Big data: is clinical practice changing?

Luigi Tavazzi

Maria Cecilia Hospital, GVM Care&Research

The widespread diffusion of the digital culture and technology, involving both individual and population, together with the fast pacing digital globalization process (far surpassing ‘political’ globalization), is radically changing the world social landscape, including medicine and clinical research. The most significant change in clinical research is the ever more frequent acceptance of observational data, both through the use of registry of rare or common conditions and the implementation of capillary networks recording the daily clinical practice (Electronic Health Recording system). By becoming ‘observational’ clinical practice should change significantly: (i) record of different data (epidemiological, clinical, and administrative) in inter-operational database, producing a dynamic map of the health demands, either met or not, that allows a reconfiguration of the health systems capable to adapt to the shifting clinical needs. Implicate the larger group possible of patients and healthy individuals, who, through smartphone technology, could participate in primary and secondary prevention projects and epidemiological analyses. (ii) Support scientific research by integrating it with the clinical practice as instrument of good government that is scientific evidence-based Public Health System: the Learning Health System. The road will be long and gruelling. A first negative by-product is the proliferation of cybercrime throughout digital medicine.

Big data: what are they?

Big data is a technical term used to measure byte in Internet. Table 1 shows the current measuring unit: Exabyte (10^18 byte), Zettabyte (10^21 byte), and Yottabyte (10^24 byte), these are only theatrical units because as vast a database is not yet available on our planet. There are, though, digital database measured in Petabyte. The Petabyte is 10^15 bite, one million billion (1 000 000 000 000 000) byte.

For Big data is intended a database of 100 Petabyte or more. As an example, YouTube handles, each month, a flow of data of about 27 Petabyte.

At present, then, medical-scientific research is not dealing with Big data. The term is more journalism than technical and is used loosely to simply indicate ‘a lots’ of digital data. Nonetheless the current production of medical data could soon reach those numbers. In fact progressive globalization, based mainly on instantaneous data transactions, is growing exponentially and would require ongoing technical and cultural adjustment to govern the processes. In the year 2000, 25% of the world information was digital, and in 2013 reached 98%. Table 2 shows data relating to the collective and individual digital involvement in the world today. In the world about 3 billion people own a smartphone or a digital device able to communicate by signals and messages. Almost half of these individuals use apps, particularly for what concerns health/disease issues. In highly developed countries, half to two-thirds of the hospital/clinical centres employs technology for remote monitoring. In the United States, telemedicine sales will reach 3 billion dollars by 2020, as compared to 572 million in 2014. The inclusive term, common to all languages, to define digital technology as it applies to health care is Digital Health, a generic term, encompassing technology for collecting, sharing, and managing health-related data, as well as initiatives devoted to its improvement. There are two main fields in Digital Health: population based and individual based.

Published on behalf of the European Society of Cardiology. © The Author(s) 2019.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Population digital health

This activity is usually supported by public founding for official data, mostly administrative, or for objective-driven networks financed by institutional Agencies [National Institute of Health (NIH), European Community (EC), Sovereign States]. The critical elements necessary to implement a functional and useful digital health are many. Some of them are evident, yet not easy to realize. Among them, the choice of the information to be gathered (dataset) should derive from a compromise between the desired information and the feasibility of the specific activity that must be incorporated in the routine; the characterization of each datum should be agreed upon and constantly updated; the interoperability of the data base, and the traceability of the information in time as well as its usefulness should also be available. In other words, it is the setup of a system for routine medical data collection, homogeneous, and with capillary dissemination. This is the Electronic Health Recording (EHR) system, which goal is to provide a comprehensive, but analytic, description of the Health Care. In other terms is the integration of observational research methodology into clinical practice. In Europe the Scandinavian Countries are at the forefront of this process, which they started 30 years ago, consisting in systematic collection, in real time, of the national clinical practice, mainly hospital-based, utilizing simple and pragmatic registries, and supported both technically and financially by the central Government. At present these countries enjoy a wealth of information, also in the long term, and not exclusively relating to the cardiovascular system, unique in Europe because provide a realistic images of those countries, which are analysed by physicians and epidemiologists delivering medical-scientific analyses, not only administrative reports.

More recently, in several countries, registries have been activated for specific conditions, addressing both hospital-based and outpatient’s practices. These databases include ten of thousands of patients, yet are far from representing Big data. But their role has change drastically. Nowadays observational medicine has become the core of the health systems, and observational scientific research is its guiding tool.

Presently the United States are among the countries most engaged in the digital restructuring of their Health System. The 21st Century Cures Act mandates the Food and Drug Administration (FDA) to integrate the use of ‘real-world evidence’ in the approval process for new drugs, explicitly defining the data as ‘derived from sources other than randomized clinical trials’. Accordingly the FDA revealed that data from ‘real-world evidence’ derived from registries, and ever more often from EHRs and portable devices, are generating significant amount of data that will complement data from conventional clinical trials in their ‘regulatory decision making’ process (Health Data Management, 24 June 2016).

Two recent statements of the American Scientific community have addressed this methodological approach which places the registries at the centre of quality based medicine. The basic principles are the following:

1. Best clinical practices based on (methodologically correct) evidences.
2. Measure of the outcomes, fatal and non-fatal, of the treatment (systematic patients follow-up)
3. Techniques for data quality control, in particular, standardization of the nomenclature (definitions, starting with the event’s definition)

Furthermore:

1. Direct the registries toward clinical data (not only administrative data), designed to improve quality of care and outcomes.
2. Develop feedback useful for clinicians (actionable)

| Table 1 | Measurement units of internet |
|---------|-------------------------------|
| Megabyte (MB): $10^6$ byte (1 000 000 byte) |
| Gigabyte (GB): $10^9$ byte (1 000 000 000 byte) |
| Terabyte (TB): $10^{12}$ byte (1 000 000 000 000 byte) |
| Petabyte (PB): $10^{15}$ byte (1 000 000 000 000 000 one million billions byte) |
| Big data: $\geq$100 petabyte |

| Table 2 | The digital world 2010-2020 |
|---------|-----------------------------|
| 2010    | 2015 | 2020 |
| World population (billion) | 6.8  | 7.2  | 7.6  |
| Number of wired Device (billion) | 12.5 | 25  | 50  |
| Device per person | 1.8  | 3.5  | 6.6  |
| Number of patients with a smartphone (billion) | 0.5  | 3.0  | 6.1  |
| Number of wireless points (billion) | 3    | 47  | 500 |
| Number of transistors (million/chip) | 16/40 | 19/16 | 22/8 |
| Number of sensors | 20 million | 10 billion | 1000 billion |
| Number of individual with genetic sequencing | <10 million | 400 000 | 5 million |
(3) Consider the complexity and the frailty of the patients (rather than universal treatment according to the 'stack' concept).

(4) Assure communication and interoperability among the components (medical, interventional, and surgical) of the same or different clinical specialties.4

The system should not rely solely on the registries, but integrate with the Electronic Health Recording (EHR) system.6 Electronic Health Recording is different and complementary to the registries. In fact whether both systems employ observational methodology, the registries have a specific focus (disease, procedure, prevalence of a condition etc.) the EHR should: (i) document the clinical activity as a whole, (hospital and outpatients clinical data; administrative data; analytic management data; and long-term therapy monitoring data), recording it in such a fashion that data could be explored by multiple parties, producing a dynamic map of the healthcare necessities both met and unmet, thus allowing a constant reconfiguration of the health system, matching the varying clinical needs, as well as the accessibility to care for the people. (ii) Engage the largest group of people, both in good health or patients, interested in their health and owning a smartphone, to get involved directly in primary and secondary prevention studies, epidemiologic analysis (population, drug, diagnostic techniques, costs of care etc.). (iii) Support scientific research and use it as an instrument for good policies, the so-called Learning Health System.7

Use of Electronic Health Recording in clinical research

A very fertile research field is the search for phenotypes of complex diseases, taking advantage of the huge analytic capability of digital technology. A typical complex condition in cardiovascular medicine is heart failure, and in particular, heart failure with preserved ejection fraction. An analysis on a limited population of patients, but with abundant biological and instrumental data, identified three phenotypes markedly differing among each other, and with very different prognosis.8 The same methodology has recently been applied, with similar results, to other cardiovascular conditions. The basic tenant of this analysis is that each phenotype has its unique pathophysiology, and responds to treatment at variance from the other phenotypes. This is the main reason why heart failure with preserved ejection fraction does not respond to neuro-hormonal treatment, whereas the low ejection fraction counterpart does. There is a negative side to the analysis that is when the calculation suggests excessive phenotypical fragmentation, not clinically relevant, or patients grouping in different stages of the disease. Also, different conditions could be combined according to phenotypical similarities not clinically relevant. Another occasion/risk is the characterization of ‘computable phenotypes’, the combination of clinical signs/symptoms and instrumental data that statistically occur more frequently than by chance only.9

There are several kinds of ‘computable phenotypes’: (i) combinations derived from simple scanning of clinical database (Natural combination). (ii) Combinations derived from longitudinal database and/or cross-talk of several systems of health data collection in the long term (epidemiological, clinical, administrative), and time-sensitive (Clinical paths). (iii) Groups of responder/non-responder to treatments or preventive and therapeutic initiatives (Retrospective therapeutic phenotypes). (iv) Testing of new phenotypes, derived from clinical experience or scientific hypothesis, likely to determine a better response to treatment (Prospective therapeutic phenotypes).10

The National Institute of Health (NIH) introduced, some time ago, an interesting initiative based on EHR: the Undiagnosed Disease Program (UDP). The programme started in 2008 as Intramural Research Program included 150 patients, every year, referred to NIH for diagnosis not reached elsewhere. In 2015, the programme was expanded by including seven centres in the United States, and providing the network with a screening centre, two genetics laboratories, a bio-repository, and a centre for metabolomics. By the year 2017, each satellite centre should contribute at least 50 patients/year, whereas NIH should continue with the expected 150 cases/year. The total number of patients studied by the network should amount to 500/year. The patients eventually receiving a diagnosis will be included in the EHR system, searching for similar cases.11 A further NIH initiative, includes studies combining genomic data and EHR, focusing on variants of 100 genetic loci to be incorporated in the EHR and compared with already existing sequences. Five year grants have been assigned for this activity in 12 clinical institutions.

Individual digital health

Another interesting field relates to the data collected form individual patients using ‘m(mobile)Health’, and based on portable devices. The main device is the smartphone with all the available health apps (more than 160 000 on the market today), which can be connected with wireless gadgets for ambulatory monitoring of physiological variables, or recording of electric (Electrocardiogram, electroencephalogram) or acoustic (digital stethoscope connected to the smartphone) signals, collection of images (mostly echography) from all body’s areas, and conventional non-invasive recording such as blood pressure, glucose levels, oxygen saturation of haemoglobin, sweating, physical activity, implantable cardiac (pacemakers and defibrillators), or vascular (CardioMEMSTM, Champion) devices. To those information, the Genome and Epigenome data should be added, as well as the Microbiome (about 0.7-2.2 kg of physiologic bacteria for a 70 kg person) when it eventually will be available. These collected data are still far from fulfilling the definition of ‘Big Data’, but their growth is rapid, enhanced by the new available biomarkers, as well as the progress of nanotechnology and automated data collection and analysis and cost reduction. These elements, along with genetics, are the basic tenants of ‘Precision Medicine’. In this contest genetics is of outstanding value. Besides the sequencing of neoplastic tissue, providing the opportunity for individualized, and more effective, therapy, the genotype of healthy people, complementing personal health information, could be very
helpful in guiding present and future interventions. This concept is gaining momentum and, in many countries, is receiving public funding support. In the United States, there are three ongoing Federal programmes aimed at enrolling one million patients each (All of Us, the Cancer Moonshot, and the Million Veterans Program). Other programmes are oriented towards disease prevention (Million Hearts EHR Optimization Guides).

Problems and risks

Although the future is filled with optimism, expectations should be realistic and possible risks outlined. First of all how this fantastic data management innovation has been received by the medical community?

In the United States, which invested 50 billion dollars for the widespread digital update of the Federal Health Care system [MEDICARE and MADICAID (involving almost half million hospitals), Veteran Hospitals, and the Pentagon], there have been many snap-shot surveys, and there is agreement that all operations (which should have been completed by 2017) have been completed with an excessive time constrain, using incentives and sanctions in a frustrating fashion. All Medical Societies criticized, sometimes harshly, the process. The vast majority of them shared the objectives but not the approach or the timing. Two interesting surveys have been conducted in 2016 and published on EHR Intelligence. One involved the nursing staff reported the following results: 92% were not satisfied with the process; 85% reported that the system had problems and dysfunctions; 84% reported that the technology interfered with productivity and work flow. The other survey involved the medical staff and reported 90% burnout, with two-thirds of the doctors seriously considered a change of career. A further survey during the second half of 2017, reported a significant improvement, and 43% of the doctors where satisfied with functioning of the EHR.

It is likely that the first impact of the system on the clinical practice is disruptive, and it requires the necessary adaptation time, rather than incentives and sanctions.

From the operational stand point there are many aspects requiring choices and decisions, to provide the necessary reliability of the data upon which the government of the Health Care System is based.

Table 3 reports some of the risk apparent today.

The ‘cyber(un)security’

When the Health Care System is based on paper documents the opportunities for thieves are limited.

The transfer of data and medical information on a digital platform has open up a huge, and mostly unexpected, avenue for cybercrime. There are two kinds of crime: one destructive (minority) and the other one blackmail (prevalent). The first form of crime uses viruses which destroy irreversibly the ‘infected’ data. The second, blocks reversibly the operating system, demanding a ransom for its restoration. This has been the case of the famous Wannacry. Few cases reported in the medical or digital press in the US will better define the problem. This is somewhat a testing field for countermeasures against a ‘new’ criminal opportunity:

The number of hacked documents during the first semester of 2017 increased by 164% as compared to the second semester of 2016 reaching the incredible number of 1.9 billion cases (Health IT Smart Brief, 21 September 2017).

Warning for all healthcare businesses regarding Mamba, a new type of blackmail procedure able to encrypt all the hard disk of the victim organization denying access to the files and Window (Health Data Management, 28 September 2016).

The three larger digital violations in health care during the first 9 months of 2017 involved the data (mostly financial, such as payments) belonging to 1 497 800 people (Health IT Security, 15 September 2017).

An hacker called Skyscraper revealed that data of 500 000 sick children and 200 000 high school students have been sold on the dark web. The information included the name of the children and their parents, phone numbers, address, and Social Security number (Healthcare IT News, 3 May 2017).

An hacker called thedarkoverlord sold personal data (names and Social Security numbers) concerning 9 278 352 US patients for 500 000 dollars (Motherboard, 26 June 2016).

Sales on more than 6 300 dark web marketplaces have increased by 2500%, from 249 287 US $ in 2016 to 6 237 248 US $ in 2017 (October). Most affected health and legal businesses.

Ventures, Cybersecurity Agency, estimated that by 2020, the number of blackmails affecting the health care industry with increase four folds, and by 2021 the cybersecurity world market will exceed 65 billion dollars (BeckersHospitalReview.com, 7 April 2017).

Some final thoughts

Is medicine changing because of Big Data? No! at least according to the current guidelines of the major

| Table 3 Possible risks of the implementation of systematic Digital Registration System for health care data (Electronic Health Recording system) |
| --- |
| • Dataset too ‘simple’ for feasibility reason |
| • Overflow of information |
| • ‘Administrative’ data access (little scientific value) |
| • Creation of ‘calculated’ conditions/computable phenotypes and ‘creation’ of drugs for inconsequential target |
| • Fragmentation of diseases in phenotypes non-clinically relevant or, aggregation of non-related conditions according to some phenotypical similarities |
| • Overflow of genomic data in a non-prepared cultural contest |
| • Crediting with casual occurrence non-casual genetic mutation and ‘creation’ of drugs for inconsequential target |

Modified from Tavazzi.
International Medical Societies. The criteria for evidence and recommendations are the same. Observational research has been, by and large, ignored for lack of accepted quality criteria necessary for its consideration in the recommendation process. Some of the position statements4,5 and the FDA position regarding the use of observational data in the regulatory process, are important, but there is a necessity for accepted and shared rules for implementation by the medical and scientific community. To this point, available published studies concern epidemiology or compliance with guidelines in medical practice. Observational research has a very important role, and theoretically with a wider scope than clinical research, but requires an established scientific methodology, not only illustrative reporting, to gain full acceptance in the medical culture.

Medicine is going through a moment in which there is overabundance of health-related data, while at the same time, is growing a strong impulse toward individualized approach to the patient, leading to the so-called precision medicine. In other words, the last few decades have witnessed the assertion of evidence based medicine, mostly relying on large trials, ever more pragmatic, hence less selective, aimed at identify dosages and treatments effective for the ‘majority of patients’ (optimal medical therapy!), but now the wave has changed, and we cherish the opposite concept, that is individuality. The rationale for this shift is solid, and we have now the means to realize it.

We should be vigilant as not to incur in errors for lack of knowledge, carelessness, and superficiality. The near future of medicine will be characterized by an overflow of information not easy to categorize or decipher for our cultural shortcomings. Furthermore we should take into account the ‘evidence’, not new, published on the British Medical Journal,12 reporting that in the US diagnostic errors affect 15% of all clinical encounters, involving 12 million adult patients annually, and being responsible for permanent damages or death of 160,000 patients every year.

This is the third most common cause of death, after cardiovascular diseases and cancer. The ‘Intelligent Health System’ (Learning Health System) will not be implemented in the short term unless some conditions are fulfilled. The first is that National Policies should consider health care as a priority, with the appropriate administrative and financial coverage. The second requires that scientific processes be at the foundation of the ‘Intelligent Health System’. In some countries these concepts are integral part of the political strategy. In the US, for instance, NIH received 30 billion dollars to invest in clinical research utilizing EHR as data source. The developing Health Care System is supported by clinical research using its data and controlling its evolution.

Conflict of interest: none declared.

References
1. Bhavnani SP, Narula J, Sengupta P. Mobile technology and the digitization of healthcare. Eur Heart J 2016;37:1428-1438.
2. Topol EJ, Steinshubl SR, Torkamani A. Digital medical tools and sensors. JAMA 2015;313:353-354.
3. Goodman SN, Schnerewiss S, Baiocchi M. Using design thinking to differentiate useful from misleading evidence in observational research. JAMA 2017;317:705-707.
4. Bhatt DL, Drozda JR, Shahian DM, Chan PS, Fonerow GC, Heidenreich PA, Jacobs JP, Masoudi FA, Peterson ED, Welke KE. ACC/AHA/STS Statement on the Future of Registries and the Performance Measurement Enterprise: a report of the American College of Cardiology/American Heart Association Task Force on Performance Measures and the Society of Thoracic Surgeons. J Am Coll Cardiol 2015;66:2230-2245.
5. Windle JR, Katz AS, Dow JP Jr, Fry ET, Keller AM, Lamp T, Lippitt A Jr, Paruche AAP, Resnic FS, Serwer GA, Slotwiner DJ, Tcheng JE, Tilkemeier PL, Weiner BH, Weinstein WS. 2016 ACC/AJES/ASNC/HRS/SCAI health policy statement on integrating the health care enterprises. J Am Coll Cardiol 2016;68:1348-1364.
6. Roe MT, Mahaffey KW, Ezekowitz JA, Alexande JH, Goodman SG, Hernandez A, Temple T, Berdan L, Califf RM, Harrington RA, Peterson ED, Armstrong PW. The future of cardiovascular clinical research in North America and beyond—addressing challenges and leveraging opportunities through unique academic and grassroots collaborations. Am Heart J 2015;169:743-750.
7. Fiuza M, Califf R. The US Food and Drug Administration and the future of cardiovascular medicine. JAMA Cardiol 2016;1:950-952.
8. Shah SJ, Katz DH, Selvaraj S, Burke MA, Yancy CW, Gheorghiade M, Bonow RO, Huang CC, Deo RC. Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation 2015;131:269.
9. MacRae CA, Vasan RS. The future of genetics and genomics: closing the phenotype gap in precision medicine. Circulation 2016;133:2634-2639.
10. Tavazzi L. Le radici della rapida e profonda evoluzione in corso della struttura e della metodologia della ricerca clinica. G Ital Cardiol 2016;17:181-185.
11. Gahl WA, Wise AL, Ashley EA. The Undiagnosed Diseases Network of the National Institutes of Health: a National Extension. JAMA 2015;314:1797-1798.
12. Khullar D, Jena AJ. Reducing diagnostic errors: a new imperative in quality health care. BMJ 2016;352:i417.