A Revolução Silenciosa do *Big Data* em Medicina

The Quiet Revolution of Big Data in Medicine

Bernardo Neves[1], Anabela Raimundo[1], Ziad Obermeyer[2]

Palavras-chave: Conjuntos de Dados; Medicina Baseada em Evidência; Registos de Saúde Electrónicos.
Keywords: *Datasets; Electronic Health Records; Evidence-Based Medicine.*

The *big data* era

The amount of data that is produced every minute grows exponentially and its economic value seems to follow the same trend. In a recent article in *The Economist*, a parallel is made between the current importance of data and that of oil at the beginning of the 20th century.[1] However, like any other raw material, in order to obtain its maximum value, data must be extracted, transported and refined. Nowadays, any industry and economic sector is looking for strategies to take full advantage of and profit from data, generating knowledge that translates into actions that will represent a competitive advantage in the market. Examples of success in this process known as *data mining* are, for example, decisions to change stocks in supermarket chains according to predicted sales or telecom marketing campaigns targeting specific customers whose probability of selling is predicted to be higher.[2] Actions such as these characterise the current paradigm of the so-called *big data* in which *big* refers not only to the volume of generated data but also to its variety, velocity of processing and complexity of its analysis.

Healthcare is no longer an exception to this phenomenon. In Portugal, like in most of other European countries, we have witnessed a great investment in health information systems and electronic health records during last years, both in public and private parties. Hospital systems like ours have been capturing all medical encounters, procedures and exams on a digital format for years, presenting a valuable data source for research that is now starting to be used with promising results. The impact of *big data* in healthcare is however, in a general way, still much focused on promises and less on the challenges we face in order to be able to properly analyse these important data sources. Surprisingly, little discussion is being promoted about this subject on the medical community, comparatively to the tremendous impact it is about to bring to the future of medicine.

Data generated in healthcare

Oe of the most visible aspects of data generated in healthcare consists on the vast amount of "unstructured data" that live in the electronic health records. Radiology reports, laboratory results or physician notes are just a few examples of clinical data stored in hospital servers. Although their use is now generalised, their adoption has not always been straightforward and many award them a significant share of the existing professional dissatisfaction and burnout, in particular because of the high administrative burdens they entail.[3] It is widely believed, however, that their adoption could lead to an improvement in the quality of healthcare provided, in addition to generate the potential of using them for other purposes, such as internal auditing, economic studies and now with particular interest, clinical research.[4] The very high costs of conducting conventional clinical trials, their often weak external validity, and the increasing availability of "real-life" data sources have contributed to a broad movement to encourage the sharing and use of various "islands" of information stored in health services worldwide.[5,6]

The great data revolution in medicine might however happen because of the use of less conventional or secondary data sources. Administrative databases, data from the internet and social networks, information generated by biometric sensors and devices are all examples of data currently being used in clinical research.[7] Patient-generated data will likely play a key role in healthcare daily life in the near future.[8] So-called wearables, biometric evaluation sensors such as watches and wristbands are now capable of continuously and reliably analyse clinical variables such as range of movement, capillary glycemia, heart rate or blood pressure and this is a field of rapid technological growth and venture capital investment. Although most of these devices still have a recreational usage and its clinical impact is still unproven, it is quite possible that we will be able to change the current health paradigm from an overly reactive and curative medicine, to a new era where preventive medicine prevails and chronic conditions are managed in a very different way with greater patient involvement. Regulation of these devices and their usage is of great importance in order to avoid the spread

[1]Departamento de Medicina Interna - Hospital da Luz, Lisboa, Portugal
[2]Brigham and Women's Hospital, Harvard Medical School, Boston, USA

http://revista.spmi.pt – DOI: 10.24950/rspmi/Perspective/2017

of poorly designed, unsafe or even fraudulent applications.[9]

Challenges with data usage in health

In contrast to the hype related to the amount of generated data in healthcare, surprisingly there has been little impact of its usage for clinical purposes so far. The scientific discipline of *machine learning* was born from the intersection between computational science and mathematics and studies the ability of computers to automatically extract information from data.[10] Being especially competent in identification and prediction tasks, "trained" machine learning algorithms have proven able to generate knowledge in many other fields, showing complex associations between parameters and outcomes. Some examples of their usage in medical issues are the identification of subgroups of patients in heart failure with preserved ejection fraction,[11] automatic recognition of diabetic retinopathy,[12] confirmation of diagnosis in radiology or dermatology sometimes with superior performance to doctors themselves, or the detection of histological patterns with prognostic impact in Oncology.[13] Their clinical utility remains however uncertain, because being able to make a good prediction does not necessarily mean there will be an impact in patient care.

The development of increasingly complex algorithms is also leading to the development of early warning and clinical decision support systems, often integrated into electronic health records, thus paving the way for more efficient medicine with fewer administrative tasks performed by clinicians and promoting reduction of medical error.[14] At the root of the construction of these artificial intelligence-based systems is however the quality of clinical records and all the remaining information that is generated about the patients, since even the most sophisticated algorithm will do little if the data that feeds it is inaccurate or have little explanatory power of the outcomes being analysed.[15]

New tools, new concerns

Besides assuring information quality, there are several other difficulties intrinsically related to medicine in particular, that if not properly taken into account, could undermine the use of valuable health data. Beginning with the very definition of many diseases and the outcomes measured, too often they are subject to significant inter-observer variability or lack of a *ground truth.* For example, too often we don't actually know when a patient has sepsis, so algorithms that predict sepsis will incorporate all the uncertainty and error on the part of doctors. This is a very relevant aspect in the use of *machine learning* algorithms, since their performance is highly dependent on the ability to define variables of interest and categories of interest since the beginning.[16]

On the other hand, the development of more efficient *machine learning* algorithms is frequently achieved at the expense of greater complexity, which makes them difficult to understand, somehow turning them into "black boxes".[16] This leads to liability problems with their usage, as the example with automatic cars illustrates where the accountability of decisions in limit situations can be complex. In medical issues it may also be difficult, for example, to ascertain responsibility for clinical decisions that depend, albeit in part, on the use of complex risk stratification and clinical decision support tools.[17]

Another important aspect has to do with individual privacy and information security. The development of information technologies has led to the current paradigm of the so-called *internet of things* in which electronic devices increasingly communicate to each other in an unprecedented way. This, coupled with the size and value of the data nowadays, explains the increasing number of computer attack attempts, with health being an especially sensitive area with unpredictable consequences, as the recent example of cyber-attack *WannaCry* demonstrates.[18] While there is now greater investment and concern for the security of health information systems and information anonymization, it is virtually impossible to assert absolute guarantees in this area.

Future perspectives

There is now great potential to generate new knowledge about diseases, new ways to treat them and possibly adopt a more preventive attitude in medicine. All this technological development brings to the present generation of doctors an important responsibility of learning and understanding these new structures and information circuits, so as to make the most of the investments made and not to under-exploit their full potential. The pressure generated by the availability of data and ease of access to the analysis tools even for newbies, makes it fundamental that physicians acquire basic knowledge in the field of *data science*, from the data acquisition process, to the machine learning algorithms used for its analysis. As with any other scientific evidence, understanding the context in which data are produced and the methodologies used in its analysis is crucial to identify biases and not draw erroneous conclusions.[19] Various stakeholders, including medical professionals, academics, hospital administrators and patient representatives should participate in an in-depth discussion on how health-related data can be used to support the development of medicine, safeguarding fundamental aspects such as privacy and safety of people.

Internal Medicine has an increased responsibility here, as hospitals of the future will certainly have multidisciplinary teams dedicated to data science and clinical elements with deep knowledge in this area will be required. Internal medicine doctors, who are responsible for the most complex patients within the healthcare system and often work as connectors and facilitators between the many stakeholders in the health care process, are thus especially well prepared for this function.

As in all our medical practice, the Hippocratic principle of non-maleficence must also guide us with the adoption of these new tools. The fascination with technological novelty must not hide the fact that even the most complete datasets will never represent the whole person to which they refer to, in their psychological, environmental, social context or even in their place in the complexity of the healthcare system. On the other hand it will also be important to prevent the ignorance and inertia of altering the status quo that may deprive our patients of a medicine based on more accurate and appropriate information: the promise of *personalised medicine*.[20] We all have the responsibility to make the most of technological development in health, believing that automation will give us more time to do what we were taught to do and robots will not quickly replace us with: to listen, observe and understand our patients. ■

Conflicts of interest: The authors have no conflicts of interest to declare.

Financing Support: This work has not received any contribution, grant or scholarship.

Conflitos de Interesse: Os autores declaram a inexistência de conflitos de interesse na realização do presente trabalho.

Fontes de Financiamento: Não existiram fontes externas de financiamento para a realização deste artigo.

Recebido: 01/09/2017
Aceite: 01/10/2017

Correspondência: Bernardo Duque Neves
bernardo.neves@hospitaldaluz.pt
Hospital da Luz-Lisboa – Av. Lusíada, 100. 1600-650 Lisboa, Portugal

REFERENCES

1. Data is giving rise to a new economy. The Economist [Internet]. [accessed 2017 Aug 28]. Available from: https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy
2. Provost F, Fawcett T. Data science and its relationship to big data and data-driven deci-sion making. Big Data. 2013r;1:51–9.
3. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, et al. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. Mayo Clin Proc. 2016 ;91:836–48.
4. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. Yearb Med Inform. 2017;26.
5. Sim I. Two ways of knowing: big data and evidence-based medicine. Ann Intern Med. 2016;164:562–3.
6. Cano I, Teny A, Vela E, Miralles F, Roca J. Perspectives on Big Data applications of health information. Curr Opin Syst Biol. E2017;3:1–13.
7. Scott PJ, Rigby M, Ammenwerth E, Brender McNair J, Georgiou A, Hyppönen H, et al. Evaluation Considerations for Secondary Uses of Clinical Data: Principles for an Evidence-based Approach to Policy and Implementation of Secondary Analysis. A Posi-tion Paper from the IMIA Technology Assessment & Quality Development in Health Informatics Working Group. Yearb Med Inform. 2017;26.
8. Topol E. Digital medicine: empowering both patients and clinicians. Lancet. 2016;388:740–1.
9. Piwek L, Ellis DA, Andrews S, Joinson A. The rise of consumer health wearables: promises and barriers. PLoS Med. 2016;13:e1001953–9.
10. Deo RC. Machine learning in medicine. Circulation. 2015;132:1920–30.
11. Shah SJ, Katz DH, Selvaraj S, Burke MA, Yancy CW, Gheorghiade M, et al. Phe-nomapping for novel classification of heart failure with preserved ejection fraction. Cir-culation. 2015;131:269–79.
12. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Devel-opment and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316:2402–10.
13. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med. 2016;375:1216–9.
14. Blumenthal D. Realizing the value (and profitability) of digital health data. Ann Intern Med. 2017;166:842–3.
15. Chen JH, Asch SM. Machine learning and prediction in medicine-beyond the peak of inflated expectations. N Engl J Med. 2017;376:2507–9.
16. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA. 2017; 318:517-8.
17. Fraccaro P, Sullivan DO, Plastiras P, Sullivan HO, Dentone C, Di Biagio A, et al. Behind the screens - Clinical decision support methodologies – A review. Health Policy and Technology. 2014;4:1–10.
18. Gordon WJ, Fairhall A, Landman A. Threats to Information security - public health implications. N Engl J Med. 2017;377:707–9.
19. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence - what is it and what can it tell us? N Engl J Med. 2016;375:2293–7.
20. Bierman AS, Tinetti ME. Precision medicine to precision care: managing multimorbidity. Lancet. 2016;388:2721–3.