OPEN ACCESS

Glossary for public health surveillance in the age of data science

Arnaud Chiolero,[1,2,3,4] David Buckeridge[4]

[1]Population Health Laboratory (#PopHealthLab), Department of Community Health, University of Fribourg, Fribourg, Switzerland
[2]Institute of Primary Health Care (BIHAM), University of Bern, Bern, Switzerland
[3]Observatoire valaisan de la santé (OVS), Sion, Switzerland
[4]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada

Correspondence to
Prof Arnaud Chiolero, Population Health Laboratory (#PopHealthLab), Department of Community Health, University of Fribourg, 1700 Fribourg, Switzerland; achiolero@gmail.com

Received 16 November 2019
Revised 15 January 2020
Accepted 29 February 2020

Check for updates

© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Chiolero A, Buckeridge D. *J Epidemiol Community Health* 2020;**74**:612–616.

ABSTRACT

Public health surveillance is the ongoing systematic collection, analysis and interpretation of data, closely integrated with the timely dissemination of the resulting information to those responsible for preventing and controlling disease and injury. With the rapid development of data science, encompassing big data and artificial intelligence, and with the exponential growth of accessible and highly heterogeneous health-related data, from healthcare providers to user-generated online content, the field of surveillance and health monitoring is changing rapidly. It is, therefore, the right time for a short glossary of key terms in public health surveillance, with an emphasis on new data-science developments in the field.

PURPOSE OF THIS GLOSSARY

'Only describe, don't explain', attributed to Ludwig Wittgenstein

Public health surveillance is the ongoing systematic collection, analysis and interpretation of data, closely integrated with the timely dissemination of the resulting information to those responsible for preventing and controlling disease and injury.[1] It is a core element of public health practice, through routine monitoring and reporting systems, and of population health science—the science that informs public health and prevention strategies—through observational evidence.[2] More specifically, surveillance aims to provide health decision-makers with timely and useful information to set priorities, to identify the need for interventions and to evaluate the effects of interventions.[3] It is related to public health research but differs in its purposes (figure 1): research aims to increase general knowledge while surveillance aims to provide information for decision and action in public health.[1]

With, on the one hand, the rapid development of data science, encompassing big data and artificial intelligence (AI), and, on the other hand, the exponential growth of accessible and highly heterogeneous health-related data, from electronic medical records used by healthcare providers to user-generated online content,[4-6] the field of surveillance and health monitoring is changing rapidly with a widening scope of application, an increasing depth and new methods. It is, therefore, the right time for a glossary for public health surveillance and monitoring, with an emphasis on new data-science developments.[7] We do not aim to cover the whole field of surveillance but rather focus on how data science is changing methods and concepts, going from data generation and collection to information dissemination for decision-making (figure 2).

ABERRATION DETECTION

In public health, aberration detection is the identification of anomalous events or patterns in data, with a clinical or public health potential relevance, that is, statistical signals in surveillance data that may be of epidemiological importance.[8] A major challenge, of growing importance with the use of highly heterogenous types of surveillance data, is to account for random variability and measurement error, which makes it difficult to tease out the 'signal' upon which the decision to intervene is based from the 'background' noise.[9] Traditionally, outbreak detection and infectious disease surveillance have relied on reports from clinicians and laboratories. At the turn of the century, surveillance expanded to consider prediagnostic or syndromic data, such as the

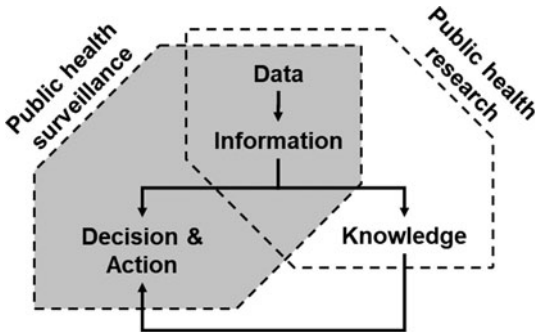

Figure 1 Health data and related information are used, on one hand, to increase general knowledge, which corresponds traditionally to a public health research activity. On the other hand, they are also key for guiding decisions and actions by stakeholders in public health, which corresponds to public health surveillance activities. The knowledge produced by research is eventually used to improve public health surveillance.

Figure 2 Steps in the data processing of public health surveillance, from data generation and collection to information dissemination for decision-making.

J Epidemiol Community Health: first published as 10.1136/jech-2018-211654 on 24 April 2020. Downloaded from http://jech.bmj.com/ on October 2, 2020 at Universite de Fribourg. Protected by copyright.

count of patients visiting an emergency room[5] (see also *Syndromic surveillance*). With the growth in volume and variety of accessible surveillance data, aberration detection methods have evolved from the analysis of time series of case counts to the complex modelling of individual-level surveillance cases with covariates drawn from multiple sources[5,8]; it is also applied beyond the field of human infectious diseases.

BIG DATA AND DATA SCIENCE

Big data refers to the massive amount of data that is more and more easily accessible through the digitalisation of all aspects of health, healthcare and related areas.[10] It is characterised by its variety, volume and velocity—the '3Vs'.[11] Multiple sources of data have become usable for public health surveillance, for example, mobile phones, online searches, social media, credit card transactions, wearable and ambient sensors, electronic health records (EHRs), medico-administrative records and pharmacy sales. While public health monitoring relies traditionally on well-defined and high-quality data, effective use of big data for surveillance requires new analytical methods such as data mining and data visualisation; data science is becoming mainstream in public health, integrating knowledge and skills from informatics and biostatistics. One major challenge in the analysis of big data is to account for the low quality, the poor data consistency across setting and time and the lack of metadata (see also *Source population and selectivity bias*). The questionable 'veracity' (the fourth 'V') of big data refers actually to its poor quality and high noise. Of critical importance is to go from big to 'smart' data, that is, data that can be transformed into information. While the development of big data and related data-science methods opens the way to data-informed or data-driven healthcare and public health,[12] it also raises major concerns about privacy protection (see also *Ethics of public health surveillance and privacy protection*). At the policy level, the use of big data for surveillance raises issues of access and benefit sharing, accountability and transparency and quality and safety.[13,14]

DATA, INFORMATION, KNOWLEDGE AND WISDOM PYRAMID

The data, information, knowledge and wisdom (DIKW) pyramid is a framework to help understand the hierarchal relationships from data to wisdom.[15] It has gained importance in public health monitoring, with the growing use of all types of data for surveillance activities, notably to highlight that data do not speak by itself and need to be transformed to become information, for example, in the form of health indicators,[16,17] with the latter having to be contextualised to become knowledge and eventually wisdom, for example, to inform health policy decisions[18] (see also figure 2). The DIKW pyramid also highlights that surveillance is not the mere collection and analysis of data, but a complex multilayer activity at the core of public health decision-making process, allowing evidence-informed policy-making[19] (see also *Evidence based and data-informed public health*). Recently, it has been proposed to review this pyramid, by deemphasising the notion of wisdom and by adding 'evidence' between information and knowledge (DIEK)[20]; evidence emerges through the comparison of information and is used to build actionable knowledge for public health.

DATA MINING

The discovery of patterns in large data sets by drawing on a range of methods from engineering, computer science and statistics is called data mining (see also *Big data and data science*). These methods are applied in an automated or semiautomated manner, usually with no a priori specification of the pattern to be detected. In a health monitoring context, some methods used for detecting aberrations or outbreaks can be considered data mining methods[5] (see also *Aberration detection*). Mining EHRs aims to gather information from unstructured narrative data[21] (see also *Electronic medical record*).

DATA VISUALISATION

Data visualisation has always been an important tool of public health surveillance. However, with the growth in available data and the improvement in statistical tools, data exploration through visualisation has gained importance for surveillance and monitoring activities. The field has evolved with contributions of computer science merging scientific visualisation, information visualisation and visual analytics, making visualisation an important part of surveillance data analyses[22]; it is a powerful tool to understand complex multilayer data, which are not easily captured by simple indicators. It has a major impact on how temporal and spatial analyses are conducted and reported. The production of continuously updated maps and atlas of diseases and risk factors has become possible by leveraging big data, thereby strengthening the surveillance of numerous conditions, notably of infectious diseases.[23] Visualisation of healthcare outputs through maps has also become a standard tool for health services research aiming to address unwarranted variation in healthcare.[24] Data visualisation is also gaining importance for displaying complex longitudinal data from EHRs[25] (see also *Electronic medical record*). One major change is the possibility of tailoring visualisation surveillance output to users' needs through interactive data visualisation.[22]

ETHICS OF PUBLIC HEALTH SURVEILLANCE AND PRIVACY PROTECTION

In 2017, the WHO issued international ethics guidelines on public health surveillance.[26,27] Surveillance activities raise ethical issues due to data collection methods, notably when the identity of individuals is recorded. More broadly, it is necessary to account for the balance between the protection of privacy and the benefits at a population level. With the development of surveillance based on the analyses of medicoadministrative,[6] social media or geospatial mobile phone data, and with growing linkage possibilities, individual privacy protection has become a major concern. The increasing sophistication and broadening possibilities for data linkage put at risk data management transparency and accountability.[13,14] The new European Union General Data Protection Regulation (GDPR) is the current legal framework for the collection of personal data in European countries[18]; it aims notably to give citizens more control over their own data and to harmonise data protection across Europe. The broad principles of GDPR include having a legitimate basis for data collection, purpose limitation, transparency, as much privacy and data minimisation as possible and accountability for all data use.[18]

ELECTRONIC MEDICAL (EMR) OR HEALTH RECORD (EHR) AND PERSONAL HEALTH RECORD (PHR)

The increasing adoption of electronic records to manage medical and health data creates new opportunities for public health monitoring.[28] An electronic medical record (EMR) is used to integrate, manage and analyse patient data collected in a clinical context, often within one clinic or institution. An EHR is intended to have a broader scope, encompassing all health-related data over the life course. A related concept is a personal health record (PHR), which is an EHR controlled by a patient. In all cases, these records are useful for population monitoring to the extent that they record concepts and health events in a consistent and unambiguous manner (eg, through the use of data standards and ontology[29]), which enables different systems to exchange data, or interoperate[30] (See also *Interoperability*). Major challenges remain such as how to define the denominators for events extracted from EHR.[31]

J Epidemiol Community Health: first published as 10.1136/jech-2018-211654 on 24 April 2020. Downloaded from http://jech.bmj.com/ on October 2, 2020 at Universite de Fribourg. Protected by copyright.

EVIDENCE-BASED AND DATA-INFORMED PUBLIC HEALTH

At the crossroad between population health science[2] and applied public health research, public health surveillance is a core element of evidence-based public health (figure 3).[32] Indeed, population assessment, production of indicators and reports and evaluations are typical activities and outcomes of public health surveillance. Monitoring the literature is also an integral part of surveillance, for example, to allow comparison and benchmarking or to challenge measurement and definition of indicators. In the age of data science, the management of surveillance data and information has gained importance in the evidence-based public heath cycle, with the policy-making process becoming not only evidence based but also data informed if not data driven. Evidence-based public health should also guide how surveillance system is designed[33] (see also *Population health record*).

FORECASTING

Data collected through surveillance are often analysed to identify important changes in population health. Inference about change requires an estimate of the expected state of population health, which is obtained through forecasting, or predicting future population health status using data collected in the past. Many methods are available for forecasting, from a simple average of historical values to multivariate time-series methods.[34] Forecasting of expected values is a critical step in routine surveillance for outbreaks and is also used to estimate the future burden from chronic diseases and other prevalent conditions. The accuracy of a forecast usually decreases as the length of the horizon increases and is usually evaluated by comparing forecasts to actual values once data become available. Because the performance of predictive models depends on the quality and stability (across eg, time and space) of data, forecasting methods must adapt to the relatively low quality and selectivity of big data (see also *Source population and selectivity bias*).

INTEROPERABILITY

Increasingly, public health surveillance draws data from a wide range of sources and makes information available to many stakeholders. This acquisition of data and dissemination of information has traditionally been a manual process, but as volumes continue to grow, automation of data and information exchange becomes necessary. Such automation requires the definition and adoption of standards that indicate clearly how information systems should interact with one another or interoperate. The term semantic interoperability is used to define the ability for one information system to receive data from another system and to reliably process this data to produce information.[35] For example, messaging standards such as Health Level Seven and Fast healthcare Interoperability Resource allow public health surveillance systems to interoperate with laboratory systems and information exchange standards such as Statistical Data and Metadata Exchange allow public health systems to interoperate with web-based systems to automate the dissemination of population-based indicators.

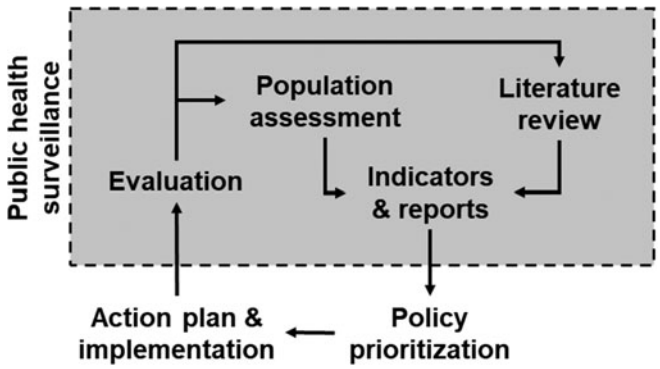

Figure 3 Public health surveillance is a central element of evidence-based public health. Inspired by Brownson *et al* 2009.[32]

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE

AI can be defined in terms of human intelligence, such that any machine that can act like a human is displaying AI.[36] The ability of a machine to perform any intellectual task is called Artificial General Intelligence or Strong AI and is thought to require a range of skills, such as natural language processing, knowledge representation, automated reasoning, machine learning, computer vision and robotics. Each of these skills is the subject of considerable research in AI, employing different connectionist (ie, data driven) or symbolic (ie, using logic and symbols) approaches. Recent algorithmic advances have enabled profound gains in the performance of neural networks for machine learning.[37] In epidemiology and public health surveillance, machine learning is used as one tool to execute causal inference analysis, diagnosis and prognosis studies, genome-wide association studies, geospatial applications or forecasting.[38] Such machine learning methods also have the potential to advance aberration detection.[5]

POPULATION HEALTH RECORD

The International Organization for Standardization (ISO) has defined a population health record (PopHR) as a system analogous to an EHR but containing aggregated and usually deidentified data for public health and other epidemiological purposes.[39] The concept of the PopHR was subsequently developed further, noting that its primary purpose is to support efficient and effective public health practice, that it should be based on an explicit population health framework and that it should make available indicators that document the current status and influences of the health of a defined population.[40] While PopHR systems have yet to be adopted widely in public health practice, researchers have developed and implemented demonstration systems,[33] along with formal ontologies to support information integration in a PopHR.[41]

PRECISION PUBLIC HEALTH

Precision public health is inspired by precision medicine with the idea that a better use of all types of data, encompassing geography, physical and sociodemographic characteristics, as well as health behaviours and biomarkers, at a local or community scale, would help design specific public health policy for a given population, and be more effective than general policy.[42,43] Some have argued that the term is problematic, causing confusion with the precision medicine movement and focusing attention on individual diagnosis and treatment.[44,45] Others have suggested that precision public health merely rebrands modern public health surveillance activities and adds little value.[45]

SECONDARY USE OF DATA

Surveillance activities are relying increasingly on the use of data not specifically collected for that purpose, including data a priori not related to health.[46,47] The secondary use of data is not new in surveillance, but it has grown in importance and depth, leading to a paradigm shift in surveillance. Indeed, the classical approach is (1) to define or choose the health problem for which surveillance is necessary, (2) to define and collect the data needed and (3) to analyse data to address your problem. Along this approach, 'designed data' specifically tailored to address surveillance goals

J Epidemiol Community Health: first published as 10.1136/jech-2018-211654 on 24 April 2020. Downloaded from http://jech.bmj.com/ on October 2, 2020 at Universite de Fribourg. Protected by copyright.

are used. The more contemporary data-driven approach is (1) to collect data from multiple source without knowing a priori what will be done with this data and (2) to analyse data to see if they could help solve surveillance problems. With this approach, 'organic data' not specifically tailored for surveillance are used (see also *Big data*).[48] Designed and organic data have specific advantages and disadvantages. On the one hand, validity and reliability of designed data are often documented. Further, designed data collection processes are defined and the ethical and legal frameworks for collection are explicit; the lack of such clear frameworks for organic data is a major current issue (see also *Ethics of public health surveillance and privacy protection*). On the other hand, resources needed to collect designed data are larger than for organic data. Also, the reporting delay can be shorter with organic data compared with designed data. However, the source population of organic data can be tricky to identify (see also *Source population and selectivity bias*).[31]

SOURCE POPULATION AND SELECTIVITY BIAS

Public health surveillance aims to gather information on the health-related characteristics of a specific population, which most often is a group of people living in a given location. More broadly, a population is a group of people sharing a characteristic, such as a medical condition or treated in specific healthcare facilities.[2,49] With some types of big data, one difficulty is to define the source population from which this data have emerged; completeness or representativeness of the supposedly source population cannot be ensured due to the non-probabilistic character of this data, resulting from the selectivity of people from which data are recorded.[50,51] Routinely collected data are often event based rather than population based, with no information on the individuals who did not experience the event,[46] and the link between the event and the individual can be difficult to establish. Further, the source population can change very rapidly, for example, for sales, online and any other user-generated data, and in an unpredictable manner. As a result, denominators cannot be easily computed, and inference beyond the study population is problematic, due to a selectivity bias (see also *Secondary use of data*). Selectivity bias is a term used to highlight the challenge of identifying and defining the source population *per se* of big data; it differs from selection bias which refers usually to a sampling issue, making the data used for the analysis problematic for inference to the source or target population.

SURVEILLANCE BIAS

Many conditions and health-related events under surveillance are sensitive to the modality and intensity of detection activities, for example, several types of cancer, thromboembolism or postoperative infections.[52,53] Surveillance bias occurs when such conditions are sought with differential intensity across populations or over time, or according to care setting and patient characteristics.[54,55] As a result, the difference in the frequency (incidence, prevalent) of the condition may not reflect a change in the risk of this condition, but instead a difference in the frequency of detection. For instance, between-hospital differences in the frequency of thromboembolism following hip surgery can reflect between-hospital differences in postsurgery screening activities (large number of cases identified in hospitals with intense screening activities vs low number in other hospitals), rather than any difference in the quality of care.[55] A related concept is the 'streetlight effect' which occurs when surveillance activities are not concentrated on what matters, but on what is measurable, even if it is not relevant.

SYNDROMIC SURVEILLANCE

Case definitions based on syndromes can enhance the sensitivity and timeliness of surveillance. Around the turn of the millennium, surveillance of syndromes was implemented on a large scale by applying automated algorithms to clinical data.[56] The automated detection of syndromes in clinical data and by automated statistical analysis to detect aberrations in the frequency of syndromes are defining characteristics of syndromic surveillance[57] (see also *Aberration detection*). Although an early motivation for syndromic surveillance was rapid outbreak detection, the use of non-specific, prediagnostic data can make it challenging to detect a signal quickly with an acceptable rate of false alerts.[58] Nonetheless, due to their potential to provide real-time information about population health, syndromic surveillance systems routinely contribute to situational awareness in many public health systems and are often deployed for mass gathering events.

CONCLUSION

Data-science and newly accessible data are driving innovation in methods for public health surveillance and monitoring, offering new opportunities. However, disappointment is also to be expected due to the challenge in extracting value from healthcare data which often lack consistent structure and clear meaning.[59]

Fostering the ability of primary data providers to improve the structure and semantics of the data they collect can make it easier to obtain meaningful information and, eventually, knowledge from these data. Stronger semantic interoperability between health information systems[35] and more consistent data structure will be essential to help moving from big to smart data, that is, data that can be used to produce information, and to transform health systems which are currently data rich but information poor into systems which are data and information rich.[60]

Finally, while many resources are directed towards data collection and processing, the resources and expertise needed to make these data truly useful for surveillance, namely background knowledge on public health and on the processes generating the data,[6] are critical more than ever in an age of data science; knowledge brokers are needed to bridge data science, health monitoring and public health.

Contributors AC and BD both drafted the paper and reviewed it before submission.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iD
Arnaud Chiolero http://orcid.org/0000-0002-5544-8510

REFERENCES

1 Thacker SB, Berkelman RL. Public health surveillance in the United States. *Epidemiol Rev* 1988;10:164–90.
2 Keyes KM, Galea S. *Population health science*. Oxford University Press, New York, 2016.
3 Nsubuga P, White ME, Thacker SB, *et al*. Public health surveillance: a tool for targeting and monitoring interventions. In: Jamison DT, Breman JG, Measham AR, *et al*. eds. 2nd ed. *Disease control priorities in developing countries*. Washington (DC): The International Bank for Reconstruction and Development / The World Bank; New York: Oxford University Press;, 2006.

J Epidemiol Community Health: first published as 10.1136/jech-2018-211654 on 24 April 2020. Downloaded from http://jech.bmj.com/ on October 2, 2020 at Universite de Fribourg. Protected by copyright.

4 Lee LM, Thacker SB. Public health surveillance and knowing about health in the context of growing sources of health data. *Am J Prev Med* 2011;41:636–40.
5 Yuan M, Boston-Fisher N, Luo Y, *et al.* A systematic review of aberration detection algorithms used in public health surveillance. *J Biomed Inform* 2019;94:103181.
6 Sarrazin MS, Rosenthal GE. Finding pure and simple truths with administrative data. *JAMA* 2012;307:1433–5.
7 Groseclose SL, Buckeridge DL. Public health surveillance systems: recent advances in their use and evaluation. *Annu Rev Public Health* 2017;38:57–79.
8 Faverjon C, Berezowski J. Choosing the best algorithm for event detection based on the intended application: a conceptual framework for syndromic surveillance. *J Biomed Inform* 2018;85:126–35.
9 Chiolero A, Anker D. Screening interval: a public health blind spot. *Lancet Pub Health* 2019;4:e171–2.
10 Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA* 2013;309:1351–2.
11 Mooney SJ, Westreich DJ, El-Sayed AMJE. Epidemiology in the era of big data. Epidemiology 2015;26:390.
12 Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* 2014;2:3.
13 Vayena E, Dzenowagis J, Brownstein JS, *et al.* Policy implications of big data in the health sector. *Bull World Health Organ* 2018;96:66–8.
14 Vayena E, Haeusermann T, Adjekum A, *et al.* Digital health: meeting the ethical and policy challenges. *Swiss Med Wkly* 2018;148:w14571.
15 Rowley J. The wisdom hierarchy: representations of the DIKW hierarchy. *J Inf Sci* 2007;33:163–80.
16 Etches V, Frank J, Di Ruggiero E, *et al.* Measuring population health: a review of indicators. *Annu Rev Public Health* 2006;27:29–55.
17 Chiolero A, Paccaud F, Fornerod L. [How to conduct public health surveillance? The example of the Observatoire Valaisan de la Sante in Switzerland]. *Sante Publique* 2014;26:75–84.
18 Verschuuren M, Van Oers H. *Population health monitoring: climbing the information pyramid*. Switzerland, Springer, 2018.
19 Oxman AD, Lavis JN, Lewin S, *et al.* SUPPORT tools for evidence-informed health policymaking (STP) 1: what is evidence-informed policymaking? *Health Res Policy Syst* 2009;7:S1.
20 Dammann O. Data, information, evidence, and knowledge:: a proposal for health informatics and data science. *Online J Public Health Inform* 2018;10:e224.
21 Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012;13:395–405.
22 O'Donoghue SI, Baldi BF, Clark SJ, *et al.* Visualization of biomedical data. *JAMIA Open* 2018;1:275–304.
23 Hay SI, George DB, Moyes CL, *et al.* Big data opportunities for global infectious disease surveillance. *PLoS Med* 2013;10:e1001413.
24 Birkmeyer JD, Reames BN, McCulloch P, *et al.* Understanding of regional variation in the use of surgery. *Lancet* 2013;382:1121–9.
25 West VL, Borland D, Hammond WE. Innovative information visualization of electronic health record data: a systematic review. *JAMA* 2015;22:330–9.
26 World Health Organization. *WHO guidelines on ethical issues in public health surveillance*. Geneva: World Health Organization 2017.
27 Fairchild AL, Haghdoost AA, Bayer R, *et al.* Ethics of public health surveillance: new guidelines. *Lancet Pub Health* 2017;2:e348–9-.
28 Klompas M, McVetta J, Lazarus R, *et al.* Integrating clinical practice and public health surveillance using electronic medical record systems. *Am J Public Health* 2012;102: S325–32.
29 Gonzalez C, Blobel BG, Lopez DM. Ontology-based framework for electronic health records interoperability. *Stud Health Technol Inform* 2011;169:694–8.
30 Moreno Conde A. *Quality framework for semantic interoperability in health informatics: definition and implementation*. London, UK: UCL (University College London), 2016.
31 Cocoros NM, Ochoa A, Eberhardt K, *et al.* Denominators matter: understanding medical encounter frequency and its impact on surveillance estimates using EHR data. *EGEMS* 2019;7:31.
32 Brownson RC, Fielding JE, Maylahn CM. Evidence-based public health: a fundamental concept for public health practice. *Annu Rev Public Health* 2009;30:175–201.
33 Shaban-Nejad A, Lavigne M, Okhmatovskaia A, *et al.* PopHR: a knowledge-based platform to support integration, analysis, and visualization of population health data. *Ann N Y Acad Sci* 2017;1387:44–53.
34 Burkom HS, Murphy SP, Shmueli G. Automated time series forecasting for biosurveillance. *Stat Med* 2007;26:4202–18.
35 Dixon BE, Vreeman DJ, Grannis SJ. The long road to semantic interoperability in support of public health: experiences from two states. *J Biomed Inform* 2014;49:3–8.
36 Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. *NPJ Digit Med* 2019;2:77.
37 Ardila D, Kiraly AP, Bharadwaj S, *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med* 2019;25:954–61.
38 Bi Q, Goodman KE, Kaminsky J, *et al.* What is machine learning? A primer for the epidemiologist. *Am J Epidemiol* 2019;188:2222–39.
39 ISO/TR 20514. *Health informatics - electronic health record - definition, scope, and context*. 2005.
40 Friedman DJ, Parrish RG 2nd. The population health record: concepts, definition, design, and implementation. *JAMA* 2010;17:359–66.
41 Shaban-Nejad A, Okhmatovskaia A, Izadi MT, *et al.* PHIO: a knowledge base for interpretation and calculation of public health indicators. *Stud Health Technol Inform* 2013;192:1207.
42 Desmond-Hellmann S. Progress lies in precision. *Science* 2016;353:731.
43 Dowell SF, Blazes D, Desmond-Hellmann S. Four steps to precision public health. *Nature* 2016;540:189–91.
44 Seeking precision in public health. *Nat Med* 2019;25:1177.
45 Chowkwanyun M, Bayer R, Galea S. "Precision" public health - between novelty and hype. *New Engl J Med* 2018;379:1398–400.
46 Jorm L. Routinely collected data as a strategic resource for research: priorities for methods and workforce. *Public Health Res Pr* 2015;25:e2541540.
47 Benchimol EI, Smeeth L, Guttmann A, *et al.* The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLoS Med* 2015;12:e1001885.
48 Ann Keller S, Koonin SE, Shipp SJS. Big data and city living: what can it do for us? Significance 2012;9:4–7.
49 Keyes KM, Galea S. Setting the agenda for a new discipline: population health science. *Am J Public Health* 2016;106:633–4.
50 Buelens B, Daas P, Burger J, *et al. Selectivity of big data: statistics Netherlands*. The Netherlands: The Hague/Heerlen, 2014.
51 Beresewicz M, Lehtonen RT, Reis F, *et al.* An overview of methods for treating selectivity in big data sources. *Eurostat.* Luxembourg: Publications Office of the European Union, 2018.
52 Welch HG, Brawley OW. Scrutiny-dependent cancer and self-fulfilling risk factors. *Ann Intern Med*. 2018;168:143–5.
53 Welch HG, Kramer BS, Black WC. Epidemiologic signatures in cancer. *New Engl J Med* 2019;381:1378–86.
54 Chiolero A, Santschi V, Paccaud F. Public health surveillance with electronic medical records: at risk of surveillance bias and overdiagnosis. *Eur J Public Health* 2013;23:350–1.
55 Haut ER, Pronovost PJ. Surveillance bias in outcomes reporting. *JAMA* 2011;305:2462–3.
56 Mandl KD, Overhage JM, Wagner MM, *et al.* Implementing syndromic surveillance: a practical guide informed by the early experience. *J Am Med Inform Assn* 2004;11:141–50.
57 Soler MS, Fouillet A, Viso AC, *et al.* Assessment of syndromic surveillance in Europe. *Lancet* 2011;378:1833–4.
58 Buckeridge DL. Outbreak detection through automated surveillance: a review of the determinants of detection. *J Biomed Inform* 2007;40:370–9.
59 Greene JA, Lea AS. Digital futures past - The long arc of big data in medicine. *New Engl J Med* 2019;381:480–5.
60 OCDE (2019), Health in the 21st Century : Putting Data to Work for Stronger Health Systems, OECD Health Policy Studies, Éditions OCDE, Paris, https://doi.org/10.1787/e3b23f8e-en.

J Epidemiol Community Health: first published as 10.1136/jech-2018-211654 on 24 April 2020. Downloaded from http://jech.bmj.com/ on October 2, 2020 at Universite de Fribourg. Protected by copyright.