RMD Open
Rheumatic & Musculoskeletal Diseases

ORIGINAL ARTICLE

Current status of use of big data and artificial intelligence in RMDs: a systematic literature review informing EULAR recommendations

Joanna Kedra,[1,2] Timothy Radstake,[3] Aridaman Pandit,[3] Xenofon Baraliakos,[4] Francis Berenbaum,[5] Axel Finckh,[6] Bruno Fautrel,[1,2] Tanja A Stamm,[7] David Gomez-Cabrero,[8] Christian Pristipino,[9] Remy Choquet,[10] Hervé Servy,[11] Simon Stones,[12] Gerd Burmester,[13] Laure Gossec[1,2]

To cite: Kedra J, Radstake T, Pandit A, *et al*. Current status of use of big data and artificial intelligence in RMDs: a systematic literature review informing EULAR recommendations. *RMD Open* 2019;**5**:e001004. doi:10.1136/rmdopen-2019-001004

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136rmdopen-2019-001004)

Received 9 May 2019
Revised 26 June 2019
Accepted 29 June 2019

For numbered affiliations see end of article.

Correspondence to
Dr Joanna Kedra;
jkedra.pro@gmail.com

ABSTRACT

Objective To assess the current use of big data and artificial intelligence (AI) in the field of rheumatic and musculoskeletal diseases (RMDs).

Methods A systematic literature review was performed in PubMed MEDLINE in November 2018, with key words referring to big data, AI and RMDs. All original reports published in English were analysed. A mirror literature review was also performed outside of RMDs on the same number of articles. The number of data analysed, data sources and statistical methods used (traditional statistics, AI or both) were collected. The analysis compared findings within and beyond the field of RMDs.

Results Of 567 articles relating to RMDs, 55 met the inclusion criteria and were analysed, as well as 55 articles in other medical fields. The mean number of data points was 746 million (range 2000–5 billion) in RMDs, and 9.1 billion (range 100 000–200 billion) outside of RMDs. Data sources were varied: in RMDs, 26 (47%) were clinical, 8 (15%) biological and 16 (29%) radiological. Both traditional and AI methods were used to analyse big data (respectively, 10 (18%) and 45 (82%) in RMDs and 8 (15%) and 47 (85%) out of RMDs). Machine learning represented 97% of AI methods in RMDs and among these methods, the most represented was artificial neural network (20/44 articles in RMDs).

Conclusions Big data sources and types are varied within the field of RMDs, and methods used to analyse big data were heterogeneous. These findings will inform a European League Against Rheumatism taskforce on big data in RMDs.

Key messages

What is already known about this subject?

► Big data and artificial intelligence are rapidly evolving fields with the potential to profoundly modify RMD research and ultimately, patient care.

What does this study add?

► This literature review showed the variety of big data sources, and the important heterogeneity in the methods used to analyse big data, with more than seven different methods, in rheumatology and in other medical fields.

How might this impact on clinical practice or future developments?

► These findings provide a current status, which will inform a European League Against Rheumatism taskforce on big data in RMDs.

INTRODUCTION

There are tremendous opportunities for health research propelled by the recent expansion of technology aimed to apply 'big data/real world data' for clinical decision support.[1 2] The growth in quantity and improvement in quality of data; the changing dynamic and scale of data collection from various sources, including health records and omics[3 4]; and the fast development in measurements, analytic methods and parallel computing of large amounts of clinical, biological and imaging data promise to dramatically transform clinical medicine and biomedical science.[5] In addition, the exponential growth in the number of publicly traded companies in this field indicates the economic potential, achievability and feasibility of digital healthcare.[5–7]

Although promising, big data raises many issues. To address these in the context of rheumatic and musculoskeletal diseases (RMDs), a European League Against Rheumatism (EULAR) taskforce was set up in 2018; in this context, information was needed on the current status of big data in the literature. The mains issues of interest included the definition of big data and the number of

datapoints corresponding to big data.[8–10] A clear definition of big data is needed as it appears that there is no consensual description, and that the meaning of this term has evolved during the last decades. The concept of big data was first defined in 1997 as 'data sets that are too large or complex for traditional data-processing application software to adequately deal with'[11]; more recently, the European Medicines Agency (EMA) defined big data as 'extremely large datasets which may be complex, multi-dimensional, unstructured and heterogeneous, which are accumulating rapidly and which may be analysed computationally to reveal patterns, trends and associations'. This definition also mentions the requirement of advanced or specialised methods to provide an answer within reliable constraints.[12] Another important issue with big data is how to collect these data—in other terms, what are the sources of big data, and how many datapoints are concerned?[13] Indeed, as mentioned above, big data may be obtained from various kinds of sources, from clinical to biological data. At present, there is no clarity where big data in the field RMDs comes from. Finally, another major question is the analysis of big data, since the use of traditional statistical methods may be difficult or inappropriate giving the complex nature of these data; new statistical methods derived from artificial intelligence (AI), such as machine learning, are often applied to big data.[14 15] However, their exact use in medicine and the different types of analyses are unknown.[16] We aimed to assess the current status of big data both in RMDs, and for comparison purposes, in other medical fields.

The objective of this systematic literature review (SLR) was to obtain an overview of the existing literature on big data in RMDs, to inform a EULAR taskforce.[17] In addition, to compare the current status of big data in RMDs with other medical fields, a 'mirror' review outside of RMDs was also performed.

MATERIAL AND METHODS

For the SLR in RMDs as well as for the mirror review in other medical fields, standardised methods were applied.[18]

Search strategy

The SLR was performed on PubMed MEDLINE on 21st of November 2018 and updated on 19th of February 2019. The key words ('big data' (All Fields) OR 'Artificial Intelligence' (MeSH Terms)) were combined with ('musculoskeletal diseases' (MeSH Terms) OR 'musculoskeletal diseases' (All Fields) OR 'rheumatology' (MeSH Terms) OR 'rheumatology' (All Fields)). The following filters were applied: English language publications and studies performed in humans. The resulting articles were included if they reported use of big data (as defined by the articles' authors) in RMDs and were original articles.

The mirror review outside RMDs was performed also in PubMed MEDLINE on 28th of November 2018 and updated on 20th of February 2019. The key words ('big data' (All Fields) OR 'Artificial Intelligence' (MeSH Terms)) and NOT RMDs were used, with the same filters as above. The resulting articles were included if they reported use of big data (as defined by the authors) in healthcare but outside RMDs, and were original articles. Since this search was performed to obtain a mirror non-systematic review outside RMDs, we performed it in a retro-chronological way, including the most recent articles corresponding to our criteria, up to the same number of articles as found by the SLR in RMDs.

One reviewer (JK) assessed titles and abstracts for suitability for inclusion in review, according to the pre-determined inclusion criteria, followed by full-text review (online supplementary table 1). Support from coauthors was provided, in particular when data scientist skills were needed.

Data extraction

Data were extracted to answer the following questions: (1) the current definition of big data; (2) data sources of big data; and (3) type of analysis used to deal with big data.

To answer the question of the current definition of big data, on the one hand, definitions provided or referenced in the included articles were collected[12]; on the other hand, the number of data points in the paper was reported. The number of data could refer to number of units of observation (eg, number of patients or number of MRI analysed) or the number of data point, it is to say the set of one or more measurements on a single member of unit of observation, if provided.[19]

To answer the question of data sources of big data, we collected data sources and types and classified them into the following categories: clinical data; data provided by registers or cohorts; electronic health records; claims databases; trials or patient generated Health Data; biological data (including various kinds of -omic data); imaging data; and other kinds of data (including data provided by text mining from publications).

To answer the question of analyses of big data, the statistical methods used were collected and classified into traditional statistical methods and AI methods. Traditional statistical methods refer to techniques such as non-parametric statistics, χ^2 test, Student's t-test, linear or logistic regression, survival analyses, longitudinal analyses or trajectory modelling.[20 21] Among AI methods, heuristics and machine learning were separated and machine learning methods were classified as follows[16]: Artificial Neural Networks (including Deep Learning), Support Vector Machine, Random Forests, Natural Language Processing, k-Nearest Neighbors and Bayesian models.[22 23]

For descriptive purposes, we also collected study characteristics (year of publication, impact factor, country of origin of the first author), underlying disease (in RMDs) or specialty (outside RMDs).

Data analysis

Findings were described and then compared between those covering RMDs and those from other medical

Figure 1 Flow-chart of the systematic literature review in RMDs.

fields by non-parametric statistics. A comparison of statistical methods used to analyse big data was also performed in subgroups of data sources (clinical vs other sources), using exact Fisher's test, and the results were considered significant if the p value was below 0.05. Meta-analysis was not appropriate; potential selection bias was not accounted for.

In June 2019, a sensitivity analysis was performed using key words referring specifically to the different rheumatic diseases and to AI methods. This sensitivity analysis aimed to assess the additional number of articles that would be found using additional and more specific key words, without using the papers by extracting data from the relevant papers.

RESULTS

Paper selection and general characteristics

The flow chart of the SLR in the field of RMDs is shown in figure 1. In RMDs, of 567 abstracts, 55 original articles met the inclusion criteria, and we screened 313 additional articles to include 55 articles outside of RMDs. The flow chart of the literature review in other medical fields is provided in online supplementary figure 1.

A general description of the 110 articles is provided in table 1. The mean year of publication was 2014 for the RMDs SLR, with 72% of the articles published between 2013 and 2018 (figure 2); whereas the articles included in the mirror non-RMD review were all published in 2018 or 2019. In the field of RMDs, first authors were mostly from North America (38%), whereas the distribution was homogenous between North America, Europe and Asia outside of RMDs (respectively, 31%, 33% and 34%). More details on the selected articles are provided in online supplementary tables 2 and 3.

Table 1 Description of 55 articles on big data in RMDs, and 55 articles for comparison outside RMDs

	RMDs	Other medical fields
Year of publication, mean (SD) (range)	2014 (4.6) (1992–2018)	2018 (0.4) (2018–2019)
Year of publication: last 5 years, N (%)	40 (72)	55 (100)
Impact factor, mean (SD) (range)	3.8 (4.0) (0.35–23.3)	5.56 (9.8) (0.56–47.7)
Geographic origin of the first author, N (%)		
North America N=15 (34%)	21 (38)	17 (31)
Europe	18 (33)	18 (33)
Asia	15 (27)	19 (34)
Australia	0 (0)	1 (2)
South America	1 (2)	0 (0)
Africa	0 (0)	0 (0)
Clear definition of big data, N (%)	2 (4)	7 (13)
N data analysed		
Units of observation, mean (SD) (range)	1 142 000 (3 990 000) (5–25 000 000)	5 298 000 (23 909 000) (40–140 000 000)
Data points, mean (SD) (range)	746 000 000 (1 660 000 000) (2000–5 000 000 000)	9 149 000 000 (39 000 000 000) (100 000–200 000 000 000)
Clinical data sources, N (%)	26 (47)	17 (31)
Registries/cohorts, N (%)	14 (25)	10 (18)
EHR	11 (20)	3 (6)
Claims databases	1 (2)	0 (0)
Trials	0 (0)	0 (0)
PGHD (sensors, etc)	0 (0)	4 (7)
Other	0 (0)	0 (0)
Biological data, N (%)	8 (15)	17 (31)
–omics	8 (15)	17 (31)
Other	0 (0)	0 (0)
Imaging, N (%)	16 (29)	16 (29)
Other data sources	5 (9)	5 (9)
Text-mining from publications	5 (9)	5 (9)
Other	0 (0)	0 (0)

EHR, Electronic Health Record; PGHD, Patient Generated Health Data.

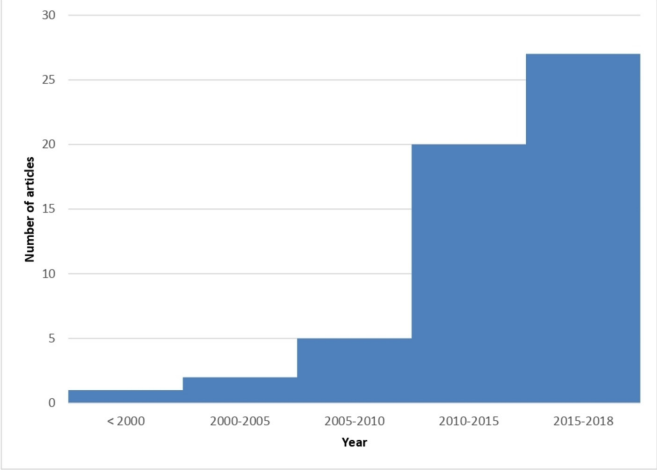

Figure 2 Evolution of the number of articles on big data in the field of RMDs.

Among RMDs, the most represented fields were inflammatory joint diseases (N=22, 40%) and osteoarthritis (N=16, 29%); other studies were on gout (N=3, 5%), osteoporosis (N=6, 11%) spine pathology (N=6, 11%) and individual pathologies not pertaining to one of these categories (N=5, 9%). The three most represented diseases were: knee osteoarthritis (N=13, 24%), rheumatoid arthritis (N=12, 22%) and postmenopausal osteoporosis (N=6, 11%).

Outside of RMDs, the most represented medical fields were: oncology (N=14, 25%), neurology (N=8, 15%), infectious diseases (N=6, 11%), ophthalmology (N=5, 9%) and psychiatry (N=5, 9%). More details are provided in table 2.

Definition of big data

Only two articles in the field of RMDs (4%) and seven articles out of the field of RMDs (13%) mentioned a clear definition of big data (table 1). Overall, 53 articles in RMDs (96%) provided a number of units of observation, and 15 (27%) provided a number of data points, whereas outside of RMDs, 52 articles (95%) provided a number of units of observation and 26 (47%) a number of data points. The mean number of data points was 746 million (2000–5 billion) in RMDs, and 9.1 billion (range 100 000–200 billion) outside of RMDs. Even if the mean number of units of observation in the SLR and in the mirror review was higher than 1 million, small numbers of units of observation were also observed; however, they corresponded to imaging data, which actually provided a huge number of data points (eg, five CT-scans in RMDs provide more than 26 million data points).[24]

Sources of big data

In RMDs, big data were mostly obtained from clinical data sources (N=6, 47%), whereas outside of RMDs, the distribution was quite homogenous between clinical, biological and imaging sources (respectively, 31%, 31% and 29%): table 1.

Table 2 Description of the diseases in RMDs and other medical fields

RMDs	
Pathology	**N (%)**
Gout	3 (5)
Inflammatory joint diseases	22 (40)
Myositis	*1 (2)*
Psoriatic arthritis	*1 (2)*
Rheumatoid arthritis	*12 (22)*
Sjögren's syndrome	*4 (7)*
Spondyloarthritis	*2 (4)*
Systemic lupus erythematosus	*2 (4)*
Systemic sclerosis	*1 (2)*
Vasculitis	*2 (4)*
Osteoarthritis	16 (29)
Osteoarthritis of the knee	*13 (24)*
Other location	*3 (5)*
Osteoporosis	6 (11)
Post-menopausal osteoporosis	*6 (11)*
Other cause of osteoporosis	*0 (0)*
Spine pathology	6 (11)
Other pathologies	5 (9)

Other medical fields	
Specialty	**N (%)**
Cardiology	2 (4)
Dermatology	1 (2)
Endocrinology	1 (2)
Genetics	2 (4)
Gerontology	1 (2)
Gynaecology	4 (7)
Hepatology and gastroenterology	3 (5)
Immunology	1 (2)
Infectious diseases	6 (11)
Neurology	8 (15)
Oncology	14 (25)
Ophtalmology	5 (9)
Pharmacology	3 (5)
Physiology	3 (5)
Psychiatry	5 (9)
Pulmonology	1 (2)

The total of pathologies in RMDs and in other medical fields is above 55, as some articles were about several diseases at the same time.

Table 3 Description of the statistical methods used to analyse big data in RMDs and in other medical fields

	RMDs	Other medical fields
General description		
AI only, N (%)	30 (55)	35 (63)
Traditional statistics only N (%)	10 (18)	8 (15)
Both AI and traditional methods N (%)	15 (27)	12 (22)
AI methods		
AI articles, N (%)	**45 (82)**	**47 (85)**
AI methods		
Machine learning, N	44	47
Other, N	2	0
Mention of supervision, N	**4**	**18**
Supervised	4	14
Unsupervised	0	7
Semi-supervised	0	1
Not reported	41	29
Types of machine learning method, N		
Not specified	4	3
Artificial Neural Networks	20	24
Deep Learning	*5*	13
Support Vector Machine	10	8
Random Forests	7	13
Natural Language Processing	7	2
k-Nearest Neighbors	3	6
Bayesian models	3	5
Traditional methods		
Traditional statistics, N (%)	**25 (45)**	**20 (36)**
Regression methods, N	11	15
Other methods, N	16	21

Supervised learning refers to the machine learning task of learning a function that maps an input to an output based on example input–output pairs, and unsupervised learning to the ability to learn without a 'teacher'.
Of note, one article in RMDs used both machine learning and another AI method (heuristic).
AI, artificial intelligence.

Statistics and AI

The methods used to analyse big data are reported in table 3. Both traditional and AI methods were used to analyse big data. In RMDs, 30 (55%) articles used traditional analysis methods and 10 (18%) articles used AI methods, while outside RMDs 35 (63%) articles used traditional analysis methods and 8 (15%) articles used AI methods. Of note, some articles used both methods (respectively, 15 (27%) and 12 (22%) articles). Within and outside of the field of RMDs, AI methods were used respectively in 45 (82%) and 47 articles (85%), and consisted of machine learning methods respectively in 98% and 100% of the articles. In both reviews, the most used machine learning method was Artificial Neural Network (N=20 in RMDs and 24 out of RMDs; respectively 44% and 51% of AI articles). Overall, four RMDs and three non-RMDs articles did not describe the kind of AI method which was used to analyse big data, which represents respectively 9% and 6% of AI articles. Usual statistical methods were exclusively used in 10 articles (18%) in the SLR and eight articles (15%) in the mirror review; respectively, 11 (20%) and 15 (27%) articles reported regression methods.

In RMDs, usual statistical methods were significantly more used to analyse clinical data than to analyse other data sources (N=8 and N=2, respectively, p=0.035); similar results were found in other medical fields (N=6 for clinical data, N=2 for other data sources, p=0.008).

Sensitivity analysis

At the time of the sensitivity analysis, 1051 articles were found using the algorithm of the SLR. Adding key words related to specific RMDs brought 27 additional articles (2.6%) and key words related to AI methods led to 71 additional articles (6.8%).

DISCUSSION

This review has brought to light important information on the current status of big data in RMDs and in other medical fields. Only a few authors clearly defined what they meant by 'big data', and the provided definitions were quite different. There were also disparities in the number of data found in the articles. Data sources were varied, and were mostly clinical in RMDs; whereas they were equally clinical, biological or radiological in the very recent publications outside RMDs. Both traditional and AI methods were used to analyse big data, and among machine learning methods, the most represented was artificial neural networks, independently of medical field and data source.

This study has strengths and weaknesses. First, only one reviewer performed the screening of the articles and extracted information from the selected articles. However, support was provided by coauthors, especially when data scientist skills were needed. Second, the research was only performed on PubMed MEDLINE with a language restriction. Ideally, multiple bibliographical databases would have been searched. The key words used to perform the SLR were restrictive. To assess the impact of the choice of key words, a sensitivity analysis showed that the use of additional key words referring to RMDs by name and/or to specific AI methods would have brought less than 10% more articles. We believe this indicates the validity of the present findings. Moreover, the aim was to obtain an overview rather than an exhaustive view of the topic. In the future, SLR in individual fields such as imaging, computational biology or clinical research, could be considered. Indeed, imaging and the other big data sources are very wide fields, and specific researches in each of them would have certainly found additional articles. We did not perform a SLR outside of the field of RMDs, however, this part of the work was only

meant to be used as a comparison with RMDs and not to be exhaustive; furthermore, giving the fact that 33 794 articles related to big data out of RMDs were found using our key words, performing a SLR would not have been feasible. Choosing the right key words for this research was an issue, and some references in the field of RMDs were not picked up, such as the ActConnect Study[25]; however, we used the best keywords available at that time, and even if a MeSH term was created in 2019 for 'big data', all articles using this concept are not referenced yet under this MeSH term. Finally, classification of statistical methods, in particular machine learning methods, may be discussed. However, it was based on accepted classifications, which are considered reliable references in this field.[22 23]

A key finding from this review is that there is no consensual definition of big data. First defined as data sets too large or complex for traditional analysis methods,[11] this concept has evolved and the '5 V' paradigm (for volume, velocity, veracity, variety and value) is more and more used.[26–28] The definition provided in recent EMA recommendations may be considered as a synthesis of all these notions.[12] Although all the authors of the selected articles agree that big data refers to a very large number of data points, there is also no consensual 'cut-off' to define what is meant by 'very large'. Some authors proposed log(n×p) superior or equal to 7 (n being the number of units of observation and p the number of variables),[29] however, giving the rapid growth of datasets in the last decade, some authors rather propose to think in terms of terabytes (10^{12}) or petabytes (10^{15}).[9 10] Nevertheless, even terabytes and petabytes will be soon too restrictive, since according to an International Data Corporation report prediction, the global data volume will grow exponentially from 4.4 zettabytes to 44 zettabytes (10^{21}) between 2013 and 2020.[30] This issue shows that the definition of big data is beyond the scope of the characteristics of data type and cannot be restricted to the size or volume of those data.[31] It confirms also the disparity in the number of data reported in the studies. Thus, the amount of data is not the same when considering clinical or imaging data, since a single radiological exam can contain millions of pixels, and some imaging techniques such as MRI can also contain several images or sequences; this point makes complex the estimation of the number of datapoints in imaging. However, beyond their volume, what makes big data a challenge is their complexity based on heterogeneity, multidimensionality and the fact that they are dynamic—in other terms, all the previous single dimensions are dynamically connected. None of the selected articles addressed clearly these issues, despite studying complex connexions between heterogeneous, multidimensional and dynamic data offer unparalleled opportunities to personalise medicine.

In this review, clinical data were the most frequent source of big data in RMDs, whereas the distribution was more spread out outside of RMDs. This could be explained by the fact that, except clinical and radiological data, other data sources may not be so well implemented in rheumatology, whereas outside rheumatology, the literature review only picked up extremely recent articles (due to the retrochronological approach) and omics are a rapidly evolving field. With the increasing amount of information collected by registries, Electronic Health Records and the increasing use of sensors collecting in real time patients' data, clinical research must evolve to take advantage of these new sources of information and implement them in routine practice.[2 32] Given the exhaustive nature of clinical big data, they could be particularly interesting in the future to study rare diseases, rare outcomes and evaluate the efficacy of treatments in non-selected populations, which are difficult to assess in usual clinical trials.[2 33] Omics is a growing field, particularly promising for personalised medicine as it supports the discovery of predictive biomarkers and therapeutic targets.[3 4 34] Imaging is also a very interesting application of big data for diagnosis and clinical decision making. Since any single radiological exam compiles a huge amount of data, medical imaging is particularly conducive to the use of AI and notably machine learning methods.[35 36] Examples of applications of big data in medical imaging are numerous and varied, from diagnosis and follow-up of cancers,[37 38] to scoliosis[39] or diabetic retinopathy.[40] As social networks and Internet-driven data are exponentially growing, text mining is becoming a relevant source for health information: recent applications were the prediction of influenza and pertussis epidemics thanks to Google searches[41 42] and prediction of depression thanks to Facebook statuses.[43] In the field of RMDs, only one paper included in the SLR was based on Google, Wikipedia and Youtube searches concerning inflammatory vasculitis,[44] contrasting with the variety of examples provided in other medical fields. This could be explained by the fact that rheumatic diseases and symptoms are less common than the examples cited above or that RMD specialists are less aware of these methods at this time. However, it is probable that this novel source of information will play an important role in health research in the years to come, particularly given the increasing focus on patient-driven and community-driven research.[45]

This work revealed the use of various statistical methods, traditional or AI-related, to analyse big data. The use of usual statistical methods for big data, such as χ^2, Student's t-test or logistic regression may seem paradoxical, because big data may be too complex to be analysed with these tools.[11] However, we found out that more than 20% of articles related to big data used traditional statistical methods (27% in RMDs and 22% in other medical fields). This may be because of lack of knowledge of AI methods, or because traditional statistics allowed to answer the clinical questions.[46] Thus, in cohorts or registries, the number of patients is higher than the number of variables collected for each of them, so even if it is big data by 'the number', it may not be too complex for usual methods. Another possibility is that more specific methods such as AI are not yet well

implemented in every research unit. Most of the selected articles used AI, and in particular machine learning methods; indeed, these methods seem more relevant to analyse huge and complex data, such as genomic or imaging-driven data.[35 47] Moreover, these methods are tolerant of poor quality of underlying data,[48] which is a common issue in big registries, were missing data are frequent. However, the risk of an inappropriate use of these methods is creating quantitative fallacy and over fitting models which could not be generalisable in clinical practice.[49] That is why AI and machine learning algorithms should be validated and regulated to be integrated in medical practice.[50] In the present review, between 5% and 10% of the selected articles did not mention explicitly the kind of machine learning algorithm which was used, indicating a need for better reporting.

In conclusion, this work gives an overview of the current status of big data in RMDs, and in medicine in general. Data sources and types are varied, and methods used to analyse them are heterogenous and not always well reported. This variety of sources and methods hold promises for potential applications of big data in rheumatology and in other medical fields, and may lead to a major change in health research for the years to come.

Author affiliations
[1]Institut Pierre Louis d'Epidémiologie et de Santé Publique (iPLESP), UMR S 1136, Sorbonne Universite, Paris, France
[2]Rheumatology Department, Hôpital Universitaire Pitié Salpêtrière, APHP, Paris, France
[3]Department of Rheumatology, Clinical Immunology and Laboratory for Translational Immunology, University of Utrecht Faculty of Medicine, Utrecht, The Netherlands
[4]Herne, Ruhr-University, Rheumazentrum Ruhrgebiet, Bochum, Germany
[5]Rheumatology Department, Hospital Saint-Antoine, APHP, Paris, Île-de-France, France
[6]Division of Rheumatology, University Hospital of Geneva, Geneva, Switzerland
[7]Section for Outcomes Research, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria
[8]Departamento de Salud-Universidad Pública de Navarra, Translational Bioinformatics Unit, Navarra Biomed, Pamplona, Spain
[9]Interventional Cardiology Department, Ospedale San Filippo Neri, Rome, Italy
[10]Orange e-Health, INSERM U1142, Paris, France
[11]e-Health Services, Sanoïa, Gardanne, France
[12]School of Healthcare, University of Leeds, Leeds, West Yorkshire, UK
[13]Department of Rheumatology and Clinical Immunology, Charité - University Medicine Berlin, Berlin, Germany

Contributors All authors have provided data for the study, participated in the data interpretation and have approved the final version.

Funding Supported by the European League Against Rheumatism, EULAR (grant SCI018).

Competing interests RC is an employee of Orange Healthcare, and HS is an employee of Sanoïa, a Digital CRO providing clinical research services including data science. There are no competing interests for the other authors.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

REFERENCES

1. Dixon WG, Michaud K. Using technology to support clinical care and research in rheumatoid arthritis. *Curr Opin Rheumatol* 2018;30:276–81.
2. Misra DP, Agarwal V. Real-World evidence in rheumatic diseases: relevance and lessons learnt. *Rheumatol Int* 2019;39:403–16.
3. PY W, Cheng CW, Kaddi CD, *et al*. Omic and electronic health record big data analytics for precision medicine. *IEEE Trans Biomed Eng* 2017:64263–73.
4. Suwinski P, Ong C, Ling MHT, *et al*. Advancing personalized medicine through the application of whole exome sequencing and big data analytics. *Front Genet* 2019;10:49.
5. Zhang X, Pérez-Stable EJ, Bourne PE, *et al*. Big data science: opportunities and challenges to address minority health and health disparities in the 21st century. *Ethn Dis* 2017;27:95–106.
6. Hoyt RE, Snider D, Thompson C, *et al*. IBM Watson analytics: automating visualization, descriptive, and predictive statistics. *JMIR Public Health Surveill* 2016;2:e157.
7. Chen Y, Elenee Argentinis JD, Weber G. IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research. *Clin Ther* 2016;38:688–701.
8. Vogel C, Zwolinsky S, Griffiths C, *et al*. A Delphi study to build consensus on the definition and use of big data in obesity research. *Int J Obes* 2019;390.
9. Eisenstein M. Big data: the power of petabytes. *Nature* 2015;527:S2–S4.
10. Schofield P. Big data in mental health research - do the ns justify the means? Using large data-sets of electronic health records for mental health research. *BJPsych Bull* 2017;41:129–32.
11. Cox M, Ellsworth D. Managing big data for scientific visualization. In: *ACM SIGGRAPH '97 course #4, exploring gigabyte datasets in real-time: algorithms, data management, and time-critical design*. Anaheim, CA, US, Los Angeles: ACM Digital Library, 1997: 5–17.
12. HMA-EMA Joint Big Data Taskforce. Available: https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf [Accessed 16 Feb 2019].
13. Alonso SG, de la Torre Díez I, Rodrigues JJPC, *et al*. A systematic review of techniques and sources of big data in the healthcare sector. *J Med Syst* 2017;41:183.
14. ICSU-IAP-ISSC-TWAS working group. Open data in a big data world. An international accord. Available: https://twas.org/sites/default/files/open-data-in-big-data-world_short_en.pdf [Accessed 16 Feb 2019].
15. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med* 2016;375:1216–9.
16. Saria S, Butte A, Sheikh A. Better medicine through machine learning: what's real, and what's artificial? *PLoS Med* 2018;15:E1002721.
17. Gossec L, Kedra J, Servy H, *et al*. EULAR points to consider for the use of big data in rheumatic and musculoskeletal diseases. *Ann Rheum Dis* 2019:annrheumdis-2019-215694.
18. The Cochrane Collaboration. Cochrane Handbook for Systematic Reviews of Interventions. Available: https://training.cochrane.org/handbook [Accessed 16 Feb 2019].
19. Wikipedia. Unit of observation. Available: https://en.wikipedia.org/wiki/Unit_of_observation [Accessed 16 Feb 2019].
20. Barkan H. Statistics in clinical research: important considerations. *Ann Card Anaesth* 2015;18:74–82.
21. Krousel-Wood MA, Chambers RB, Muntner P. Clinicians' guide to statistics for medical practice and research: Part I. *Ochsner J* 2006;7:3–7.
22. Wakefield K. SAS insights - A guide to machine learning algorithms and their applications. Available: https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html [Accessed 16 Feb 2019].
23. Towards Data Science. Machine learning. Available: https://towardsdatascience.com/machine-learning/home [Accessed 16 Feb 2019].
24. Abidin AZ, Deng B, DSouza AM, *et al*. Deep transfer learning for characterizing chondrocyte patterns in phase contrast X-ray computed tomography images of the human patellar cartilage. *Comput Biol Med* 2018;95:24–33.
25. 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* 2018.
26. Genovese Y, Prentice S. Pattern-based strategy: getting value from big data. *Gartner* 2011 June 17. Available: https://www.gartner.com/doc/1727419/patternbased-strategy-getting-value-big [Accessed 16 Feb 2019].
27. Jin X, Wah BW, Cheng X, *et al*. Significance and challenges of big data research. *Big Data Res* 2015;2:59–64.
28. Moscatelli M, Manconi A, Pessina M, *et al*. An infrastructure for precision medicine through analysis of big data. *BMC Bioinformatics* 2018;19:351.
29. Baro E, Degoul S, Beuscart R, *et al*. Toward a Literature-Driven definition of big data in healthcare. *Biomed Res Int* 2015;2015:639021
30. Sh. Hajirahimova M, S. Aliyeva A. About big data measurement methodologies and indicators. *Int J Mod Educ Comp Sci* 2017;9:1–9.
31. Lee CH, Yoon H-J. Medical big data: promise and challenges. *Kidney Res Clin Pract* 2017;36:3–11.
32. Mayo CS, Matuszak MM, Schipper MJ, *et al*. Big data in designing clinical trials: opportunities and challenges. *Front Oncol* 2017;7:187.
33. Monti S, Grosso V, Todoerti M, *et al*. Randomized controlled trials and real-world data: differences and similarities to untangle literature data. *Rheumatology* 2018;57(Suppl 7):vii54–8.
34. Topol EJ. The big medical data miss: challenges in establishing an open medical resource. *Nat Rev Genet* 2015;16:253–4.
35. Morris MA, Saboury B, Burkett B, *et al*. Reinventing radiology: big data and the future of medical imaging. *J Thorac Imaging* 2018;33:4–16.
36. Landewé RBM, van der Heijde D. "Big Data" in rheumatology: intelligent data modeling improves the quality of imaging data. *Rheum Dis Clin North Am* 2018;44:307–15.
37. Wang L, Yang S, Yang S, *et al*. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. *World J Surg Oncol* 2019;17:12.
38. Park HJ, Kim SM, La Yun B, *et al*. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound: added value for the inexperienced breast radiologist. *Medicine* 2019;98:e14146.
39. Duong L, Cheriet F, Labelle H. Automatic detection of scoliotic curves in posteroanterior radiographs. *IEEE Trans Biomed Eng* 2010;57:1143–51.
40. Khojasteh P, Aliahmad B, Kumar DK. Fundus images analysis using deep features for detection of exudates, hemorrhages and microaneurysms. *BMC Ophthalmol* 2018;18:288.
41. Gianfredi V, Bragazzi NL, Mahamid M, *et al*. Monitoring public interest toward pertussis outbreaks: an extensive Google Trends-based analysis. *Public Health* 2018;165:9–15.
42. Zhang Y, Bambrick H, Mengersen K, *et al*. Using Google trends and ambient temperature to predict seasonal influenza outbreaks. *Environ Int* 2018;117:284–91.
43. Eichstaedt JC, Smith RJ, Merchant RM, *et al*. Facebook language predicts depression in medical records. *Proc Natl Acad Sci U S A* 2018;115:11203–8.
44. Fleurence RL, Beal AC, Sheridan SE, *et al*. Patient-powered research networks aim to improve patient care and health research. *Health Aff* 2014;33:1212–9.
45. Bragazzi NL, Watad A, Brigo F, *et al*. Public health awareness of autoimmune diseases after the death of a celebrity. *Clin Rheumatol* 2017;36:1911–7.
46. Tan SS-L, Gao G, Koch S. Big data and analytics in healthcare. *Methods Inf Med* 2015;54:546–7.
47. He Y, Jiang Z, Chen C, *et al*. Classification of triple-negative breast cancers based on Immunogenomic profiling. *J Exp Clin Cancer Res* 2018;37.
48. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol* 2016;13:350–9.
49. Kerr D, Klonoff DC. Digital diabetes data and artificial intelligence: a time for humility not hubris. *J Diabetes Sci Technol* 2019;13:123–7.
50. Price WN. Big data and black-box medical algorithms. *Sci Transl Med* 2018;10:471.