What's new in ICU in 2050: big data and machine learning.

Sébastien Bailly, HP2 laboratory, University of Grenoble Alpes, Grenoble, France and Department of Physiology and Sleep, Grenoble Alpes University Hospital (CHU de Grenoble), Grenoble, France

Geert Meyfroidt, Department and Laboratory of Intensive Care Medicine, University Hospitals Leuven and KU Leuven, Leuven, Belgium

Jean-François Timsit, Inserm UMR 1137-IAME Team 5-DeSCID : Decision SCiences in Infectious Diseases, control and care INSERM/Paris Diderot, Sorbonne Paris Cité University, Paris, France and Medical ICU, Paris Diderot University/Bichat University Hospital, APHP, Paris, France and Inserm, U1042, Grenoble, France

The era of big data

The amount of digitalized data that the world produces today is by all measures unseen and spectacular. Social media, e-commerce, and the Internet of things generate approximately 2.5 quintillions of bytes per day, an amount that equals 100 million Blu-ray discs, or almost 30,000 GB per second. Data grows exponentially, and 90% of all data on the Internet has been created since 2016. This trend will continue in the next decades [1]. Such datasets of unimaginable size cannot be maintained with traditional database management technology, or examined with traditional statistical techniques. The general term for methods to manage and analyze such unstructured datasets is "big data". Although the term is often ill-defined and improperly used, the five "Vs" concept is a good summary: Volume, Velocity, Variety, Veracity, and Value referring to, respectively, the large quantity of data; the speed of acquisition; the diversity of data sources; the uncertain data quality; and the possible valorization. The last of these is without a doubt the "V" that matters most. In medicine, the first datasets amenable to big data were generated when techniques to process genomic data became available, for instance for tumor genotyping in oncology [2].

From big data to structured data in medicine and ICU

Computerization of medical activities has lagged behind compared to other fields of human activity, such as industry, trade, or aviation, but is now widespread. Medical data, previously available only on paper, are now being collected continuously at the bedside [3]: demographic data, clinical observations, physiological signals, and results from laboratory analyses constitute data that combine large quantity (Volume) and high speed of acquisition (Velocity). In addition, non-numeric or unencoded data such as diagnostic information, annotations from medical staff in free text, radiology and imaging data, text, audio or video files, etc., can be stored. Unlike big data, medical data are not completely unstructured and can be considered as a huge multilevel database, which can be structured (Fig. 1).

The high-tech intensive care unit (ICU) is at the forefront of these new developments. In their daily practice, ICU clinicians deal with these different data sources on a patient or departmental level, but are less familiar with advanced analytic methods that could exploit these data to generate new medical knowledge. By applying big data techniques to the large amounts data that are continuously generated in daily healthcare practice, a continuously learning healthcare system could be designed [4, 5] (Fig. 1) that is able to generate knowledge from all these different sources of big data, such as genome sequencing, omics, administrative database, social networks, connected objects, telemedicine, etc.

In 2017, such a learning healthcare platform does not exist for many reasons: first, the technological challenge to set up such a large and yet accessible system is huge; second, inherent problems of artefactual, spurious, and incomplete data will need to be dealt with; third, data standards will need to be developed; and fourth, and most importantly, proprietary, legal, and privacy issues need to be solved [6]. Large-scale acquisition of medical data raises the question about personal consent of the patient. Furthermore, many legal, property and responsibilities issues must be solved. In particular, the ownership of the data, the intellectual property of the structuration of the metadata created will need legal clarifications. Similarly, the conditions of sharing the ownership of new scientific discovery or knowledge will be a barrier for international collaborations.

Machine learning

Pattern recognition, predictions, and generating data-driven hypotheses from large amounts of multidimensional data can be done with automatically learning algorithms, where computers build models that are not explicitly programmed in advance, based on the data being fed to them. These algorithms are known as machine learning, a broad term for an exciting new statistical discipline which encompasses a variety of methods that have been developed together with the data revolution of the past decades. However, overconfidence in the data, and an incomplete understanding of the underlying relationships, could lead to dubious results and spurious interpretations. Unmeasured confounders, missing data, and inherent noise can generate false positive associations [7]. The problem of multiple testing and false positive associations can be amplified in automatically learning algorithms, known as data dredging. Obviously, a statistical association can never be interpreted as a proof of a causal relationship.

From big data to personalized medicine

Large multicenter randomized clinical trials (RCT) are considered the gold standard to demonstrate efficacy or safety, and are still the only way to demonstrate causal inference. However, they are not devoid of problems. First, RCTs are difficult to undertake, often require a large sample size, and are expensive. Second, they have a low external validity and a sometimes limited applicability to clinical practice [8]. In particular for comparative effectiveness research, big data has been proposed as a complementary approach to RCTs [8]. Precision or personalized medicine, by tailoring diagnostic and therapeutic strategies for each specific patient, is a particular challenge [9]. This approach accounts for individual variability, particularly in the case of rare diseases. In addition, it can be used to classify patients into homogeneous subgroups, especially in conditions that are known to represent a heterogeneous spectrum of underlying clinical conditions, such as sepsis [10], and maybe identify specific subgroups which can be recruited for a targeted RCT. Current pioneering initiatives, such as the OutcomeREA® database, which has prospectively collected data since 1997 from over 22,000 patients or 190,000 patient days, have demonstrated that it is possible, with new statistical approaches for causal inference [11], to investigate clinical assumptions such as the impact of empirical antifungal treatment on patient prognosis in critically ill non-neutropenic patients [12]. This hypothesis was later confirmed in the EMPIRICUS RCT [13], performed in a more specific population. Another dimension of personalized medicine is the development of clinical prediction tools [14]. Current, early versions of such decision support systems are expected to become more sophisticated as the volume, content, completeness, and organization of data improve. A random forest model based on clinical data was able to outperform an existing biomarker in the prediction of acute kidney injury [15]. Such models could be used as an early warning system in the ICU, or as a research tool to detect those high-risk patients that would benefit from a certain intervention to be included in an RCT. The ultimate dream of personalized medicine would be the integration of genomics data, through a stable, safe, and

yet open community where these data can be shared. Finally, analysis of structured medical data can be used in the case of adaptative designs of RCT to test multiple interventions simultaneously while minimizing the sample size [16]. Moreover, given the incomplete external applicability of RCTs, non-RCTs based on large high-quality databases or well-structured metadata will be considered of major value.

Conclusion

The fourth industrial revolution, the data revolution, has just begun. Digital data are and will be a big source of information, and possibly of innovation. This field is truly translational, and transdisciplinary collaborations of domain experts, epidemiologists, biostatisticians, and bio-informaticians are key. The success of this revolution in medicine, and in the ICU in particular, will depend on the ability to create a stable and open community where data can be shared. Another main concern to ensure relevant results is the quality, structure, and correctness of the initial data. A community of responsible stakeholders, consisting of patients and researchers, will have to manage and oversee the correct use of data. In addition, we need to be aware of the inherent limitations of big data research, and realize that it can improve or guide, but never substitute experimental research [17].

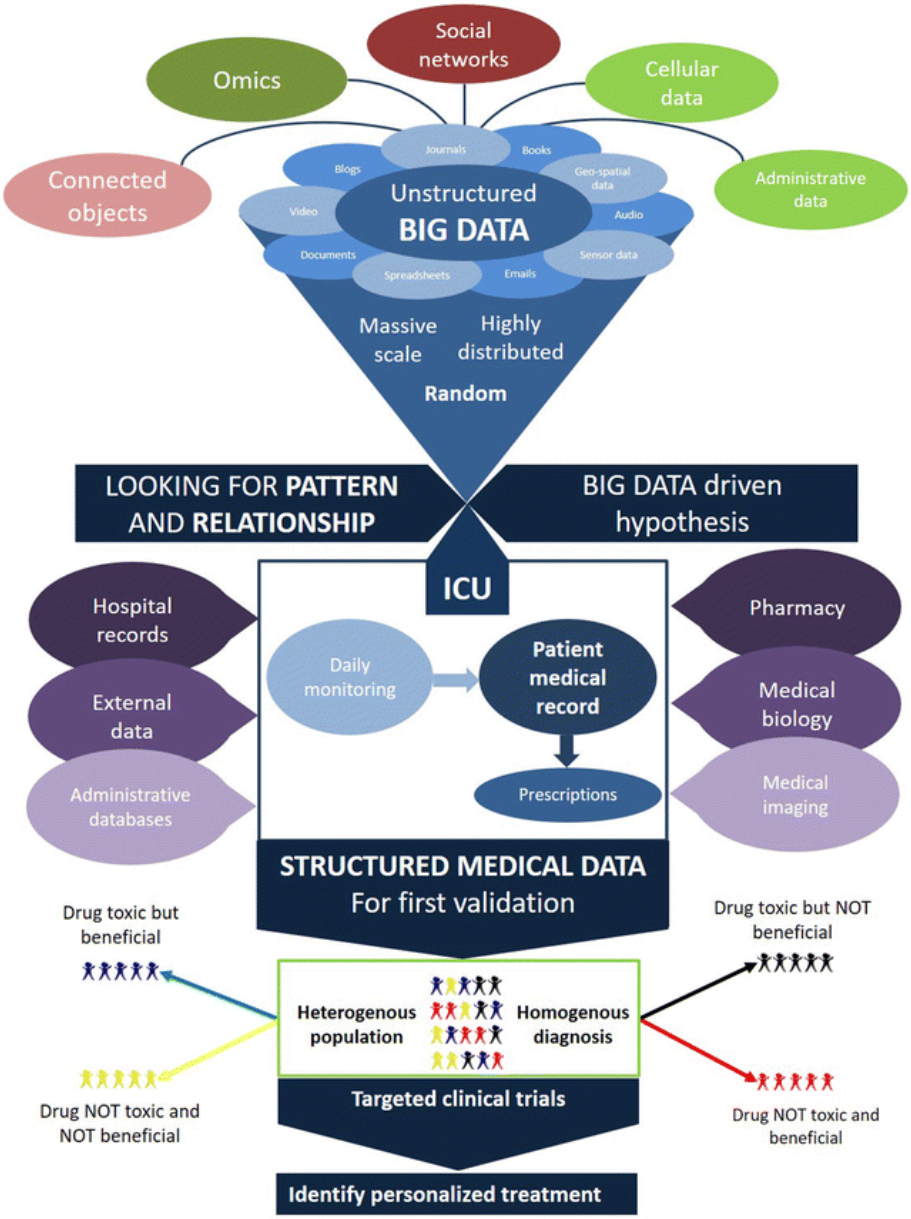

Fig. 1

Big data comprise unstructured data from various origins with different file formats and can be explored by using machine learning methods to define "big data driven hypotheses". These hypotheses can be validated by using structured medical data collected during the ICU stay or by merging ICU data with other medical databases. Finally, it would be possible, from a population of ICU patients with the same diagnosis, to identify specific subgroups who can be included in a targeted randomized trial. This is the way to develop personalized medicine from big data to targeted randomized clinical trial

References

1. Ffoulkes P (2017) InsideBIGDATA guide to the intelligent use of big data on an industrial scale. InsideBIGDATA, Massachusetts

2. Booth CM, Tannock IF (2014) Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. Br J Cancer 110:551–555

3. Apkon M, Singhaviranon P (2001) Impact of an electronic information system on physician workflow and data collection in the intensive care unit. Intensive Care Med 27:122–130

4. Raghupathi W, Raghupathi V (2014) Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2:3

5. Murdoch TB, Detsky AS (2013) The inevitable application of big data to health care. JAMA 309:1351–1352

6. Flechet M, Grandas FG, Meyfroidt G (2016) Informatics in neurocritical care: new ideas for Big Data. Curr Opin Crit Care 22:87–93

7. Simpkin AL, Schwartzstein RM (2016) Tolerating uncertainty—the next medical revolution? N Engl J Med 375:1713–1715

8. Angus DC (2015) Fusing randomized trials with big data: the key to self-learning health care systems? JAMA 314:767–768

9. Mirnezami R, Nicholson J, Darzi A (2012) Preparing for precision medicine. N Engl J Med 366:489–491

10. Perner A, Gordon AC, Angus DC, Lamontagne F, Machado F, Russell JA, Timsit JF, Marshall JC, Myburgh J, Shankar-Hari M, Singer M (2017) The intensive care medicine research agenda on septic shock. Intensive Care Med 43:1294–1305

11. Robins JM, Hernan MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11:550–560

12. Bailly S, Bouadma L, Azoulay E, Orgeas MG, Adrie C, Souweine B, Schwebel C, Maubon D, Hamidfar-Roy R, Darmon M, Wolff M, Cornet M, Timsit JF (2015) Failure of empirical systemic antifungal therapy in mechanically ventilated critically ill patients. Am J Respir Crit Care Med 191:1139–1146

13. Timsit JF, Azoulay E, Schwebel C, Charles PE, Cornet M, Souweine B, Klouche K, Jaber S, Trouillet JL, Bruneel F, Argaud L, Cousson J, Meziani F, Gruson D, Paris A, Darmon M, Garrouste-Orgeas M, Navellou JC, Foucrier A, Allaouchiche B, Das V, Gangneux JP, Ruckly S, Maubon D, Jullien V, Wolff M, EMPIRICUS Trial Group (2016) Empirical micafungin treatment and survival without invasive fungal infection in adults with ICU-acquired sepsis, candida colonization, and multiple organ failure: the EMPIRICUS randomized clinical trial. JAMA 316:1555–1564

14. Guiza F, Van Eyck J, Meyfroidt G (2013) Predictive data mining on monitoring data from the intensive care unit. J Clin Monit Comput 27:449–453

15. Flechet M, Guiza F, Schetz M, Wouters P, Vanhorebeek I, Derese I, Gunst J, Spriet I, Casaer M, Van den Berghe G, Meyfroidt G (2017) AKIpredictor, an online prognostic calculator for acute kidney injury in adult

critically ill patients: development, validation and comparison to serum neutrophil gelatinase-associated lipocalin. Intensive Care Med 43:764–773

16. Bhatt DL, Mehta C (2016) Adaptive designs for clinical trials. N Engl J Med 375:65–74

17. Pocock SJ, Stone GW (2016) The primary outcome fails—what next? N Engl J Med 375:861–870