the website (Kosinski, Stillwell and Graepel, 2013[22]).

Social media infoveillance can also allow public health entities to detect whether specific communities may need certain health or social services, particularly in the case of stigmatised health issues, such as drug use. This awareness can enable more targeted surveillance and enhanced interventions. For instance, one study used Twitter data to identify online communities of illicit, recreational, and medical cannabis users connected to specific dispensary accounts (Baumgartner and Peiper, 2017[23]).

“Real-life digital trails” are “signals produced by people’s everyday actions, recorded digitally through devices and sensors measuring individuals’ movements and behaviours” (Balicer et al., 2018[24]). The ways in which an individual interacts with digital technologies – such as through texts, calls, and social media posts – are markers of his/her “digital phenotype”, which, when combined with other sources of data, such as clinical data, can allow for early disease detection and intervention (Jain et al., 2015[25]). For instance, in one study, behavioural indicators obtained from phone usage data were strongly linked to depression severity (Saeb et al., 2015[26]). Table 5.2 summarises the advantages, challenges and potential contribution to disease surveillance of sources of traditional health data, virtual digital trails and real-life digital trails.

In addition, sensor data from wearable technologies and environmental sensors can provide insights on a variety of lifestyle risk factors and aid in chronic disease management (Balicer et al., 2018[24]). Smart devices represent a promising source of crowdsourced big data that can be leveraged to track the spread of certain infectious diseases. For instance, “smart thermometers” connected to a mobile phone application provide de-identified fever data that can help track influenza activity in real time. This crowdsourced data can be used to improve influenza surveillance and forecasting by complementing traditional models, which rely on data from hospitals and clinics and which tend to lag behind real-time influenza activity. One study, which analysed data from over 8 million temperature readings generated by almost 450 000 “smart thermometers”, showed that the data were highly correlated with information obtained from traditional disease surveillance systems and could potentially predict influenza activity up to two to three weeks in advance (Miller et al., 2018[27]).

The use of such smart devices is generally limited to individual purchasers who can afford to buy them, making the data susceptible to socioeconomic biases. Given the potential of these smart devices to both capture real-time infectious disease activity at a population level and help generate improved disease forecasts, policy makers should explore strategic partnerships with the private sector that could facilitate a more widespread use of such devices – provided that they have been validated against traditional surveillance methods – and a systematic integration of their data into national communicable disease surveillance systems.

Table 5.2. Big data sources: advantages, challenges, and potential contributions to non-communicable disease (NCD) surveillance

	Advantages for NCD surveillance	Challenges for NCD surveillance	Potential contribution to NCD surveillance
Health organisation databases	• Passively recorded, clinically based (credible source of clinical data) • Comprehensive EHR databases (wide range of diseases and clinical information) • Clinically representative for reporting on epidemiology, morbidity, and health service use related to NCDs • Ability to assess outcomes in relation to explanatory and risk factors • Some EHR databases contain longitudinal data with continuous membership	• Poor standardisation and harmonisation in coding and data structure pose challenges to linking and comparing data from various EHR sources • Issues of coding validity and consistency • Might not be representative of populations outside of the system • Often do not have information on behavioural risk factors, disability, and functional status	• Add breadth and depth to NCD surveillance • Ability to assess risk factors in relation to outcomes • Identify small-area variation and subgroups for intervention targeting • Flexibility to identify and respond to new or emerging problems • Identification of long-term trends • Identification of trends in health service use and correlation of utilisation with epidemiology
Virtual digital trails	• Rich, accessible, and inexpensive source of quantifiable qualitative information • Subjective, and representative of individuals' perceptions and perspectives • Some social network data offer the opportunity for technological leapfrogging and inclusion of previously excluded populations, particularly in urban settings	• Biases in who is represented, because only some segments of the population will participate • Biases in the types and accuracy of content that users are communicating publicly on social media • Dependence on changing platforms and technologies • Potential for "ecological fallacy" errors	• A situational awareness tool that can be integrated into existing surveillance frameworks as complementary data • Provide indicators of behavioural factors and functional status
Real-life digital trails	• High-resolution, real-time data • Highly sensitive to detection of abrupt changes or seasonal patterns • Some real-life digital trail data offer the opportunity for technological leapfrogging and inclusion of previously excluded populations, particularly in urban settings	• Difficult to identify the factors that cause or influence the observed trends • Sensitive to issues of individual privacy • Reliability, validity, and accuracy of these data sources for health surveillance have yet to be determined • Often proprietary data, owned by industry stakeholders • Operationalisation is contingent on data-sharing frameworks that uphold individual privacy and the competitive advantage of the data providers	• Complementary source that offers insights on new aspects of health behaviours • Can detect abrupt changes or seasonal patterns in risk factors • Information about new population segments not captured through traditional health data surveillance • Can enhance epidemiological research that monitors the association between environmental exposures and health outcomes • Additional time points and granular, local-level data

Source: Balicer et al. (2018[24]), "Using big data for non-communicable disease surveillance", http://dx.doi.org/10.1016/S2213-8587(17)30372-8.

Wearable sensors, including mobile phone accelerometers, can provide valuable data about individual behaviour and lifestyle factors, such as patterns of physical activity and sleep, which are linked to a variety of health outcomes. Combined with smart technology, the data collected by these sensors can enable personalised health promotion interventions, such as mobile phone applications that prompt a user to engage in exercise if an unhealthy pattern of physical inactivity is detected.

Data from such wearable devices, however, tends to suffer from a large amount of noise (e.g. a wrist-worn accelerometer may not be able to differentiate whether different types of arm movement indicate the user is exercising or not (Gelfand, 2019[28])). Complex statistical methods are thus needed to analyse these data. Despite these issues, wearable sensor-based interventions hold the potential to enable health policy makers to implement large-scale, highly personalised, low-cost health promotion interventions, in partnership with researchers and the private sector.

5.3.2. Big data approaches can complement traditional disease surveillance methods

Traditional infectious disease surveillance is typically based on epidemiological data collected by health institutions. While these data have a high degree of veracity, they also suffer from several disadvantages, including: time lags, due to lack of human resources or problems when aggregating data from different sources; high cost; and insufficient local granularity. In contrast, big data streams – such as internet queries, social media data, and crowdsourced data – can be tracked in real time and at a local level with minimal cost, but have their own biases that need to be accounted for. (Bansal et al., 2016[7]; Simonsen et al., 2016[29])

Communicable disease surveillance is therefore one of the most exciting opportunities created by big data in the realm of public health. These novel data streams can improve the timeliness and the spatial and temporal resolution of infectious disease tracking, as well as provide access to "hidden" populations (Bansal et al., 2016[7]; Simonsen et al., 2016[29]). Some recent successes – including Health Map's successful identification of a haemorrhagic fever outbreak in West Africa in 2014, which was subsequently identified as Ebola – point to the potential of complementing traditional surveillance methods with new approaches. Big data streams can also go beyond disease surveillance and provide information on specific behaviours and outcomes related to vaccine or drug use.

Nevertheless, while such new methods of infectious disease surveillance are promising, they may not always be mature enough and should be validated against established infectious disease surveillance systems. Policy makers and researchers must remain vigilant to avoid "big data hubris", which refers to the assumption that "big data are a substitute for, rather than a supplement to, traditional data collection and analysis" (Lazer et al., 2014[30]) – in other words, the assumption that high-volume, high-velocity data can replace smaller, high-veracity data and traditional data analysis methods (Fuller, Buote and Stanley, 2017[12]). Some past examples of big data-driven surveillance systems that suffered from big data hubris – such as, notably, Google Flu Trends (Box 5.3) – failed to deliver reliable information.

Using digital data for public health surveillance presents a set of challenges, such as: lack of demographic information; issues of representativeness, as these data tend to represent a limited segment of the population that may not include certain age categories, or may include fewer elderly individuals; and spatial and temporal uncertainty – for instance, someone may be researching a family member's illness that occurred in a different city several weeks earlier. (Bansal et al., 2016[7])

Furthermore, before relying on these types of novel data sources, public health authorities should assess the impact of local conditions on the reliability of predictive algorithms. For instance, a dengue surveillance algorithm that used Internet query data to create an index of dengue incidence worked well in areas with high incidence of dengue and favourable vector climate conditions, but was not a reliable predictor of dengue incidence in areas with low incidence and an unfavourable climate (Gluskin et al., 2014[31]).

Novel methods should therefore complement – rather than replace – traditional methods (Bansal et al., 2016[7]). Policy makers should aim for hybrid tools that combine traditional methods and big data analytics to enhance communicable disease surveillance. Hybrid tools make use of the advantages of novel big data sources – such as timeliness, scale, and fine granularity – while maintaining a direct link to disease through traditional surveillance systems (Simonsen et al., 2016[29]). Where well-performing prediction models are already in place at the national level (e.g. the CDC flu prediction model in the US), big data analytics could be used to enhance local-level data.

Box 5.3. The rise and fall of Google Flu Trends: lessons learned from a big data failure

Google Flu Trends (GFT) began functioning in 2008. The idea behind it was simple, yet revolutionary: it would monitor flu outbreaks worldwide based on Internet searches for flu-related terms. It was initially heralded as an innovative way to harness online search data to predict flu trends faster than traditional prediction systems – approximately two weeks ahead of the Centers for Disease Control and Prevention data.

In April 2009, GFT missed the onset of the non-seasonal influenza A-H1N1 pandemic, a failure attributed to changes in people's online search behaviour due to the exceptional nature of the pandemic (Cook et al., 2011[32]). Then, in February 2013, "Nature" reported that GFT was predicting more than double the proportion of influenza-like illness (ILI) doctor visits compared to the Centers for Disease Control and Prevention (CDC), whose estimates are based on surveillance reports from across the US – in spite of the fact that GFT had, in fact, been built precisely to predict CDC ILI report (Butler, 2013[33]; Lazer et al., 2014[30])

GFT likely suffered from two main problems: "big data hubris" and algorithm dynamics. "Big data hubris" refers to the "often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis (Lazer et al., 2014[30])." Algorithm dynamics refer to the continuous modification data engineers make to the algorithm, which can cause unforeseen effects, as well as to changes in user behaviour, which can be driven by the algorithm modifications themselves.

Key lessons that can be drawn from the failure of GFT include:

- **Big data prediction models need to be transparent and replicable.** This can be achieved through collaborations between academia, public health policy makers, and the private entities that create the algorithms and own the data. Closing off the methods and data can make it difficult to validate them and rely on their predictions for decision-making (Lazer and Kennedy, 2015[34]).
- **Algorithm dynamics should be well understood and continuously analysed for potential systematic measurement errors.** This should be done particularly carefully when intentional changes are made to the algorithm for commercial purposes, and when these changes prompt changes in users' search behaviour over time.
- **Big data analytics should complement – rather than attempt to replace – traditional public health surveillance methods.** For instance, if a well-performing flu prediction model is already in place at the national level (e.g. the CDC model in the US), big data analytics can be used to enhance awareness about flu prevalence at local levels, where national data models may not be as useful. The high volume and velocity of the data should not supplant existing "smaller, slower data" if issues of reliability and measurement validity are at stake.

Source: Lazer, D. et al. (2014[30]), "The Parable of Google Flu: Traps in Big Data Analysis", http://dx.doi.org/10.1126/science.1248506, Lazer and Kennedy (2015[34]), "What We Can Learn From the Epic Failure of Google Flu Trends", https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/.

Participatory surveillance systems that use crowdsourcing or crowdmapping have been growing, but have yet to be used at a large-scale by public health authorities. Examples include Influenzanet – a crowdsourcing system for self-reported flu symptoms used in ten European countries, ResistanceOpen – which aggregates publicly available data on antimicrobial resistance, and Health Map – which uses online data to track infectious disease outbreaks around the world. An extension of HealthMap is "Flu Near You", which provides a crowdsourced, real-time "flu map" that shows influenza activity in the United States. As

the integration of participatory surveillance with traditional surveillance systems can significantly improve infectious disease surveillance, policy makers should explore ways in which they can develop and contribute to both national and international efforts in this area.

Surveillance systems that track infectious diseases form the backbone of communicable disease monitoring and controlling, but tend to suffer from time lags and insufficient spatial resolution. They remain primarily based on manually collected data, which is then aggregated into national or regional reports, thus lacking local-level granularity. Novel surveillance approaches that use big data streams, including electronic health (e-health) patient records, unstructured digital data sources, and participatory surveillance should be leveraged to help strengthen infectious disease surveillance systems. (Bansal et al., 2016[7])

5.3.3. Policy makers are starting to explore big data for precision public health

Precision medicine is quickly moving to the forefront of many health systems' vision for the future, driven by significant advances in genetic research. Combining traditional medical data with novel data and technologies from fields such as genomics, epigenomics, transcriptomics, proteomics, metabolomics, and phenomics is enabling a better understanding of disease pathogenesis and more targeted diagnoses and treatments, especially for cancer.

In 2016, the United States launched the USD 215 million Precision Medicine Initiative. This initiative includes, among other projects, the *All of Us* research programme, a 1-million participant study whose mission is "to accelerate health research and medical breakthroughs, enabling individualized prevention, treatment, and care" by studying "individual differences in lifestyle, environment, and biology" (National Institutes of Health, 2019[35]).

While precision medicine is seen as "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle in each person" (National Alliance Of Healthcare Purchaser Coalitions, 2018[36]), most efforts in this area have thus far focused on improving treatment, rather than prevention. The next step for health systems is to begin to use big data to transform precision medicine into precision health, by applying new insights to not only improve diagnosis and treatment, but also health promotion and disease prevention. In the United States, for example, the program *Connecting Public Health Information Systems and Health Information Exchange Organizations* gathered information on how public health jurisdictions use existing health information exchange (HIE) organizations as a means of sharing information with providers (The Office of the National Coordinator for Health Information Technology, 2017[37]).

Public health policy makers should leverage research discoveries to enable more targeted disease prevention strategies at a population level. Such prevention strategies can become an important part of "wellness planning" and health management, and a stepping stone towards providing "truly anticipatory "health care," instead of the responsive "sick care" that has long been the health care system's default" (Willard, Feinberg and Ledbetter, 2018[38])

In addition, current precision health approaches mostly focus on individual genetic variability. Most "personalised prevention" has been based on hereditary risk factors, such as cancer-causing mutations, that can be detected with genetic testing. In the United Kingdom, for example, genetic testing is made available to the family members of people with certain mutations.

In the United States, Geisinger's MyCode/ GenomeFIRST initiative provides screening for genetic variants linked to a higher risk for certain medically actionable conditions, such as the BRCA1 and BRCA2 variants associated with an increased risk of developing breast cancer (Williams et al., 2018[39]). Based on ongoing results, it is estimated that approximately 3.5% of study participants will be found to carry risky gene variants (Trivedi, 2017[40]) (Box 5.4).

Box 5.4. The Geisinger MyCode Community Health Initiative: implementing precision health by using genetic screening to prevent disease

The MyCode Community Health Initiative is a precision medicine project started by the Geisinger Health System in Pennsylvania and New Jersey. Geisinger serves approximately 4.2 million residents, with about 1.5 million unique patient visits annually. About one-third of Geisinger patients are insured by the provider-owned Geisinger Health Plan.

MyCode began in 2007 as a biorepository for discovery research, as part of Geisinger's mission to be a "learning health care system". The initiative now includes a system-wide biobank designed to store blood and other samples for research use by Geisinger and Geisinger collaborators.

Over 220 000 Geisinger patients have consented to participate in the initiative and approximately 100 000 whole exome sequences have been completed, as of August 2018. The findings are used for both research discovery (Geisinger mines DNA data and anonymised electronic health records for links between gene variants and diseases), as well as disease prevention, through the GenomeFIRST program, which was launched in 2013 as part of the MyCode initiative. If a patient-participant is found to carry one of 76 gene variants known to be causally linked to higher risk for one or more of 27 conditions, the patient-participant and their care provider are notified. Patient-participants can then opt to follow up with their primary care or specialist provider, meet with a member of Geisinger's clinical genomics team, or both. All of the 27 conditions are "medically actionable": they can be treated, managed, or prevented. Gene variants that raise the risk for certain conditions that cannot be treated or prevented, such as Alzheimer's, are not disclosed to patients-participants who carry them.

It is estimated that approximately 3.5% of study participants will be found to carry risky gene variants. What remains to be seen is whether this number is significant enough for the initiative to be cost-effective in the long run: whether the cost savings from patients who do not go on to develop the diseases for which they carry the risky gene variants will outweigh the cost of sequencing so many patients' exomes. If so, widespread genomic screening could become an important strategy for implementing precision health at a population level, particularly if the cost of genome sequencing continues to decline.

Source: Williams, M. et al. (2018[39])), "Patient-Centered Precision Health In A Learning Health Care System: Geisinger's Genomic Medicine Experience", http://dx.doi.org/10.1377/hlthaff.2017.1557, Trivedi (2017[40]), "Is health care ready for routine DNA screening? A massive new trial will find out", http://www.sciencemag.org/news/2017/10/health-care-ready-routine-dna-screening-massive-new-trial-will-find-out.

It remains to be seen whether the intervention will be cost-effective in the long run. If so, it would provide evidence that routine population level genetic screening for variants that increase the risk for medically actionable conditions should be leveraged as an important strategy for implementing precision health at a population level, particularly as the cost of genome sequencing continues to decline. Other considerations for such programs, in addition to cost effectiveness, include the potential for false positive results, as well as privacy concerns.

A precision health approach to health care, however, should move beyond just looking at genetic testing; it should take into account individual variability not only in genes, but also in environment and lifestyle, as well as their interaction. This model has been adopted by a number of large-scale research projects, including the Human Project in New York City in the United States (Box 5.5). Public-private partnerships are essential for this approach. In the United States, the State of Nevada's "Healthy Nevada Project" aims to improve population health and personalized medicine by integrating clinical, genetic and environmental data with socioeconomic determinants to better understand the interplay between these factors; for this project, health care network Renown Health has partnered with the Desert Research Institute and Helix, a personal genomics company, with support from the State of Nevada (Renown Institute for Health Innovation, 2019[41]).

Ongoing large-scale, big data-driven population studies, like the United States 1-million participants "All of Us" study and the Danish "Harnessing the Power of Big Data to Address the Societal Challenge of Ageing" research project, also hold the potential to yield valuable insights, such as identifying which types of environments are more likely to facilitate healthier behaviours.

Box 5.5. The Human Project: an atlas for the human experience

Started in 2018, the **Human Project** will use big data analytics to aggregate and analyse a variety of measurements gathered over at least 20 years from 10 000 individuals in all five boroughs of New York City. The project aims to capture the dynamic interplay of biology, behaviour, and the environment, as well as their impact on health and disease.

In addition to undergoing physical examinations, participants will need to give researchers access to medical, financial and educational records, as well as cell phone data. In total, the project will extract and aggregate approximately 50 000 data points. Participants will also receive wearable activity trackers, special scales, and surveys via smart phones. Follow-up physical examinations will be requested every three years.

The project will thus drive a better understanding of the dynamic links between behavioural phenotypes, disease, and the broader environment, as well as how human health and behaviour co-evolve over the lifecycle, and will ultimately lead to new ways of improving health promotion and disease prevention.

Source: The Human Project (2019[42]), https://www.thehumanproject.org/about/; Peltz (2017[43]) 'Human Project' study will ask 10,000 to share life's data", https://www.apnews.com/12129cb7cab542248e83c9709e2ee7d0, Santora (2017[44]), "10,000 New Yorkers. 2 Decades. A Data Trove About 'Everything.'", https://www.nytimes.com/2017/06/04/nyregion/human-project-new-york-city-study.html.

5.4. Big data can be leveraged to implement more targeted public health interventions

5.4.1. Big data approaches can help translate knowledge to practice

The development of better population profiles offers the opportunity to develop new approaches to implementing prevention strategies. Using big data to better target public health initiatives may help to improve their effectiveness. One particular area in which big data can yield valuable insights is the translation of knowledge into effective public health policy. Translation arguably represents the next frontier in public health: how to distil the vast public health knowledge yielded by public health research studies into actionable steps to inform public health policy and decision-making.

The issue of "Much is known, but little is done" is at the forefront of many countries' public health policy discussions. The European Commission, for instance, recently formalised a new mechanism, called the Steering Group on Health Promotion, Disease Prevention and Management of Non-Communicable Diseases (the "Steering Group for Promotion and Prevention"), to facilitate the implementation of evidence-based best practices by EU countries to help prevent and manage non-communicable diseases (European Commission, n.d.[45]).

There are two main ways in which policy makers can leverage big data to address this issue. Firstly, big data can enable a faster, real-time monitoring of the impact of public health interventions. Changes in behaviour, public attitudes, public attention, or health status are often reflected in real-time online information sharing and communication patterns (Kim, Huang and Emery, 2016[19]). These data points can

give decision makers valuable feedback on the effectiveness of public health interventions and thus inform policy making.

Secondly, big data can facilitate a better understanding of the interaction between human behaviour/lifestyle ("behavioural phenotypes"), genetics, and the physical and social environment. Understanding which of these interactions have the greatest causal impact on public health can inform policy making – not only in public health, but also in other areas that influence health (e.g., various socioeconomic issues). Policy makers should therefore leverage big data to help move towards a Health in All Policies approach.

5.4.2. Many promising uses of big data for public health have emerged from the municipal level

In many cases, promising uses of big data for public health have emerged not from traditional public health actors, but, for instance, from initiatives to transform urban areas into "smart cities" that use data to improve urban planning, as well as policy making more generally.

As a result of accelerating urbanisation, cities face both the challenge and opportunity of being "first responders" to key global issues, including in public health, especially given many cities' status as hubs of data traffic and new technology applications. As centres for novel, big data-informed solutions and services in an increasingly decentralised world, "smart cities" have moved to the forefront of public health innovation.

Smart cities are "cities strongly founded on information and communication technologies that invest in human and social capital to improve the quality of life of their citizens by fostering economic growth, participatory governance, wise management of resources, sustainability, and efficient mobility, whilst they guarantee the privacy and security of the citizens." (Pérez-Martínez, Martínez-Ballesté and Solanas, 2013[46]). In the area of public health, smart cities can leverage their ability to develop and implement innovative, data-driven public health policies informed by big data analytics and move towards "smart health", without having to wait for national-level action – although such action can complement city-driven innovation by enhancing collaboration between cities, as discussed later in this section.

Smart health ("s-health") is "the provision of health services by using the context-aware network and sensing infrastructure of smart cities" (Solanas et al., 2014[47]). As this definition implies, ICT, big data and big data analytics are a key driver of smart health approaches.

The City of Chicago's food safety inspection, *E. coli* prediction on Lake Michigan beaches, and lead poisoning prevention programmes (Box 5.6), driven by predictive analytics and machine learning models, are examples of big data-driven, smart health solutions to public health problems. A partnership between the cities of Chicago and Las Vegas, Google, and Harvard University found that using location data and foodborne illness online searches predicted potentially unsafe restaurants better than traditional restaurant inspection methods and data-mining approaches (Sadilek et al., 2018[48]). What makes this example unique is that machine learning was used to improve accuracy, and that the linkage of these personal data also ensured that individuals remained unidentifiable.

The development and dissemination of such smart health solutions, however, has been slow. For instance, data gathered by sensors that measure environmental variables such as air pollution and airborne allergens could enhance the delivery of personalised prevention and care to asthma patients; a study that analysed data from wireless environmental sensor networks for air pollution measurement in eight cities across Europe, however, found that the performance of most of the sensors was unreliable, and they required frequent calibrations due to the interference of various environmental factors (Broday and Collaborators, 2017[49]).

Box 5.6. Chicago: Pioneering predictive analytic models for food protection and lead inspection programs

The Chicago Department of Public Health, as part of Chicago's Smart Data Project, has pioneered predictive analytics to identify the households with children most at risk for lead poisoning, as well as to more effectively monitor food establishments that are most likely to have food safety violations.

In 2014, Chicago's Department of Innovation and Technology used publicly available city data to build an algorithm to predict which restaurants were most likely to be in violation of health codes. The model aggregates data from a variety of sources (such as ZIP codes, business licenses, building code violations, and 311 complaints) to formulate a risk score, which allows inspectors to identify potential issues before they occur. The algorithm identified violations significantly earlier than business-as-usual did. Importantly, the team also made it easy for other cities to replicate Chicago's approach. More recently, the City of Chicago partnered with Google and Harvard University to test a novel machine learning-based approach that uses location data and foodborne illness online searches to predict potentially unsafe restaurants; this model is more effective than the original one (Sadilek et al., 2018[48]).

In collaboration with the University of Chicago, the Chicago Department of Public Health also developed a model that uses two decades of blood lead-level tests, home lead inspections, property value assessments, and census data to predict which households are most likely to have the greatest risk of causing lead poisoning in children. The model allows inspectors to prioritise house inspections and identify children who are at the highest risk.

An important component of Chicago's innovation strategy is liberating data: "making data accessible, discoverable, and usable by the public so that it can spur entrepreneurship, innovation, and discovery." (Choucair, Bhatt and Mansour, 2015[50]). As such, the code and data for these projects is publicly available on Chicago's "Open Source Projects" website.

The Clear Water collaborative, open source project further illustrates the data liberation approach: the City partnered with the Chicago Park District, volunteer data scientists, and local graduate students to build a better predictive model for forecasting beach water *E. coli* breakouts and help prevent infection. The model has tripled *E. coli* prediction rate on Lake Michigan beaches. The Clear Water code is publicly available and is written in R, an open source statistical programming language, which allows other scientists to test and potentially improve the current method.

Source: Choucair, B., J. Bhatt and R. Mansour (2015[50]), "A Bright Future: Innovation Transforming Public Health in Chicago", http://dx.doi.org/10.1097/PHH.0000000000000140; Sadilek et al., (2018[48]), "Machine-learned epidemiology: real-time detection of foodborne illness at scale", http://dx.doi.org/ 10.1038/s41746-018-0045-1; Spector (2016[51]), "Chicago Is Using Data to Predict Food Safety Violations. Why Aren't Other Cities?", https://www.citylab.com/solutions/2016/01/chicago-is-predicting-food-safety-violations-why-arent-other-cities/422511/; City of Chicago (2017[52]), "Clear Water", https://chicago.github.io/clear-water/; City of Chicago Developers (2017[53]), "Open source projects", http://dev.cityofchicago.org/projects/.

As noted in the Chicago case study, open data sharing represents a key driver of big data innovation and developing smart health approaches. Open data sharing helps connect smart city innovators with the relevant data. As data collection is often the most difficult part of researching and developing a solution to a particular problem, limiting data access can result in missed opportunities for "non-insiders" to develop potentially successful solutions. Open data also needs to be organized into databases with user-friendly search tools that allow easy data filtering (Smith, 2017[54]). It is also important to ensure that when data are made widely available, it is interpreted correctly by the wider set of users. Data privacy and security are, of course, particularly important when data is made publicly available.

While the cost and effort required for such projects may seem daunting, the benefits of enabling the development of innovative solutions driven by big data analytics can far outweigh these costs. Further, in the case of data crowdsourcing projects open data sharing can serve as an incentive for the public to participate in these initiatives. Examples of open data projects include La Base Adresse Nationale (France), Trafikverket (Sweden), and Data.gov (United States).

Inter-city collaboration can significantly speed up the rate of big data solutions in public health. One of the key issues of within-city innovation is that, typically, each city designs and implements its own good practices, with other cities finding out about them at a later stage, after they are successful (if ever). But co-ordination that can scale up successful solutions more effectively is needed; for instance, countries should organise partnerships between cities to facilitate tackling common issues together. Some promising examples are emerging: the Netherlands' "Smart City Strategy," for instance, aims to create a "Smart City collective" that will link cities, companies, and the research sector, functioning as a catalyst of knowledge sharing and change (Institute for Future of Living, 2017[55]).

At the same time, the important role of cities in testing and implementing "smart" approaches to public health runs the risk of exacerbating rural-urban health inequities. Policy makers should ensure rural areas are included in smart health programmes.

Another key aspect needed to advance towards smart city collaborations in the area of "smart health" are cross-sector and public-private partnerships. The "quadruple helix" collaboration between government, academia, industry, and citizens is essential for smart city and smart health innovations. Inter-sectoral collaboration is another key driver, particularly in public health, given the diversity of the causes of various diseases. As an example, Chicago's lead poisoning prevention programme involves collaboration between health care providers, lead inspectors, and housing providers, among others. Further, partnerships between cities (and other local governments) and the private sector should be explored, which can allow cities that do not yet use predictive analytics methods in-house, due to lack of resources or expertise, to contract them out.

5.5. Clear and consistent policies designed to safeguard private data are needed

Big data are increasingly allowing public health researchers and private companies to identify and link personal data across a variety of sources, many of which (e.g. smartphone data, credit card purchases, electronic medical records, GPS data) may contain sensitive health- and non-health related personal information. The implications of how this data could be used are considerable, and data protection policies that protect people from discrimination based on their health-related data is critical. While linking data that reflects health, genetic, and other personal information can provide valuable information about an individual's disease risk, it also poses the risk of uncovering potentially discriminatory personal health-related findings. (Salerno et al., 2017[56]) In addition, DNA-sequencing data can potentially be used to identify individuals by third parties.

Linking multiple data sources related to personal, socioeconomic, or other determinants of health without the individual's informed consent presents a particularly significant privacy risk, especially when the data linkage is not done for the specific purpose of answering a relevant research question or providing a clear public health benefit (Salerno et al., 2017[56]). Such data linkages, resulting in increased data dimensionality, can produce individual "data fingerprints", which can allow third parties to re-identify individuals in de-identified data sets through deductive disclosure techniques (Mooney and Pejaver, 2018[57]).

As such, wide-scale linkage of big data in public health needs to be accompanied by policies and regulations designed to safeguard privacy (e.g. sufficient de-identification of personal data), data security, confidentiality, and informed consent (Salerno et al., 2017[56]). In many cases, approaches to safeguarding privacy will require regulations that go beyond protecting health data alone and instead apply to the broader data landscape. Given

the quickly evolving nature of health data, data security and privacy risks are quickly changing, and best practices to ensure data privacy is safeguarded must be regularly assessed (OECD, 2015[58]).

5.6. Conclusion

Big data have the potential to enhance public health research, surveillance, and interventions to promote health and prevent disease, but are currently under-used. Applying big data for public health remains at a nascent stage, even when compared to other uses of data and digital technology in the health sector. However, new developments are likely to emerge in the coming years.

As emphasised throughout this report, advancing this area relies on good data governance. First, strong data governance is critical to ensure that security and privacy risks are managed – an important end as well as a key enabler of trust. Second, it promotes the development of policy and legal frameworks that enable secondary use of data in the first place. Third, it helps to maximise the utility and quality of available data (by harmonising data exchange formats enabling more data sources to be pooled, and promoting completeness of data) to generate knowledge for public health and other purposes.

Smart cities are key innovators of big data analytics solutions in public health. Inter-city collaborations can significantly speed up the development of big data-driven public health advancements. The role of cities in testing and implementing “smart” approaches to public health, however, can exacerbate rural-urban health inequities. Policy makers should ensure rural areas are included in smart health programmes. Policy makers should also leverage big data to move public health from a reactive to a predictive, precision health model.

Data transparency represents a key element that can facilitate the success of public health initiatives based on big data. Sharing data and algorithms with other stakeholders (e.g. collaborations between academia, public health departments, industry, and citizens) enables a more effective use of data and facilitates the early detection of any problems, and allows other public health authorities to implement similar successful interventions.

Despite the opportunities presented by big data, their inherent limitations and challenges suggest that their use should *complement* – not replace – traditional public health surveillance methods. Nevertheless, big data can enable a better understanding of the interaction between behaviour, genetics, and the physical and social environment. They should be put to work to generate and translate valuable knowledge into effective public health policy for better population health outcomes.

References

Andreu-Perez, J. et al. (2015), "Big Data for Health", *IEEE Journal of Biomedical and Health Informatics*, Vol. 19/4, pp. 1193 - 1208, http://dx.doi.org/10.1109/JBHI.2015.2450362. [20]

Balicer, R. et al. (2018), "Using big data for non-communicable disease surveillance", *The Lancet Diabetes & Endocrinology*, Vol. 6/8, pp. 595-598, http://dx.doi.org/10.1016/S2213-8587(17)30372-8. [24]

Bansal, S. et al. (2016), "Big Data for Infectious Disease Surveillance and Modeling", *The Journal of Infectious Diseases*, Vol. 214/Suppl 4, pp. S375–S379, http://dx.doi.org/10.1093/infdis/jiw400. [7]

Baumgartner, P. and N. Peiper (2017), "Utilizing Big Data and Twitter to Discover Emergent Online Communities of Cannabis Users", *Substance Abuse: Research and Treatment*, Vol. 11, pp. 1-9, http://dx.doi.org/10.1177/1178221817711425. [23]

Boyd, D. and K. Crawford (2012), "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon", *Information, Communication & Societ`y*, Vol. 15/5, pp. 662–679, http://dx.doi.org/10.1080/1369118X.2012.678878. [14]

Broday, D. and T. Collaborators (2017), "Wireless Distributed Environmental Sensor Networks for Air Pollution Measurement—The Promise and the Current Reality", *Sensors*, Vol. 17/10, p. 2263, http://dx.doi.org/10.3390/s17102263. [49]

Butler, D. (2013), "When Google got flu wrong", *Nature*, Vol. 494/7436, pp. 155-156, http://dx.doi.org/10.1038/494155a. [33]

Chen, P. and C. Zhang (2014), "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", *Information Sciences*, Vol. 275, pp. 314-347, http://dx.doi.org/10.1016/j.ins.2014.01.015. [11]

Choucair, B., J. Bhatt and R. Mansour (2015), *A bright future: Innovation transforming public health in Chicago*, http://dx.doi.org/10.1097/PHH.0000000000000140. [50]

City of Chicago (2017), *Clear Water*, https://chicago.github.io/clear-water/. [52]

City of Chicago Developers (2017), *Open source projects*, http://dev.cityofchicago.org/projects/. [53]

Cook, S. et al. (2011), "Assessing Google Flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic", *PLoS ONE*, Vol. 6/8, http://dx.doi.org/10.1371/journal.pone.0023610. [32]

De Mauro, A., M. Greco and M. Grimaldi (2015), "What is big data? A consensual definition and a review of key research topics", *AIP Conference Proceedings*, Vol. 1644/1, pp. 97-104, http://dx.doi.org/10.1063/1.4907823. [4]

DOMO (2018), *Data Never Sleeps 6.0*, https://www.domo.com/learn/data-never-sleeps-6. [3]

DOMO (2017), *Data Never Sleeps 5.0*, https://www.domo.com/learn/data-never-sleeps-5. [2]

Eichstaedt, J. et al. (2015), "Psychological Language on Twitter Predicts County-Level Heart Disease Mortality", *Psychological Science*, Vol. 26/2, pp. 159-169, http://dx.doi.org/10.1177/0956797614557867. [21]

European Commission (n.d.), *Steering Group on Health Promotion, Disease Prevention and Management of Non-Communicable Diseases*, https://ec.europa.eu/health/non_communicable_diseases/events/ev_20190607_fr. [45]

Eysenbach, G. (ed.) (2016), "Garbage in, garbage out: Data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection", *Journal of Medical Internet Research*, Vol. 18/2, p. e41, http://dx.doi.org/10.2196/jmir.4738. [19]

Eysenbach, G. (ed.) (2015), "Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study", *J Med Internet Res*, Vol. 17/7, p. e175, http://dx.doi.org/10.2196/jmir.4273. [26]

Eysenbach, G. (2011), "Infodemiology and Infoveillance: Tracking Online Health Information and Cyberbehavior for Public Health", *Am J Prev Med*, Vol. 40/5 Suppl 2, pp. S154-S158, http://dx.doi.org/10.1016/j.amepre.2011.02.006. [18]

Fuller, D., R. Buote and K. Stanley (2017), "A glossary for big data in population and public health: discussion and commentary on terminology and research methods", *J Epidemiol Community Health*, Vol. 71/11, pp. 1113-1117, http://dx.doi.org/10.1136/jech-2017-209608. [12]

Gelfand, A. (2019), *How Wearable and Implantable Technology is Changing the Future of Health Care*, Hopkins Bloomberg Public Health, https://magazine.jhsph.edu/2019/how-wearable-and-implantable-technology-changing-future-health-care. [28]

IBM (2017), *10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations*, https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN (accessed on 23 August 2018). [1]

IBM (n.d.), *Big Data Analytics*, https://www.ibm.com/analytics/hadoop/big-data-analytics (accessed on 26 August 2018). [6]

IBM (n.d.), *The Four V's of Big Data*, https://www.ibmbigdatahub.com/infographic/four-vs-big-data (accessed on 24 August 2018). [8]

Institute for Future of Living (2017), *NL Smart City Strategy*, https://instituteforfutureofliving.org/wp-content/uploads/NL_Smart_City_Strategie_EN_LR.pdf (accessed on 25 September 2018). [55]

Jain, S. et al. (2015), "The Digital Phenotype", *Nat Biotechnol.*, Vol. 33/5, pp. 462-463, http://dx.doi.org/10.1038/nbt.3223. [25]

Keyes, K. and S. Galea (2015), "What matters most: quantifying an epidemiology of consequence", *Annals of epidemiology*, Vol. 25/5, pp. 305–311, http://dx.doi.org/10.1016/j.annepidem.2015.01.016. [13]

Kosinski, M., D. Stillwell and T. Graepel (2013), "Private Traits and Attributes are Predictable from Digital Records of Human Behavior", *Proc Natl Acad Sci U S A*, Vol. 110/15, pp. 5802-5805, http://dx.doi.org/10.1073/pnas.1218772110. [22]

Lazer, D. and R. Kennedy (2015), *What We Can Learn From the Epic Failure of Google Flu Trends*, Wired, https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/. [34]

Lazer, D. et al. (2014), "The Parable of Google Flu: Traps in Big Data Analysis", *Science*, Vol. 343/6176, pp. 1203-1205, http://dx.doi.org/10.1126/science.1248506. [30]

Leinweber, D. (2007), "Stupid data miner tricks: overfitting the S&P 500", *The Journal of Investing*, Vol. 16/1, pp. 15–22, http://dx.doi.org/10.3905/joi.2007.681820. [15]

LexisNexis (2019), *LexisNexis Socioeconomic Health Scores*, https://www.lexisnexis.com/risk/downloads/literature/health-care/Socioeconomic-Health-Risk-Score-br.pdf (accessed on 5 August 2018). [17]

Marotta, S. (2019), *Ottawa-backed Carrot Rewards app shutting down after failing to find a buyer - The Globe and Mail*, https://www.theglobeandmail.com/business/article-ottawa-backed-carrot-rewards-app-shutting-down-after-failing-to-find-a/ (accessed on 21 October 2019). [16]

Miller, A. et al. (2018), "A Smartphone-Driven Thermometer Application for Real-Time Population- and Individual-Level Influenza Surveillance", *Clinical Infectious Diseases*, Vol. 67/3, pp. 388-397, http://dx.doi.org/10.1093/cid/ciy073. [27]

Mooney, S. and V. Pejaver (2018), "Big Data in Public Health: Terminology, Machine Learning, and Privacy", *Annual Review of Public Health*, Vol. 39, pp. 95-112, http://dx.doi.org/10.1146/annurev-publhealth-040617-014208. [57]

Mooney, S., D. Westreich and A. El-Sayed (2015), "Commentary: Epidemiology in the era of big data", *Epidemiology*, Vol. 26/3, pp. 390-394, http://dx.doi.org/10.1097/ede.0000000000000274. [10]

National Alliance Of Healthcare Purchaser Coalitions (2018), *Employers' perceptions & actions related to healthcare waste*. [36]

National Institutes of Health (2019), *All of Us Research Program*, https://allofus.nih.gov/ (accessed on 15 August 2018). [35]

OECD (2015), *Health Data Governance: Privacy, Monitoring and Research*, OECD Health Policy Studies, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264244566-en. [58]

Peltz, J. (2017), *'Human Project' study will ask 10,000 to share life's data*, https://www.apnews.com/12129cb7cab542248e83c9709e2ee7d0 (accessed on 21 October 2019). [43]

Pérez-Martínez, P., A. Martínez-Ballesté and A. Solanas (2013), *Privacy in Smart Cities: A case study of smart public parking*. [46]

Remais, J. (ed.) (2014), "Evaluation of Internet-Based Dengue Query Data: Google Dengue Trends", *PLOS Neglected Tropical Diseases*, Vol. 8/2, p. e2713, http://dx.doi.org/10.1371/journal.pntd.0002713. [31]

Renown Institute for Health Innovation (2019), *Healthy Nevada Project*, https://healthynv.org/ (accessed on 2 March 2019). [41]

Richards, C. et al. (2017), "Advances in Public Health Surveillance and Information Dissemination at the Centers for Disease Control and Prevention", *Public Health Reports*, Vol. 132/4, pp. 403-410, http://dx.doi.org/10.1177/0033354917709542. [9]

Sadilek, A. et al. (2018), "Machine-learned epidemiology: real-time detection of foodborne illness at scale", *npj Digital Medicinevolume*, Vol. 1/36, http://dx.doi.org/10.1038/s41746-018-0045-1. [48]

Salerno, J. et al. (2017), "Ethics, big data and computing in epidemiology and public health", *Annals of Epidemiology*, Vol. 27/5, pp. 297–301, http://dx.doi.org/10.1016/j.annepidem.2017.05.002. [56]

Santora, M. (2017), *10,000 New Yorkers. 2 Decades. A Data Trove About 'Everything.'*, The New York Times, https://www.nytimes.com/2017/06/04/nyregion/human-project-new-york-city-study.html (accessed on 21 October 2019). [44]

Simonsen, L. et al. (2016), "Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems", *The Journal of Infectious Diseases*, Vol. 214/Suppl 4, pp. S380-S385, http://dx.doi.org/10.1093/infdis/jiw376. [29]

Sivarajah, U. et al. (2017), "Critical analysis of Big Data challenges and analytical methods", *Journal of Business Research*, Vol. 70, pp. 263-286, http://dx.doi.org/10.1016/j.jbusres.2016.08.001. [5]

Smith, L. (2017), *Benefits of Open Data for Smart Cities*, https://hub.beesmart.city/solutions/benefits-of-open-data-for-smart-cities (accessed on 2 October 2018). [54]

Solanas, A. et al. (2014), "Smart Health: A Context-Aware Health Paradigm within Smart Cities", *IEEE Communications Magazine*, Vol. 52/8, pp. 74-81, http://dx.doi.org/10.1109/MCOM.2014.6871673. [47]

Spector, J. (2016), *Chicago Is Using Data to Predict Food Safety Violations. Why Aren't Other Cities? - CityLab*, Citylab, https://www.citylab.com/solutions/2016/01/chicago-is-predicting-food-safety-violations-why-arent-other-cities/422511/ (accessed on 4 November 2019). [51]

The Human Project (2019), *About The Human Project*, https://www.thehumanproject.org/about/ (accessed on 21 October 2019). [42]

The Office of the National Coordinator for Health Information Technology (2017), *Connecting Public Health Information Systems and Health Information Exchange Organizations: Lessons from the field*, http://www.healthit.gov/sites/default/files/ltpac_value_prop_factsheet_6-21-16.pdf. [37]

Trivedi, B. (2017), *Is health care ready for routine DNA screening? A massive new trial will find out*, http://www.sciencemag.org/news/2017/10/health-care-ready-routine-dna-screening-massive-new-trial-will-find-out (accessed on 24 August 2018). [40]

Willard, H., D. Feinberg and D. Ledbetter (2018), *How Geisinger Is Using Gene Screening to Prevent Disease*, https://hbr.org/2018/03/how-geisinger-is-using-gene-screening-to-prevent-disease (accessed on 23 August 2018). [38]

Williams, M. et al. (2018), "Patient-Centered Precision Health In A Learning Health Care System: Geisinger's Genomic Medicine Experience", *Health Affairs*, Vol. 5, pp. 757–764, http://dx.doi.org/10.1377/hlthaff.2017.1557. [39]

From:

Health in the 21st Century

Putting Data to Work for Stronger Health Systems

Access the complete publication at:
https://doi.org/10.1787/e3b23f8e-en

Please cite this chapter as:

OECD (2019), "Big data: A new dawn for public health?", in *Health in the 21st Century: Putting Data to Work for Stronger Health Systems*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/f24cb567-en