European Journal of Epidemiology (2018) 33:245–257
https://doi.org/10.1007/s10654-018-0385-9

ESSAY

Epidemiology in wonderland: Big Data and precision medicine

Rodolfo Saracci[1]

Received: 20 March 2018 / Accepted: 30 March 2018 / Published online: 5 April 2018

Abstract
Big Data and precision medicine, two major contemporary challenges for epidemiology, are critically examined from two different angles. In Part 1 Big Data collected for research purposes (Big research Data) and Big Data used for research although collected for other primary purposes (Big secondary Data) are discussed in the light of the fundamental common requirement of data validity, prevailing over "bigness". Precision medicine is treated developing the key point that high relative risks are as a rule required to make a variable or combination of variables suitable for prediction of disease occurrence, outcome or response to treatment; the commercial proliferation of allegedly predictive tests of unknown or poor validity is commented. Part 2 proposes a "wise epidemiology" approach to: (a) choosing in a context imprinted by Big Data and precision medicine—epidemiological research projects actually relevant to population health, (b) training epidemiologists, (c) investigating the impact on clinical practices and doctor-patient relation of the influx of Big Data and computerized medicine and (d) clarifying whether today "health" may be redefined—as some maintain in purely technological terms.

Keywords Big data · Datome · Doctor-patient relation · Epidemiological research · Epidemiology training · Health definition · Population health · Precision for commerce · Precision medicine · Validity · Wise epidemiology

Part 1

A wonderland of science and technology

I entered medical school in 1954, a year after the epoch-making paper in Nature by Watson and Crick on the double helix structure of DNA [1] had ignited a profound and long-lasting revolution across all biology and medicine, continuing today with the "omics". After graduating with a thesis on blood groups genetics I began in academic internal medicine: clinical trials were a focus of interest and to train in trial methodology I joined in 1964 the MRC Statistical Research Unit, sited at the 115 Gower Street in London and directed by dr. (at that time) Richard Doll. An exciting surprise was waiting for me in the basement office of the MRC Unit: epidemiology, of which I knew nothing.

My first exercise in epidemiology was a cohort study of cancer in pernicious anaemia patients [2]. The diagnostic data were manually abstracted from clinical records. The vital status ascertainment was done by a combination of follow-up contacts, writing letters to patients (or failing that, relatives or neighbours) and consultation of cumbersome nominal registers of deaths at the central statistical office: causes of death were then coded after obtaining by post paper copies of individual death certificates. Last but not least the entire data analysis, starting by computing person-years for each sex-age-calendar year group, was manually done using an electro-mechanical calculating machine whose maximum capability was to carry sums of squares and products. To control for errors the whole procedure had to be repeated at least once.

The study results showed a marked increase in mortality for pernicious anaemia and stomach cancer and a slight increase in risk of myeloid leukaemia. That particular cohort was not large, 1625 patients, all traced, but the same technical procedures and human operator requirements applied to large studies conducted at the MRC Unit, such as the prototype cohort study of 40,637 British doctors (34,445 males and 6192 females) on the health effects of tobacco smoking, of which the 10 years follow-up results

Rodolfo Saracci—Former President, International Epidemiological Association, Lyon, France.

✉ Rodolfo Saracci
saracci@hotmail.com

[1] 7 rue St.Hippolyte, 69008 Lyon, France

were published just in 1964 [3]. Elsewhere studies now regarded as classic like Framingham's [4] or the "Seven countries" [5] were at the time under way, involving more complex designs and logistic: but they too had available similar technical procedures, time consuming and heavily demanding in personnel.

One front however was moving fast ahead, computing. The need of writing one's own programs for statistical data analysis rapidly disappeared when around the mid 1960's general use software became available for the mainframe computers of the time. Even a desktop computer was launched in 1965, the Olivetti Programma 101, now exhibited for its aesthetic industrial design at the MoMA in New York [6], for which programs, including for epidemiological data analyses, were written on magnetic cards. Over the subsequent half a century the advance in computer science, information technology (IT) and applications has been vertiginuous: processing capability has escalated from one or few FLOPS (floating point operations per second) with a mechanical machine to the near 20 TeraFLOPS (10^{12}) of the most recent iMac and near 50 PetaFLOPS (10^{15}) of the most powerful machines. Data storage capacity has similarly expanded, as witnessed by the difference between two familiar memory supports, a floppy disk that could store less than a MB (10^6 bytes) and a USB flash drive capable of up to 1 TB (10^{12} bytes).

Data storage and processing capabilities have been the key enabling factors of major changes in the analysis of epidemiological data. Different methods to select variables in multiple regressions became implementable in early stages of the computing development while procedures based on extensive re-sampling from empirical or assumed data distributions in Monte Carlo simulations have become current only in the last decade or two. For their general usefulness at least three, not existent in actual practice when I started, appear now in standard textbooks [7]: multiple imputation methods for missing data, a practically inevitable and relevant hurdle as soon as one starts to analyse any large data set; uncertainty analysis (probabilistic sensitivity analysis) allowing to give quantitative expression to biases of various origins, the key thorny problem in observational studies; and the general bootstrapping approach to calculate confidence limits for any kind of statistical estimate, however complicate the estimating function may be, or to internally validate a model. Ever advancing data storage capacity and processing speed are now at the core of the Big Data "deluge" [8] or "explosion" [9]. Compared to the mid 1960s the present, at the eve of a "fourth industrial revolution" [10], technology-driven in massive data processing as in frontline biology and medicine, shines as a wonderland with prospects of lasting progress. On closer scrutiny this techno-scientific wonder world raises a variety of questions of scientific nature, with social and philosophical ramifications.

Validity first

Early characterizations of Big Data included three attributes: *Volume, Variety*, *Velocity.* A fourth fundamental V was missing, *Validity,* now usually added alongside other V's as *Variability* and *Veracity*. Validity was the principle that Paul Meier (at the origin of the Kaplan–Meier method) kept stressing within an early 1970s committee of the International Biometric Society (IBS) we were both members of. The committee was set up on the initiative of Thomas Chalmers, Associate Director of the USA National Institutes of Health and of Peter Armitage, President of the IBS, to review the biometric aspects, in fact all methods, procedures, protocols and documents of the controversial University Group Diabetes Program trial of oral hypoglycaemic drugs [11]: Paul kept our work on track by repeating "validity first", namely "let us focus on investigating sources of bias, if any" and put aside all the rest for secondary consideration.

In a "validity first" perspective data are worth not primarily for their "bigness" but for their measurement validity, the foundation without which internal and external validity of studies cannot be built. Measurement validity depends both on the intrinsic measurement method validity and on the way the method is actually applied in a specific population in the concrete circumstances of a study: moderately sized data sets with negligible systematic and restraint random (from uncontrollable sources) errors of measurement may turn out more valuable than large sets with sizable errors of both kinds. Inflating sample size may to some extent compensate for poor measurements with known error structure, as these can be taken into account in the data analysis model, but can be hardly a remedy for poor measurements with ill-defined error structure in models involving a multiplicity of variables. For this basic reason the "validity first" principle guides all collections, small or big, of data for research purposes, in which methods with maximal and quantitatively known validity are as a rule employed.

Big research data

The EPIC study (European Prospective Investigation into Cancer and Nutrition) initiated by Elio Riboli in my Unit of Analytical Epidemiology at the International Agency for Research on Cancer (IARC) in Lyon [12, 13] offers a clear example of the high priority granted to measurement validity in a Big Data collection for research purposes. In the early phase of the study a major investment was made in designing and testing for validity all methods of

measurements, from questionaries to anthropometric variables to a brand-new system for blood specimens collection and long-term storage, suitable to large scale use. The system has proved effective to maintain with no deterioration for more than 20 years in the IARC central biobank 3.8 million blood aliquots from 276,000 subjects. Similar investments in developing validated methods of measurement, including some as sophisticated as imaging techniques, have been made within another internationally accessible Big research Data resource for prospective investigations, the UK Biobank [14].

Measurement validity in its various facets, development, quantitative evaluation, improvement, comparison and alignment between studies to be combined, features as a central issue in a variety of epidemiological investigations involving Big research Data, such as:

(a) Gene-wide association studies (GWAS) and metabolomics studies. The immediately obvious issue of distinguishing signals from noise in the massive data of such studies has expanded into broader issues of measurement validity as soon as the accrual of actual measurements has shown that they were not error-free. Systematic programmes of quality assurance are now available to put measurement validity in GWAS on a firm footing [15] and similar protocols, made more complex by the great variety of molecules involved, are being developed for metabolomics [16];
(b) The European Cohort Consortium, a recently started partnership of 40 cohorts ranging in size from 10,000 up to 520,00 individuals and with a total potential sample size of over 2.5 million individuals destined to research on chronic diseases [17];
(c) The 'Exposomics' project aimed at developing a novel approach to exposure assessment chiefly to air pollution and water contaminants using a combination of personal exposure monitoring (PEM) with portable devices and "omic" measurement technologies [18];
(d) The MR-Base Collaboration that uses published genetic association from GWAS investigations to infer causal relationship between phenotypes (e.g. triglycerides and coronary heart disease) using the two steps mendelian randomization method [19].

In these as in many other research projects Big Data introduce, besides IT issues of data management, a novel aspect, the validity of Big Data bioinformatics software. A recent review [20], focused in particular on software dealing with next-generation sequencing (NGS) data, states "another very important, yet largely ignored, area of Big Data bioinformatics is the problem of software validation" in the light of studies indicating that the results produced by different bioinformatics pipelines based on the same sequencing data can differ substantially.

Big secondary data

Data generated for a purpose different from the research activities in which they are used are often generically designated as "secondary data" [21]. They are produced by a great variety of sources and in many presentations and discussions on Big Data they are completely confused with data collected for research purposes. Secondary data are and have been advantageously used at least since 1662, when John Graunt analysed London's "Bills of mortality" [22], in all types of demographic and epidemiological studies either as the sole material or together with data collected "ad hoc" for a research project. Big Data greatly expand these secondary databases, often in digital form, and the scope for research. For example in environmental epidemiology exposure data have been most often available from fixed central monitors in relatively large area units for which aggregated health outcomes data are provided by administrative records. Today data of high-resolution (in space and time) measurements of environmental exposures are acquired from remote sensors in satellites or by modelling data of emissions and dispersion of pollutants [23]: they can be linked to EHR (electronic health record) databases with information on health outcomes and possibly other individual variables (life style habits, occupation, previous diseases etc.) on large populations. Higher resolution and better control of confounding enhance the ability of validly detecting and quantifying risks caused by environmental pollutants. In a similar way the scope and validity of social epidemiology studies is improved by high resolution measurements of socio-economic variables (unemployment proportion, deprivation index) by small geographical units, typically census tracts [24].

Whatever the type and size of a big database the "validity first" principle applies. It entails that before jumping to any data analysis procedures of quality check and data editing, well established in epidemiology, need to be completed [25]. They may prove challenging when the procedures by which the data have been acquired and stored are not traceable. Large volumes of biological samples from clinical activities are today stored in repositories at hospitals and public and private health centers: before analyses these secondary samples must be submitted to the same quality control protocols already mentioned for research samples. More generally attention has to be given to how well a repository satisfies technical, ethical and legal requiremen [26]. Exploratory Data Analysis (EDA) is another area familiar to epidemiologists, who have been often searching for associations to be later tested in analytical investigations by correlating, for example, incidence

of cancers with nutrients consumption by geographical areas [27] or with a large number of occupations [28]. The general statistical problem of multiple hypotheses testing, especially critical in EDA, has prompted methodological approaches [29] that apply "a fortiori" when exploring massive data of unknown quality.

However data quality and the inherent pitfalls of blind data dredging do not figure prominently in hyped (mis)representations of Big Data as the brand new tool capable of answering fast and sure any scientific question. An extreme example is a 2008 essay [30] titled "The end of theory. The data deluge makes the scientific method obsolete". The author, at the time editor-in-chief of Wired Magazine, maintains that causal analysis is not any more required as statistical algorithms applied to immense databases can find patterns where the human mind operating by the scientific method cannot: correlations will be found and correlations stable in time is all what is needed for actions. The argument has been rebutted in its mathematical implications [31] and is epistemologically rough, failing to distinguish observable relations between events and mental logical relations like causality, that does not exist as an observable material "glue" sticking together events (Hume said it long ago [32]). Causal thinking is at the core of epidemiology, has guided its successful work in identifying disease etiological agents, and is now undergoing vigorous conceptual and analytical developments, relevant also to Big Data use and value. As stated in a recent paper [33]: "More recently, and looking to the future, the advent of omics technologies, electronic health records and other settings that leads to high-dimensional data, means that machine learning approaches to data analysis will become increasingly important in epidemiology. For this to be a successful approach to drawing causal inference from data, the predictive modelling aspects (to be performed by the machine) must be separated from the subject matter considerations, such as the specification of the estimand of interest and the encoding of plausible assumptions concerning the structure of the data generating process (to be performed by humans)". With this is as a general premise, three specific applications of Big secondary Data to epidemiological research can be discussed.

Real time surveillance

Surveillance systems monitoring disease occurrence, spread and outcomes as well as identifying pathogens has always been the key to infectious diseases control. Classical systems rely on professional sources such as "sentinel" general practitioner, involve manual operations and are relatively expensive and slow—strictly speaking not in real time-to accrue in numbers sufficient for robust analyses. Big Data in electronic form, rapidly transferrable and generated by different sources, for instance internet clicks on a vast array of terms possibly related to a disease that jointly can detect its presence and progression, seem a fast and cheaper alternative approach. In essence this was the Google Flu Trends (GFT) method, built to predict slower surveillance reports from the US CDC (Centre for Disease Control and Prevention). As it happened the method proved fast in reporting and fast in being dismissed as it was predicting more than double the proportion of doctor visits for influenza-like illness than the CDC estimate, based on surveillance reports from laboratories across the United States [34]. Looking today for 'flutrends' in Google.org one reads [35]: "It is still early days for nowcasting and similar tools for understanding the spread of diseases like flu and dengue-we're excited to see what come next". Reasons for the failure have been discussed and a 2016 special issue of the Journal of Infectious Diseases [36] provides in a series of articles a thorough review of strengths and weaknesses of the computerized Big Data approach to transmissible diseases surveillance. Validity—to be evaluated on a continuous basis against established systems—is here again stressed as the prime requirement. If a reference standard is not available, as may be the case in developing countries, at least a cross-validation of two or more Big Data indicators against each other should be performed. In addition representativeness, biases and data volatility (a general problem with web based data) need to be well controlled to justify a "cautious hope for digital technologies to deliver useful information for infectious diseases surveillance" [34]. Similar general considerations apply to another surveillance sector, pharmacovigilance, the key instrument to monitor post-marketing drug safety in populations [37, 38].

In clinical medicine surveillance is daily practiced at individual rather than population level, typically in intensive care, an ever expanding field that in the United States represents in monetary terms 1% of the gross domestic product [39]. As a young epidemiologist I inclined to Cochrane's scepticism [40] on the efficacy of the first coronary care units. In face of accumulating favourable evidence I changed my mind, and if today in need I would demand to be treated in an intensive care unit. I am now perplexed reading that in a systematic review [41] of seventy-two randomized controlled trials evaluating the effect on mortality of interventions in intensive care units (of all kinds in adults) only ten trials showed benefit, seven reported a detrimental effect and fifty-five showed no effect on mortality. The very high complexity and variability within and between patients of interventions in intensive care units might explain this finding. Here again a cautious hope can be expressed that the technical feasibility of creating complete and highly detailed databases coupled

with evolving methods of data analysis, including artificial intelligence, may lead to a better understanding of the evidence.

Real world evidence: effectiveness

"Real world evidence" is a fashionable expression that designates a genuine need for gathering evidence on a large number of issues related to health, for instance on how interventions successful in the experimental context of randomized controlled trials (RCT) work in ordinary practice within a given population. Even when the evidence from RCTs comes from a pragmatic trial in the same population where the treatment is to be used changes in practice may intervene with the passing of time capable of modifying the effectiveness. More often however the evidence comes from RCTs of the explanatory type, implying that treatment efficacy as measurable in the trial becomes affected by differences of patient profiles, compliance levels, and co-treatments between the patient populations of the trial(s) and of the routine use (not to mention off-label uses of drugs). Even more important is the occurrence of multiple pathologic conditions and associated treatments, today the norm in an ever expanding aging populations, that interfere with each other in responses to therapy. In these common circumstances resorting to observational studies to investigate treatment effectiveness is a solution that Big secondary Data, if easily accessible, render fast and relatively cheap, while a RCT even if practically feasible would take substantial time and large resources. This solution is however not exempt of pitfalls.

I recently participated (in France) in a clinical journal club discussing a paper [42] that reported a Danish nationwide cohort study investigating a suggested different effect of beta-blockers in patients with atrial fibrillation (AF) with or without heart failure (HF). AF and HF constitute a "dual epidemic" of burdensome chronic cardiac diseases in the elderly, most of whom have in addition other chronic ailments. The study could be rated of good quality as to population recruitment, follow-up, co-morbidities and treatment information (via well operating nationwide disease and prescriptions registries more than 200,000 patients were enrolled) and of excellent quality for data analysis (using propensity-score adjustment for confounders [including co-morbidities] in Cox regression, complemented by sensitivity analyses for potential biasing factors). The results indicated an advantage of including beta-blockers in the treatment of AF in both patients with and without HF. The cardiologists holding the journal club questioned the result validity and usefulness for their practice—the very purpose of effectiveness research—essentially on the ground that missing information on the ejection fraction had prevented stratifying patients by hemodynamic severity, insufficiently controlling for confounding by indication and potentially obscuring differential effects of beta-blockers. The example of this study prompts three considerations relevant to the use of Big secondary Data to measure treatment effectiveness:

(a) in general population-based health registries (of diseases, malformations, prescriptions, twins etc.) quality controlled to variable degrees are the best sources of Big secondary Data, but unlike in Denmark they are not always nor everywhere existing. However even with such registries information important to investigate a detailed and specific clinical question (as is often the case) may be missing;
(b) this plus other limitations (people selection, recording errors etc.) affect the most current sources of Big Data, i.e. routine EHR of irregular quality collected with different procedures and format in a variety of hospitals and health facilities;
(c) real world studies have "local value" and are indispensable to measure treatment effectiveness within specific local (in time and place) circumstances. Differences in effectiveness are to be expected in different settings, which makes inherently uncertain any generalization of results one may wish to attempt, typically when a study is intended to surrogate a hardly feasible RCT. This limitation is critical when the focus of interest is trying to detect and estimate treatment effects modifications, usually not large but important for adapting treatments to patients, induced by multiple co-morbidities and co-treatments.

Real world evidence: prediction

Quantitative risk of disease determination has considerably developed since the seminal work of the Framingham study predicting the probability of occurrence of coronary heart disease (CHD) as a function of a number of risk factors [43]. The Framingham Heart Study [4] included 5209 persons, aged 30–62, from a town in Massachusetts and the CHD risk equation went in use not only in the white USA population but more widely within the USA as well as in other countries. However the absolute risk estimates, usually at 10 years, were often not matching the incidence actually observed in a given population: this led to the development of many and larger prospective studies to obtain population-specific estimates of risk, not only of CHD but of other cardiovascular endpoints, notably stroke. Deriving risk equations, scores and charts by real world research using extensive data from ordinary clinical practice rather than from "ad hoc" studies was a further step

towards instruments adhering to the conditions of actual practice. Such equations and charts may be revised to improve prediction accuracy by including new predictor variables (not necessarily causal factors), as in the UK QRISK, based on Big secondary Data of more than ten million persons from representative general practices across the whole country, that is now in its third version [44]. The same approach is being extended to other diseases, particularly cancers for many of which however the long latency represents a main problem for accurate prediction [45].

In sum, valuable knowledge directly relevant to medical practice can be produced in an economically efficient way using Big secondary Data. However only the attitude labelled (by computer science specialists) "Big Data hubris" [34] may deceive into to believing that observational studies using Big secondary Data can abolish the need for (1) observational studies based on data collected for research purposes and (2) randomized controlled trials. Big Data are also often portrayed as the condition "sine qua" for precision medicine.

Precision medicine

Precision medicine embraces all steps of disease recognition (prediction of occurrence and of time course) and treatment (preventive or therapeutic) based on the hypothesis that massive data available for every individual will allow tailoring each step to each individual with maximal precision. The crux of the hypothesis, by itself totally generic, is quantifying precision in specific well-defined settings. In this respect at least two main issues need consideration: precision in prediction and aetiology, and precision for population health and for commerce.

Precision in prediction and aetiology

A central feature of precision medicine is the ability to accurately predict outcomes. The distinction between causal and predictive models is receiving increasing attention, particularly with the flourishing developments in causal analysis [46], but some key issues had been highlighted already time ago. Table 1 by Wald et al. [47] makes clear the point in the context of predictive screening tests based on continuously distributed Gaussian (or Gaussian by transformation) variables. The argument however applies to any Gaussian predictor variable, be it of disease occurrence (e.g. total cholesterol as a predictor of coronary heart disease) or of disease outcome in a precision clinical medicine setting (e.g. a cancer biomarker as a predictor of relapse). The table considers a predictor variable with approximately the same standard deviation in both people who do and do not develop the condition to be predicted. It shows how among the people who develop the disease the percentage of people correctly classified in advance, i.e. detected/predicted, by the predictor variable depends on the relative risk (or relative odds for a disease rare within the time horizon considered) between the highest and the lowest fifth of the variable distribution in the unaffected individuals. To qualify as a good predictor the variables should detect in advance most or all cases and at the same time not falsely predict as cases people who remain disease-free: in the table the percentage of false positives is assumed fixed at an optimal low 5%.

It can be seen that relative risks of the order of 100 are required to reach even a moderate percentage (around 50%) of correct predictions; and good levels of correct predictions, say at least 80%, would demand very high relative risks in the range of thousands. Taking a cruder approach not involving the Gaussian assumption a little arithmetic on a two-by-two table (varying disease risk in a population, proportion of subjects positive for a biomarker and associated relative risk) can show that in general relative risks of 5–10 or more are necessary to attain a reasonable percentage (say at least 50%) of correct predictions, unless one is prepared to accept that a substantial proportion, up to a large majority, of the subjects identified as bound to develop the disease turn out false positives. In plain words good predictor variables demand high relative risks, and failing this frequent misclassifications are unavoidable. The same applies to combination of variables such as the previously mentioned and widely popular risk scores and risk charts. A review [48] of three different cardiovascular disease risk functions (Framingham Risk Function, SCORE and the CVD Risk Score derived from the North Karelia project) applied to a large population risk factor survey database in Finland reported wide variations between the three functions in sensitivities and specificities for classifying high-risk subjects. A more recent study [49] used a UK clinical practice database to compare four risk scoring functions for cardiovascular diseases (ASSIGN, Framingham Risk Function, QRISK2 and a competing risk Cox model): it showed similar problems in the classification of high risk subjects, the primary purpose of all risk estimation systems.

In sharp contrast with prediction in etiological research even "small" relative risks (say below 2), if unbiasedly established, are of interest as they contribute to identifying disease causes. A relative risk of 2 also means that half of the disease cases among the people exposed to a cause are attributable to it, an obviously important point for modifiable causes. "Small" relative risks are most often encountered in studies of both environmental and genetic factors, and are an element in the hitherto limited performance of genetic tests to predict diseases arising from complex interplays, incompletely known, of multiple genes

Table 1 The percent of subjects correctly identified in advance out of all those who develop a disease depends on the relative risk between the highest and lowest fifth of the distribution of the predictor variable (modified from Wald et al. [47])

Relative risk	% of subjects who develop a disease correctly predicted
1	5
2	8
3	11
4	13
5	14
10	20
100	48
1000	74
10,000	89

and environmental agents [50, 51]. Sizable relative risks are encountered, but much less commonly, as for example in a case–control of hepatocellular carcinoma nested within the EPIC cohort [52]. Odds ratios ranging from near 6 to 9, depending on the analytic model, were found between the top and bottom fourths of the distribution of a biomarker, the enzyme gamma-glutamyl transferase, in blood samples collected at the entry in the cohort. A key question in this as in many similar studies is whether the biomarker is an antecedent, causal or at least predictive, of the disease or a product of the disease. Adopting an etiological viewpoint the authors tried to rule out the latter alternative by excluding cases with less than 2 to 4 years of follow-up. From a purely predictive viewpoint this would be irrelevant, the biomarker could be an antecedent, causal or not, allowing because of a sizable odds ratio a satisfactory prediction or could be a product of the disease, also allowing good prediction (but perhaps on the shorter term horizon of an early diagnosis). From a third viewpoint that considers possible effective interventions, the distinction between antecedents and consequences would however come back as relevant, because antecedents and disease would usually need different treatments.

Three important conclusions emerge from this discussion:

(a) High relative risks are detectable with small to moderate sample sizes, while as just seen high or very high relative risks are necessary for good predictive ability: it follows that Big Data involving millions of people are not at all indispensable for predictive precision medicine. More relevant may be the 'variety' dimension of Big Data, namely the spectrum of variables explored on each subject, including repeated measurements capturing random and systematic changes, periodic or directional.

(b) Like other advertising expressions 'precision medicine' conveys and suggests ideas, a main one being a medicine able to correctly classify in all circumstances and every occasion each individual's status, present (in diagnosis) and future (in outcome prediction). Useful as it may be to the marketing of a variety of projects that idea represents an unattainable objective, both in practice—among other reasons for the high relative risks usually required—and in theory for any system that is not completely deterministic and has not zero measurement errors.

(c) What is feasible is to refine the subjects stratification by risk level, namely risk prediction for *groups of people*, as internally homogeneous as possible, validly identifying strata with different risks not arising by chance due to extensive splitting of the population into subgroups. Subgroup analysis carries forward to patient populations in the form of studies of prognosis and, especially, of responses to treatments in RCTs of which it represents a major aspect [53]. Practically for any major disease, prediction of occurrence or outcome is a vast and burgeoning field of research, fruitful to the extent that it is not carried astray by mere addition of newly measurable biomarkers even if they do not appreciably contribute to stratification refining.

Precision for population health and for commerce

One definition of precision medicine recites [54]: "Precision medicine is a revolutionary approach for disease prevention and treatment that takes into account individual differences in lifestyle, environment and biology" and other more triumphant descriptions [55] stress the revolutionary character seemingly overlooking that since ancient Greece it has been a basic tenet of clinical medicine, however much science-based, that "each patient is unique". The description of precision oncology by DR Lowy, Acting Director of the USA National Cancer Institute, has a different ring [56]: "Interventions to prevent, diagnose, or treat cancer, based on a molecular and/or mechanistic understanding of the causes, pathogenesis, and/or pathology of the disease. Where the individual characteristics of the patient are sufficiently distinct, interventions can be concentrated on those who will benefit, sparing expense and side effects for those who will not ". This definition, extensible without change to precision medicine, portrays it as a phase in the incremental progress of scientific medicine rather than as a magic epochal revolution. Imaging, surgical and radiotherapy techniques have been radically transformed over the last 50 years and

the whole field of intensive care was born and developed even without the banner of precision medicine. Today primary solid cancer and metastases can be localized and selectively removed or destroyed with high accuracy and minimal lesions to surrounding tissues. For infectious diseases the concept and practice has been long established of basing therapy on the sensitivity to specific antibiotics, tested in the laboratory, of the etiological agent(s) isolated from an individual patient, a concept now translated into research on sensitivity to drugs of cancer cells with specific biomarker profiles from the tumour, metastases or circulating in the blood.

All these developments contribute to better tune interventions to the individual patient: however precision medicine as a universal recipe and approach to health is a mirage easily unveiled as soon as a naive question is asked: "Can health on the world scale result from intervening throughout life, on each occasion in precisely individualized way, on each of the seven and a half billions humans, one by one?". Patently population, sub-population and group interventions targeted on "average" persons, healthy and ill, will be indispensable for a very long time. This applies first of all to preventive measures that by definition must reach large sections or the totality of a population. Apart from the basic advantage of being healthy through prevention rather than ill and then treated, preventive interventions often offer the prospect of being economically more advantageous than treatment: the more individualized is in fact is a therapy, e.g. using drugs targeted on protein biomarkers specific to the tumorigenesis in a single or few patients, the narrower is the potential market for the drugs and higher the cost.

Popularly entertained mirages usually conceal some realistic but different driving motivation. For precision medicine commercial interests are a powerful driving force. Precision medicine is inherently geared to high-tech tools, as platforms for the various "omics" or high-resolution mass spectrometers, costly in development for which expenses recovery plus profits should then accrue from as large markets as possible. The promotional pressure of a fast growing number of new tests creates situations, for example in the cancer biomarkers assay, described as in need of "bringing order to chaos" [57]. Outside the market of health professionals the general public is also solicited as many companies engage in "consumer genomics", a market potentially worth from 2 to 7 billion dollars a year in the USA [58]. They advertise with success tests, often of unknown or poor validity, to trace genealogy, to identify gene-related tasting preferences or to detect alleged genetic predispositions to a variety of disorders, a potentially dangerous exercise against which the USA CDC has been issuing repeated, insistent warnings [59].

Time, maybe not too long, will tell whether precision medicine turns out to be just another name, perhaps needed like fashions are in society, for science-based medicine or whether it substantiates –as enthusiasts claim—a momentous change of pace and direction of its progress, bringing tangible benefits to patients and populations at large.

Part 2

The datome

> Une société' fondée sur des signes est, dans son essence, une société artificielle où la vérité charnelle de l'homme se trouve mystifiée (Albert Camus [60])

Unlike precision medicine the revolutionary jump, in respect to even recent past, of the current flow of all kinds of data is paramount. The daily production of data is estimated at 2.5 exabytes (10^{18} bytes) and more data are now produced every one-two years than during all preceding years in humankind history [61], bringing the cumulative volume of 4.4 zettabytes (10^{21}) in 2014 to a projected 44 zettabytes by early 2020, not without problems of storage capacity [62]. Health data represent a small fraction of the huge volume generated by all activities in society, mostly by individuals in such forms as movie downloads, VoIP calls, e-mails, Google searches, cellphone location readings. Substantial portions of the analysed data, including more or less stringently anonymized personal data, are incessantly traded between companies to create customer profiles for marketing purposes. Yet only a small percent of the total data are currently analysed: machine learning and artificial intelligence are deemed as essential tools to enhance the harvest of information from the data. It is inherently speculative where and when exponential rhythms of growth, like the one of data, can land. For medicine some medium-term scenarios [63] foresee a "disruption", in several areas of practice where algorithms will displace much of the physician's work. In this view clinical and imaging diagnosis (in radiology and pathology), prognosis, critical care monitoring may within years be largely taken over by algorithms entailing improvements beneficial to patients in technical, organizational and economic performance. This will redeploy the work of physicians who should be trained, or re-trained, in the data science, statistics, and behavioural science required to develop, evaluate and competently apply algorithms in clinical practice. Others express reservations [64], thinking that for machine learning in medicine we are already "beyond the peak of inflated expectations".

Scenarios apart, present reality and recent experience deserve close look and reflection. As three Boston physicians from the Massachusetts General and the Lown Institute wrote in May 2017 [65]: "It happens every day, in exam rooms across the country, something that would have been unthinkable 20 years ago: doctors and nurses turn away from their patients and focus their attention elsewhere—on their computer screens". From what I learn talking to clinician colleagues in Europe "across the country" holds "across countries" as well, at least the economically advanced ones (the less advanced encounter other and heavier difficulties). An extended quotation of Harvard's clinicians P. Hartzband and J. Groopman aptly describes a common situation prevalent in hospitals [66]: "Meanwhile, the electronic health record (EHR)—introduced with the laudable goals of making patient information readily available and improving safety by identifying dangerous drug interactions—has become a key instrument for measuring the duration and standardizing the content of patient–doctor interactions in pursuit of "the one best way". Encounters have been restructured around the demands of the EHR: specific questions must be asked, and answer boxes filled in, to demonstrate to payers the "value" of care. Open-ended interviews, vital for obtaining accurate clinical information and understanding patients' mindsets, have become almost impossible, given the limited time allotted for visits—often only 15–20 min. Instead, patients are frequently given checklists in an effort to streamline the interactions and save precious minutes. The EHR was supposed to save time, but surveys of nurses and doctors show that it has increased the clinical workload and, more important, taken time and attention away from patients". It can be added that whatever time might be saved is usually reallocated for economic efficiency to "process" more patients per unit time rather than to more time per patient. Ironically ample attention and literature is currently devoted to patient empowerment and involvement in co-decisions, which become void shells with inadequate time for nurse and doctor—patient interactions. I heard an internist trainee saying: "In my rotations the only place where I was not tightly constrained on my time with patients was the terminal care department", i.e. time was available only for the dying.

The current state frustrates the work of health professionals, with burn outs looming large [67], and denatures the relation to the patient. With a trend towards digital health data acquisition and processing similar to the one of the last couple of decades a sick person will be increasingly reduced and equated to a pack of data, many recorded by systems of direct sensors. The time honoured term "patient" had the condescending ring of paternalistic medicine and saw everyone as a could-be patient but it acknowledged the fundamental subjective trait of a person's suffering. Its popularity has declined in favour of an impersonal "user" of health services, a contractual "client" or "customer" and a more or less assiduous "consumer" of services. There is now a non-trivial risk of having to name the patient a "datome", to be dealt with essentially by algorithms (it would please post-humanistic "dataists" [68] that see not only patients but all organisms as algorithms defining physiological and mental processes and the universe as a flow of data).The disembodied datome will be a dominant reality to the extent that notwithstanding good intentions and the rhetoric of patient-centred and "personalized" medicine every health system and sub-system will be built, organized and run as a data-centric system, optimized (when successful) for the data flow and processing more than for the patient needs and attending health personnel functions. Wisdom warns against letting this trend to develop until the point of transforming medicine into datomes processing, efficiently performed and managed by artificial intelligence with minimal human intervention. It would put an end to the caring relation of humans to suffering humans that for thousands of years has been the constant inspiration and moral motivation of medicine, however short of ideal the actual practices and results [69, 70].

Wise epidemiology

> Where is the wisdom we have lost in knowledge?
> Where is the knowledge we have lost in information?
> (Thomas Stern Eliot [71])

….and information may be lost in data. Eliot's verses carry a message: in the conduct of life wisdom ranks as the most valuable asset. Wisdom is today most needed in medicine, as witnessed by the "Choosing wisely "clinical initiative [72], as well as in epidemiology. Choices in epidemiological research have become much more delicate today than when I started in 1964 because of the vast knowledge accrued in biology, medicine and epidemiology over the last half a century. With a much wider and varied knowledge-base from which to start projects there are many more opportunities for epidemiological research both pertinent and, equally, less or only remotely pertinent to public health. In practice more opportunities means more difficult choices. Together with research projects an agenda of "Wise epidemiology" for health in the digitized era should embrace several issues. I selected four for discussion and questions.

Research projects

On one side all new biomedical and biosocial knowledge may *ultimately* contribute to health and on this account

epidemiological studies essentially technology-driven or centred on investigating unexplored biomarkers and conjectural biological mechanisms or purely curiosity motivated are well justified. Epidemiology is however the most *proximate* knowledge base for possible public health actions and epidemiological studies capable of pertinently feeding results into such broad base deserve higher priority. A wise research agenda should not be mostly occupied by the former type of studies—for which funds and big datasets may often be more easily at hand—at the expense of the latter that keep the axis of epidemiological research firmly oriented to specific public health objectives visible within a time not far away and framed within the overarching goal of reducing the large health gaps prevalent within and between countries.

Training epidemiologists

In my view a mature epidemiologist, the reference for education and training, should first of all possess two broad perspectives: on how genetic, environmental and social factors interplay in conditioning the dynamic evolution of health and diseases in populations and on how science, i.e. epidemiology, articulates with applications in public health and clinical medicine, involving ethical and political issues. Second (s)he should have a good mastering of methodology, understanding of causal thinking and knowledge of a substantive field. Third, and equally important, (s)he should have acquired direct experience of all operations of an epidemiological study, conception, planning and feasibility testing, resource gathering, study conduct with field data collection, data analysis, interpretation, writing up and presentation of results. The prevalent reality appears different. Kenneth Rothman recently exposed [73] in this journal a "growing rift between epidemiologists and their data" referring to the absence in published papers of tables displaying the distribution of subjects by categories of key study variables, replaced most often by tables of effect estimates calculated from regression coefficients. The rift seems to me to be even wider and deeper. Many of the young (nominally) epidemiologists I meet are "de facto" data analysts that take for granted the availability of data sets, now often flowing in as Big Data. The validity of measurements I discussed earlier in this paper can hardly be appreciated if one has never being involved in the dirty job of data collection. When conducting studies [74, 75] in environmental and occupational epidemiology I inspected (with environmental measurements specialists) places and workplaces, examined records, engaging in conversation with and asking questions to residents, workers, administrative officers, technicians. Direct field practice in developed and developing countries has been indeed one of the strengths of epidemiology at the International Agency for Research on Cancer [76], where I spent most of my career. A key bonus of field involvement, whether in the population or clinical setting, is a tangible perception of what people experience and of what they expect from a study: it is a source of ideas on what is worth investigating. The need to train an increasing number of data scientists is today obvious as is the need for teamwork of an expanding range of specialists: does this however imply that the epidemiologist profile I have sketched is now obsolete? Whatever the answer the issue is vital for education and training, and worth discussing.

Doctors-and-patients

The expansion of computerized medicine is inducing major changes in its practices. There is room for directing these changes, either accelerating the trend towards the patient as a datome examined earlier on, or according to the spirit of a January JAMA editorial by Abraham Verghese and Nigam Shah [77] titled "What this computer needs is a physician". They advocate the working together of human and artificial intelligence arguing that "a well-informed empathetic physician armed with good predictive tools and unburdened from clerical drudgery" by artificial intelligence can come closer to optimally caring for the patient. Epidemiologists can sustain this development including in their evaluations of procedures and outcomes—objectively measured or assessed by health staff or patients—those depending to different degrees on artificial intelligence. But they can do even better by focussing on the nurse-patient and doctor-patient relations as key variables, usually the subject of much rhetoric while not being even measured. Pertinent measures, qualitative and quantitative, need to be developed capturing these variables in their multiple dimensions and permitting to investigate their dependence on the organization of a hospital or health facility. These structures can exploit the increasingly sophisticated techniques of data processing either to enhance economic efficiency and profitability or to pursue the objectives of measurable benefits to patients *and* quality of the nurse and doctor-patient relation, the kernel of humanistic medicine, the only one worth the name.

Health concepts

A newspaper article published at the time of the Davos 2018 World Economic Forum, in which technologies were a central concern, neatly titles [78]: "Health is redefined by technologies". Is this all right? Referring to the WHO definition of health ("Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" [79]) is of not much

assistance. It sets a perfection ideal that needs operational indications and limitations if it has to be concretely approached [80]. There is nothing in the definition itself that would prevent perfect health to be pursued via genetic engineering manipulations, functions augmentation, artificial construction of cyborgs or search of immortality. Simply ignoring the issue as a fad may work in normal times but not when the health paradigm is actually undergoing change through continuous technological innovations of all kinds. Discussing the issue is pertinent and timely for epidemiologists no less than is for geneticists, anthropologists, sociologists and ethicists, as a simple example illustrates. Data acquisition by multiple sensors of physiological functions can be of value when implemented in participants in an epidemiological study designed to investigate obesity. The same would hold for data acquired by personal devices sampling multiple air pollutants in a study of asthma. However there are doubts, unless one accepts that "data = good" by axiom, on the need for healthy joggers and sport amateurs in general to wear batteries of sensors (beyond a minimum possibly requested for safety). With such practices, or similar ones in other ordinary activities, amateur exercises in spontaneity, freedom and enjoyment of bodily sensations lost when sitting most of most days in offices, cars, planes, promote compulsory habits of instrumental self-documentation, surveillance and performance monitoring, as if one's personal identity could only be real if defined from the exterior by objective data, whenever possible Big Data. The relation to the crave for "selfies", the datome, the consumer genomics ("know your genes"), and more generally to the "data-driven world" [81] is evident and opens basic questions on human nature and cultures, of which the concepts of health underpinning all medicine and epidemiology are an integral component. To explore this rough territory a luminous thought by someone hardly suspect of being unscientific is a needed guide [82]: "It would be possible to describe everything scientifically but it would make no sense. It would be without meaning—as if you described a Beethoven symphony as a variation in wave pressure" (Albert Einstein).

Acknowledgements I wish to thank Albert Hofman for his invitation to write this essay and for his patience in waiting for it.

References

1. Watson JD, Crick FHC. Molecular structure of nucleic acids—a structure for deoxyribose nucleic acid. Nature. 1953;171:737–8.
2. Blackburn EK, Callender ST, Dacie JV, Doll R, Girdwood RH, Mollin DL, Saracci R, Stafford JL, Thompson RB, Varadi S, Wetherley-Mein G. Possible association between pernicious anaemia and leukaemia: a prospective study of 1625 patients with a note on the very high incidence of stomach cancer. Int J Cancer. 1968;3:163–7.
3. Doll R, Hill AB. Mortality in relation to smoking: ten years' observations of British doctors. BMJ. 1964;1:1399–1410, 1460–1467.
4. http://www.framinghamheartstudy.org/about-fhs/history.php. Accessed 9 Mar 2018.
5. Keys A, editor. Seven Countries: a multivariate analysis of death and coronary heart disease. Cambridge, MA: Harvard University Press; 1980.
6. https://www.moma.org/collection/works/3607. Accessed 9 Mar 2018.
7. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.
8. Hu H, Galea S, Rosella L, Henry D. Big Data and population health: focusing on the health impacts of the social, physical, and economic environment. Epidemiology. 2017;26:759–62.
9. Holmes DA. Big Data. A very short introduction. Oxford: Oxford University Press; 2017.
10. Schwab K. The fourth industrial revolution. Geneva: World Economic Forum; 2016.
11. Gilbert JP, Meier P, Rumke CL, Saracci R, Zelen M, White C. Report of the Committee for the assessment of biometric aspects of controlled trials of hypoglycemic agents. JAMA 1975; 231:583–608.
12. Margetts BM, Pietinen P, Riboli E, editors. European prospective investigation into cancer and nutrition: validity studies on dietary assessment methods. Int J Epidemiol. 1997;26(suppl 1):S1–89.
13. http://epic.iarc.fr. Accessed 9 Mar 2018.
14. http://www.ukbiobank.ac.uk. Accessed 9 Mar 2018.
15. Anderson CA, Petterson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010;5:1564–73.
16. Dunn WB, Broadhurst DI, Edison A, Guillou C, Viant MR, Bearden DW, Beger RD. Quality assurance and quality control processes: summary of a metabolomics community questionnaire. Metabolomics. 2017. https://doi.org/10.1007/s11306-017-1188-9.
17. Brennan P, Perola M, van Ommen GJ, Riboli E. European cohort Consortium. Chronic disease research in Europe and the need for integrated population cohorts. Eur J Epidemiol. 2017;32:741–9.
18. Vineis P, Chadeau-Hyam M, Gmuender H, Gulliver J, Herceg Z, Kleinjans J, Kogevinas M, Kyrtopoulos S, Nieuwenhuijsen M, Phillips DH, Probst-Hensch N, Scalbert A, Vermeulen R, Wild CP. The EXPOsOMICS Consortium. The exposome in practice: design of the EXPOsOMICS project. Int J Hyg Environ Health. 2016;220:142–51.
19. The MR-Base Collaboration. MR-Base: a platform for systematic causal inference across the phenome of genetic associations. BioRxiv. 2016. https://doi.org/10.1101/078972.
20. Yang A, Troup M, Ho JWK. Scalability and validation of Big Data bioinformatics software. Comput Struct Biotechnol J. 2017;15:379–86.
21. Olsen J. Using secondary data. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008. p. 481–91.
22. Graunt J. Natural and political observations mentioned in a following index, and made upon the Bills of Mortality. Facsimile ed. New York: Arno Press; 1975.
23. Stafoggia M, Schwartz J, Badaloni C, Bellander T, Alessandrini E, Cattani G, De Donato F, Gaeta A, Leone G, Lyapustin A, Sorek-Hamer M, de Hoogh K, Di Q, Forastiere F, Kloog I. Estimation of daily PM10 concentrations in Italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. Environ Int. 2017;99:234–44.
24. Krieger N. A century of census tract: health and the body politic (1906–2006). J Urban Health. 2006;83:355–61.

25. Greenland S, Rothman KJ. Fundamentals of epidemiologic data analysis. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008. p. 213–8.
26. CIOMS. International ethical guidelines for health-related research involving humans. Geneva: CIOMS; 2016. p. 41–5.
27. Armstrong B, Doll R. Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. Int J Cancer. 1975;15:617–31.
28. Pukkala E, Martinsen JI, Lynge E, Gunnarsdottir HK, Sparén P, Tryggvadottir L, Weiderpass E, Kjaerheim K. Occupation and cancer—follow-up of 15 million people in five Nordic countries. Acta Oncol. 2009;48:646–790.
29. Benjamini Y. Simultaneous and selective inference: current successes and future challenges. Biom J. 2010;52:708–21.
30. Anderson C. The end of theory: the data deluge makes the scientific method obsolete. http://www.wired.com/2008/06/pb-theory/. Accessed 9 Mar 2018.
31. Calude C, Longo G. The Deluge of spurious correlations in Big Data. https://hal.archives-ouvertes.fr/hal-01380626/document. Accessed 9 Mar 2018.
32. Hume D. In: Sellby-Bigge LA, editors. A treatise of human nature. Oxford: Oxford University Press; 1978.
33. Daniel RM, De Stavola BL, Vansteelandt S. Commentary: the formal approach to quantitative causal inference: misguided or misrepresented? Int J Epidemiol. 2016;45:1817–29.
34. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in Big Data analysis. Science. 2014;343:1203–5.
35. http://www.google.org/flutrends/about. Accessed 9 Mar 2018.
36. Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C. Big Data for infectious disease surveillance and modeling. J Infect Dis. 2016;214(suppl 4):S375–9.
37. Wang X, Hripcsack G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc. 2009;16:328–37.
38. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, Upadhaya T, Gonzalez G. Utilizing social media data for pharmacovigilance. A review. J Biomed Inform. 2015;54:202–12.
39. Cell LA, Mark RG, Stone DJ, Montgomery RA. "Big Data" in the intensive care unit—closing the data loop. Am J Respir Crit Care Med. 2013;187:1157–9.
40. Cochrane A. Effectiveness and efficiency: random reflections on health services. London: The Nuffield Trust; 1972. p. 51–3.
41. Ospina-Tascón GA, Buchele GL, Vincent JL. Multicenter, randomized, controlled trials evaluating mortality in intensive care: doomed to fail? Crit Care Med. 2008;36:1311–22.
42. Nielsen PB, Larsen BL, Gorst-Rasmussen A, Skjoth F, Lip GYH. Beta-blockers in atrial fibrillation patients with or without heart failure. Association with mortality in a nationwide study. Circ Heart Fail. 2016;9:e002597. https://doi.org/10.1161/CIRCHEARTFAILURE.115.002597.
43. Truett J, Cornfield J, Kannel WB. A multivariate analysis of the risk of coronary heart disease in Framingham. J Chron Dis. 1967;20:511–24.
44. Hippisley-Cox J, Coupland C, Brindle P. NIHR CLAHRC West. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017;357:j2099.
45. Thrift AP, Whiteman DC. Can we really predict risk of cancer? Cancer Epidemiol. 2013;37:349–52.
46. Authors Various. Special section: causality in epidemiology. Int J Epidemiol. 2017;45:1776–2206.
47. Wald NJ, Hackshaw AK, Frost CD. When can a risk factor be used as a worthwhile screening test? BMJ. 1999;319:1562–5.
48. Ketola E, Laatikainen T, Vartiainen E. Evaluating risk for cardiovascular diseases-vain or value? How do different cardiovascular risk scores act in real life. Eur J Pub Health. 2009;20:107–12.
49. van Staa TP, Gulliford M, Ng ES-W, Goldacre B, Smeeth L. Prediction of cardiovascular risk using Framingham, ASSIGN and QRISK2: how well do they predict individual rather than population risk ? PLoS ONE. 2014;9:e106455.
50. Janssens ACJW, van Duijn CM. Genome-based prediction of common diseases: advances and prospects. Hum Mol Genet. 2008;17:R166–73.
51. Hopper JL. Genetics for population and public health. Int J Epidemiol. 2017;45:8–11.
52. Stepien M, Fedirko V, Duarte-Salles T, Ferrari P, Freisling H, Trepo E, Trichopoulou A, Bamia C, Weiderpass E, Olsen A, Tjonneland A, Overvad K, Boutron-Ruault MC, Fagherazzi G, Racine A, Khun T, Kaaks R, Aleksandrova K, Boeing H, Lagiou P, et al. Prospective association of liver function biomarkers with development of hepatobiliary cancers. Cancer Epidemiol. 2016;40:179–87.
53. Tanniou J, van der Tweel I, Teernstra S, Roes KCB. Sub-group analyses in confirmatory trials: time to be specific about their purposes. BMC Med Res Methodol. 2016;16:20.
54. National Institutes of Health. About-all-of-us-research-program. https://allofus.nih.gov/about/about-all-us-research-program. Accessed 9 Mar 2018.
55. Naylor S. What's in a name? The evolution of "P-medicine". http://www.thejournalofprecisionmedicine.com/. Accessed 9 Mar 2018.
56. Lowy DR. The potential cost-effective precision medicine in low and middle-income countries. In: Presentation at the IARC 50th anniversary conference, Lyon, June 8, 2016.
57. Salgado R, Moore H, Martens JWM, Lively T, Malik S, McDermott U, Michiels S, Moscow JA, Tejpar S, McKee T, Lacombe D. IBCD-Faculty. Societal challenges of precision medicine: bringing order to chaos. Eur J Cancer. 2017;84:325–34.
58. Gavin T. The second coming of consumer genomics with 3 predictions for 2018. Posted at Medcitizens 26/7/2017. http://www.medcitynews.com. Accessed 9 Mar 2018.
59. Khoury MJ. Direct to consumer genetic testing: think before you spit, 2017 edition! Posted at CDC 18/4/2017. https://blogs.cdc.gov/genomics/2017/04/18/direct-to-consumer-2/. Accessed 2 Feb 2018.
60. Camus A. Discours de Suède. Paris: Gallimard; 1958. p. 33.
61. Marr B. Big Data: 20 mind-boggling facts everyone must read. Forbes Tech. Posted September 30, 2015. https://www.forbes.com/sites/bernardmarr/2015/09/30/. Accessed 2 Feb, 2018.
62. Rizzati L. Digital data storage is undergoing mind-boggling growth. EETimes. Posted 14/9/2016. http://www.eetimes.com/author.asp?. Accessed 9 Mar 2018.
63. Obermeyer Z, Emanuel EJ. Big Data, machine learning, and clinical medicine. N Engl J Med. 2016;375:1216–9.
64. Chen JH, Asch SM. Machine learning and prediction in medicine-beyond the peak of inflated expectations. N Engl J Med. 2017;376:2507–9.
65. Levinson J, Price BH, Saini V. Death by a thousand clicks: leading Boston doctors decry electronic medical records. http://www.wbur.org/commonwealth/2017/05/12/boston-electronic-medical-records. Accessed 9 Mar 2018.
66. Hartzband P, Groopman J. Medical taylorism. N Engl J Med. 2016;374:106–8.
67. Catalyst NEJM. Physician burnout: the root of the problem and the path to solutions. Waltham MA: Catalyst.njem.org; 2017.

68. Harari YN. Homo Deus. London: Vintage; 2016. p. 427–62.
69. Porter R. The greatest benefit to mankind. London: Fontana Press; 1997.
70. Wootton D. Bad medicine. Oxford: Oxford University Press; 2007.
71. Eliot TS. Collected poems 1909–1962. London: Farber & Farber; 1963. p. 161.
72. ABIM Foundation. Choosing Wisely. http://abimfoundation.org/what-we-do/choosing-wisely. Accessed 9 Mar 2018.
73. Rothman JK. The growing rift between epidemiologists and their data. Eur J Epidemiol. 2017;32:863–5.
74. Saracci R, Simonato L, Acheson ED, Andersen A, Bertazzi PA, Claude J, Charnay N, Estève J, Frentzel-Beyme RR, Gardner MJ. Mortality and incidence of cancer of workers in the man made vitreous fibres producing industry: an international investigation at 13 European plants. Brit J Ind Med. 1984;41:425–36.
75. Baris YI, Saracci R, Simonato L, Skidmore JW, Artvinli M. Malignant mesothelioma and radiological chest abnormalities in two villages in Central Turkey, An epidemiological and environmental investigation. Lancet. 1981;1:984–7.
76. Saracci R, Wild C. International Agency for Research on Cancer. The first fifty years, 1965–2015. Lyon: International Agency for Research on Cancer 2015. http://www.iarc.fr/en/publications/books/iarc50/index.php.
77. Verghese A, Shah NH. What this computer needs is a physician—humanism and artificial intelligence. JAMA. 2018;319:19–20.
78. Gogniat V. La santé redéfinie par les technologies. Genève: Le Temps. 28 Jan 2018.
79. World Health Organization. Basic documents. 47th ed. Geneva: WHO; 2009. p. 1.
80. Saracci R. The World Health Organization needs to reconsider its definition of health. BMJ. 1997;314:1409–10.
81. McKinsey Global Institute. The age of analytics: competing in a data-driven world. McKinsey Global Institute 2016. http://www.mckinsey.com/. Accessed 2 Feb 2018.
82. Einstein A. In: The ultimate quotable Einstein. Calaprice A, editor. Princeton: Princeton University Press; 2010. p. 409.