Big Data and Artificial Intelligence: Will They Change Our Practice?

Joanna Kedra, Laure Gossec

Big Data and Artificial Intelligence:

Will they change our practice?

Authors: Joanna Kedra[a,b], Laure Gossec[a,b]

A : Sorbonne Université, Institut Pierre Louis d'Epidémiologie et de Santé Publique, INSERM, 75646 Paris.

B : Service de rhumatologie, Hôpital Pitié Salpêtrière, AP-HP, 75013 Paris.

Corresponding author:

Dr. Joanna Kedra. Service de rhumatologie, Hôpital Pitié Salpêtrière, AP-HP, 47-83 boulevard de l'Hôpital, 75013 Paris.

@ : jkedra.pro@gmail.com

Tel +33142178421

Fax +33142177959

The corresponding author undertakes to ensure that each author agrees with the entire manuscript and has participated in the study.

Word count: 2366 words, 18/20 references, 2 tables et 1 figure

Key words: big data, artificial intelligence, machine learning, rheumatology, medicine

1. Introduction

The increase of amounts of digital data since the late 1990s has led researchers - in computer science as in health research – to transform the way they perceive and analyze the world. In the field of biology, these same decades have also seen the expansion of massive data in the form of "omics" (genomics, proteomics, metabolomics), which today constitute an important approach in personalized medicine. In this context, it is therefore a question of apprehending new ways of capturing, searching, sharing, storing, analysing and presenting data whose amounts are growing exponentially (1). These so-called Big Data are generally analyzed using artificial intelligence methods, and both concepts are increasingly being applied to medical research.

So what does the rheumatologist need to know about these concepts?

2. Big Data and artificial intelligence: what is it all about?

The term Big Data is defined in the literature in many ways: the general idea is that big data are extremely large, often complex and multidimensional data sets that grow rapidly over time. (2-4) ***(Table 1)***. Sources of medical Big Data are diverse: they can be clinical data from databases such as those of health insurance, private health insurance or cohort data, or digital traces (i.e.: keywords typed into a web server); but also imaging data (as a single imaging exam can contain millions of pixels), or biological data, more particularly those from so-called "omics" approaches (2, 3).

Big Data is often considered to be analysed by methods derived from artificial intelligence and its subspecialty, machine learning ***(Table 1)***. (5, 6) However, artificial intelligence and machine learning can also be used to analyze "small" data sets ("Small Data"). (7) In addition, "classical" statistical methods (such as parametric or non-parametric tests or regression

analyses) can also be used in some studies involving large amounts of data. (2) This shows the lack of a clear consensus on how to analyze Big Data.

3. What applications of Big Data in medicine?

Our systematic literature review on the use of Big Data in rheumatology and other medical specialties and covering 110 articles revealed that Big Data studies in our specialty mainly concern inflammatory joint diseases (N=22, 40%) and osteoarthritis (N=16, 29%) (2). Let's see a few examples.

The Big Data Sjögren Project Consortium is an example of international collaboration, with a database common to 22 countries on 5 continents, and gathering clinical, biological and histological data from 10,500 subjects with Sjögren syndrome. (8) This is the first example of data sharing in the field of autoimmune diseases, with a joint effort to harmonize data architecture and to collaborate with statisticians specialized in Big Data analyses. The objective of this consortium is to deepen knowledge about Sjögren's syndrome, and thus improve its management around the world. To date, this consortium has confirmed the influence of immunological markers on the clinical phenotype of primary Sjögren's syndrome, and in particular the greater correlation between systemic manifestations and hypocomplementemia and cryoglobulinemia compared to anti-nuclear and anti-Ro/La antibodies (8).

The French study ActConnect aimed to evaluate the association between rheumatoid arthritis or spondyloarthritis flares reported by patients and physical activity recorded in real time in number of steps per minute using a physical activity tracer (bracelet) (9, 10, 11). The analyses of the 224,952 hours of physical activity were performed using a machine learning algorithm, and the relapse prediction model thus generated was well correlated with the flares reported by patients (mean sensitivity 96%, mean specificity 97%). This study therefore paves the way

for future studies on remote monitoring of disease activity, with high accuracy and minimal burden on the patient.

The analysis of Internet search histories could also be an interesting source of information: an example is Google Flu Trend, which aims to detect influenza epidemics before they occur, thanks to search peaks on the Google search engine. The use of research histories could thus make it possible to assess the health needs of populations and set up targeted campaigns. Regarding imaging data, an interesting application of artificial intelligence techniques is the automated detection of diabetic retinopathy: based on fundus images, a method applying convolutional neural networks (a machine learning algorithm) has been developed to automatically and simultaneously detect exudates, haemorrhages and microaneurysms (12). This approach has a sensitivity of 0.96, 0.84 and 0.85 respectively to detect the anomalies mentioned above. It is therefore a reliable technique, which also has the advantage of saving the practitioner time. In osteoarticular imaging, some studies using convolutional neural networks have obtained interesting results for the automated diagnosis of myositis on ultrasound (sensitivity 66 to 82%, specificity 65 to 92%) (13).

Omics have revolutionized traditional approaches in biology, and are currently the basis for research in the field of so-called precision medicine. Applications in oncology are diverse: prediction of tumour response in breast cancer based on gene expression and proteomic and micro-RNA analysis (14), identification of genes associated with the diagnosis and prognosis of pancreatic cancer to name but a few examples. In rheumatology, only 15% of recent Big Data articles were based on "omics" data, all of which have been published over the past decade (2). This shows that this is an emerging field in our specialty, but potentially promising, particularly in the early diagnosis and targeted management of autoimmune diseases (15).

4. The limitations of Big Data

Despite its potential interest, Big Data has several limitations to its use ***(Figure 1)***. These limitations have been discussed in the recent EULAR recommendations for the use of Big Data in rheumatology, which are summarized in ***Table 2*** (3). ***Figure 1*** also summarizes some of the solutions proposed by the authors of this article.

The first of these limitations is ethical: indeed, large-scale data are not always properly anonymized, they are often transferred abroad, and the question arises as to who uses them and how they are used. In European Union, the use of personal data has been regulated since 2018 by the General Data Protection Regulations (GDPR), which emphasise in particular the right of users to dispose of their data, the right to erasure, and the need for "explicit" and "positive" consent before any use of data (16). In order to respect the rights and privacy of users, researchers working on Big Data, whether in medical research or any other field, must therefore be aware of and apply these regulations (3).

Another limitation of Big Data is logistical and organizational. Indeed, Big Data sources are numerous, and within each source, data may not be collected in the same way. (17) Moreover, the massive collection of data does not guarantee their quality, stability or consistency. Finally, once the data has been collected, the question arises of being able to store and access it. The recent EULAR recommendations propose, to address these issues, to standardise data collection and to use open data platforms for data storage and access (3).

Another issue related to Big Data is the lack of expertise of researchers in this emerging field. Indeed, our research teams are not yet fully trained in artificial intelligence methods and their interpretation. Additionally, the inappropriate use of these specialized methods could lead to erroneous results, by finding unexpected relationships between two parameters, without being able to prejudge their causal link. (18) In this context, EULAR emphasized the need for

bilateral training and close collaboration between clinicians and data scientists specializing in Big Data, in order to ensure the coherence and relevance of research projects and their results. (3)

5. What opportunities in rheumatology?

Big Data is a source of opportunities in rheumatology. Thus, in the context of personalized medicine, omics approaches could make it possible to predict the response to treatments and thus to adapt the therapeutic management of our patients on a case-by-case basis. The analysis of "digital traces" (i.e. data collected by applications or posted online) could also allow earlier diagnosis of rheumatic diseases, and therefore earlier management. Finally, automated analysis of imaging exams could be useful in cases where it is difficult to decide between inflammatory and mechanical pathology, such as in the analysis of sacroiliacs MRI.

Of course, these are not exhaustive examples, some of which could still be considered "science fiction". Nevertheless, it is undeniable that their implementation in our current practice would transform the diagnostic and therapeutic management of our patients.

6. Conclusion

Big Data is by its very nature an extremely vast, dynamic and constantly evolving field, which has the potential to transform medical research but which can also have an impact on our daily practice. Indeed, Big Data implies that we must change the way we collect, store and analyse data, while guaranteeing an ethical framework that respects international regulations. Big Data also enables innovative and promising diagnostic and therapeutic applications. However, this is still an emerging area, this is why it is important, as rheumatologists or researchers, that we take ownership of this field to better implement it in the care of our patients.

Disclosures of interest:

The authors have no disclosures of interest to declare.

Acknowledgements:

None.

.

References

[1] Hajirahimova MS, Aliyeva AS. About Big Data Measurement Methodologies and Indicators. International Journal of Modern Education and Computer Science. 2017; 9(10): 1–9.

[2] Kedra J, Radstake T, Pandit A, et al. Current status of the use of Big Data and Artificial Intelligence in RMDs: a systematic literatura review informing EULAR recommendations. RMD Open 2019 (On Press)

[3] Gossec L, Kedra J, Servy H, et al. European League Against Rheumatism points to consider for the use of big data in rheumatic and musculoskeletal diseases. Ann Rheum Dis. Epub ahead of print: [27/06/2019]. doi:10.1136/ annrheumdis-2019-215694

[4] HMA-EMA Joint Big Data Taskforce: summary report. https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf [accessed Feb 16, 2019].

[5] Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach (3rd ed.). Upper Saddle River, NJ: Prentice Hall 2009.

[6] Koza JR, Bennett FH, Andre D, Keane MA. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In: Gero JS, Sudweeks F, eds. Artificial Intelligence in Design. Dordrecht (NL): Elsevier Academic Publishers 1996.

[7] Pasini A. Artificial neural networks for small dataset analysis. J Thor Dis. 2015; 7(5): 953–960.

[8] Brito-Zerón P, Acar-Denizli N, Ng WF, et al. How immunological profile drives clinical phenotype of primary Sjögren's syndrome at diagnosis: analysis of 10,500 patients (Sjögren Big Data Project). Clin Exp Rheumatol. 2018;36 Suppl 112(3):102-112.

[9] Gossec L, Guyard F, Leroy D, et al. Detection of flares by decrease in physical activity, collected using wearable activity trackers, in rheumatoid arthritis or axial spondyloarthritis: an application of Machine-Learning analyses in rheumatology. Arthritis Care Res (Hoboken). 2018 doi: 10.1002/acr.23768.

[10] Jacquemin C, Servy H, Molto A, et al. Physical Activity Assessment Using an Activity Tracker in Patients with Rheumatoid Arthritis and Axial Spondyloarthritis: Prospective Observational Study. JMIR Mhealth Uhealth. 2018;6(1):e1.

[11] Jacquemin C, Maksymowych WP, Boonen A, Gossec L. Patient-reported Flares in Ankylosing Spondylitis: A Cross-sectional Analysis of 234 Patients. J Rheumatol. 2017;44(4):425-430.

[12] Khojasteh P, Aliahmad B, Kumar DK. Fundus images analysis using deep features for detection of exudates, hemorrhages and microaneurysms. BMC Ophthalmol. 2018;18(1):288.

[13] Burlina P, Billings S, Joshi N, Albayda J. Automated diagnosis of myositis from muscle ultrasound: Exploring the use of machine learning and deep learning methods. PLoS One. 2017;12(8):e0184059.

[14] Xia F, Shukla M, Brettin T, et al. Predicting tumor cell line response to drug pairs with deep learning. BMC Bioinformatics. 2018;19(Suppl 18):486.
[15] Chocholova E, Bertok T, Jane E, et al. Glycomics meets artificial intelligence - Potential of glycan analysis for identification of seropositive and seronegative rheumatoid arthritis patients revealed. Clin Chim Acta. 2018;481:49-55.
[16] GDPR Key Changes with the General Data Protection Regulation – EUGDPR https://eugdpr.org/the-regulation/ [accessed Dec 2 2018].
[17] Townend D. Conclusion: harmonisation in genomic and health data sharing for research: an impossible dream? Hum Genet. 2018;137(8):657-664.
[18] Price WN. Big data and black-box medical algorithms. Sci Transl Med. 2018;10:471.

Table 1: Definitions

Concept	Definition
Big Data	The term "big data" refers to extremely large datasets, which can be complex, multidimensional, unstructured, can come from heterogeneous sources and can accumulate rapidly. Big Data can come from a variety of sources, including clinical, biological, social or environmental sources. (3)
Artificial Intelligence	Ability of a machine to "mimic" human cognitive functions such as learning or problem solving. (5)
Machine Learning	Artificial intelligence subspecialty that uses statistical approaches to give computers the ability to learn from data, i.e. to improve their performance without being explicitly programmed for that. (6)

Table 2: Summary of EULAR recommendations for the use of Big Data in rheumatology (3)

Issue	Number of recommendations	Summary of the recommendations
Data collection	2	- Data harmonization - Application of « FAIR*» principles
Data storage	2	- Application of « FAIR*» principles - Use of open data plateforms
Ethics	1	Privacy by design
Interdisciplinary collaboration	1	Collaboration between researchers, data scientists, clinicians and patients
Statistical analyses	2	- Need of clear reporting of the analytical methods - Benchmarking of the methods
Validation of models and results	1	Validation of the models and results before their implementation in practice
Implementation of the results in daily practice	1	Implementation of the results by researchers
Interdisciplinary training	1	Interdisciplinary training or both health professionals and data scientists

*FAIR : « Findable, Accessible, Interoperable, Reusable »

NB : the sum of recommendations is 11 instead of 10 because the 2nd recommendation refers to both data collection and storage.

Figure 1: limitations and potential solutions in Big Data studies