EULAR points to consider for the use of big data in rheumatic and musculoskeletal diseases

Laure Gossec ,[1,2] Joanna Kedra ,[1,2] Hervé Servy,[3] Aridaman Pandit ,[4] Simon Stones,[5] Francis Berenbaum ,[6] Axel Finckh,[7] Xenofon Baraliakos,[8,9] Tanja A Stamm ,[10] David Gomez-Cabrero ,[11] Christian Pristipino,[12] Remy Choquet,[13] Gerd R Burmester,[14] Timothy R D J Radstake[4]

Handling editor Professor Josef S Smolen

For numbered affiliations see end of article.

Correspondence to
Professor Laure Gossec, Institut Pierre Louis d'Epidémiologie et de Santé Publique (iPLESP), UMR S 1136, Sorbonne Universite, Paris 75013, France; laure.gossec@gmail.com

LG and JK contributed equally.

Received 10 May 2019
Revised 7 June 2019
Accepted 7 June 2019
Published Online First 22 June 2019

ABSTRACT

Background Tremendous opportunities for health research have been unlocked by the recent expansion of big data and artificial intelligence. However, this is an emergent area where recommendations for optimal use and implementation are needed. The objective of these European League Against Rheumatism (EULAR) points to consider is to guide the collection, analysis and use of big data in rheumatic and musculoskeletal disorders (RMDs).

Methods A multidisciplinary task force of 14 international experts was assembled with expertise from a range of disciplines including computer science and artificial intelligence. Based on a literature review of the current status of big data in RMDs and in other fields of medicine, points to consider were formulated. Levels of evidence and strengths of recommendations were allocated and mean levels of agreement of the task force members were calculated.

Results Three overarching principles and 10 points to consider were formulated. The overarching principles address ethical and general principles for dealing with big data in RMDs. The points to consider cover aspects of data sources and data collection, privacy by design, data platforms, data sharing and data analyses, in particular through artificial intelligence and machine learning. Furthermore, the points to consider state that big data is a moving field in need of adequate reporting of methods and benchmarking, careful data interpretation and implementation in clinical practice.

Conclusion These EULAR points to consider discuss essential issues and provide a framework for the use of big data in RMDs.

INTRODUCTION

The recent expansion of big datasets and advanced computational techniques led to tremendous opportunities for health research.[1] As elegantly elaborated by Topol, the use of big data in medicine is going to disrupt the medical system as we know it.[2] Big data include both clinical data (eg, originating from electronic health records, healthcare system claims data or patient-generated data such as from apps), biological data issued from the development of molecular research leading to multi-omics complex molecular data,[3] social data (eg, originating from social networks, Internet of Things, physical social connexions or economic data repositories), imaging data and environmental data (eg, urbanistic data, pollution or atmospheric conditions).[4 5] In parallel, artificial intelligence–based methodologies allowing computer systems to 'learn' from data (ie, progressively improve performance on a specific task without being explicitly programmed) are more and more accessible.[6 7] The collection of big data combined with such information processing techniques (computational modelling, machine learning) led to an opportunity for progress in medical research, which should ultimately modify patient care and clinical decision-making.

Some recent applications of big data show interesting potential. These include the correct detection of skin lesions suspect of melanoma,[8–10] prediction of cancer treatment response based on imaging[11] and the correct interpretation of eye fundus pathologies.[11] However, big data is an emergent area in need of guidelines and general recommendations on how to move this field forward in a collaborative and ethical way. Some of the challenges presented by big data and artificial intelligence include data sources and data collection: how to collect and store the data, while guaranteeing ethics and data privacy[12]; how to interpret data models of complex analyses[13 14]; and what are the clinical implications of big data: how to go from big data to clinical decision-making.[3 15 16]

To our knowledge, no academic societies have developed consensus guidelines dealing with big data.[17] Very recently, the European Medicines Agency (EMA) released recommendations focused on the acceptability of evidence derived from big data in support of the evaluation and supervision of medicines by regulators[18]; however, these recommendations deal mainly with the interpretation of drug-related big data. The European League Against Rheumatism (EULAR) has recently formulated as one of its key strategic objectives, the advancement of high-quality collaborative research and comprehensive quality of care for people living with rheumatic and musculoskeletal disorders (RMDs).[19] Thus, EULAR naturally takes an interest in big data and its applications.

The objective of this project was to develop EULAR 'points to consider' (PTC) for the collection, analysis and use of big data in RMDs.

METHODS

After approval by the EULAR Executive Committee, the convenors (LG, TRDJR) and the project fellow (JK) led a multidisciplinary task force guided by the 2014 updated EULAR Standardised Operating

To cite: Gossec L, Kedra J, Servy H, *et al*. *Ann Rheum Dis* 2020;**79**:69–76.

Procedures,[20] which were modified for this specific task force. In October 2018, the main questions to be addressed in the preparatory work for the task force were defined as (1) data sources and collection, (2) data analyses, and (3) data interpretation and implementation of findings. These questions were addressed in subsequent months leading up to the face-to-face meeting by the project fellow and the convenors. A systematic literature review (SLR) was performed between November 2018 and February 2019, regarding publications employing big data in RMDs, with a comparison in other medical fields.[21] Additionally, a narrative review of unpublished data on websites on big data and artificial intelligence was performed to inform the task force[12 17 18 22–26] and expert opinions were obtained from four selected persons through individual telephone interviews.

In February 2019, during a 1-day face-to-face task force meeting, overarching principles and PTC were developed. The process was both evidence based and consensus based, through discussions of the international task force of experts from a range of disciplines including computer science and artificial intelligence. The task force consisted of 14 individuals from 8 European countries: 6 rheumatologists, 4 data scientists/big data experts, 1 cardiologist specialised in systems medicine, 1 patient research partner, 1 health professional with expertise in outcomes research and 1 fellow in rheumatology. Furthermore, feedback was obtained after the meeting from two additional experts. This inclusive approach aimed to obtain broad consensus and applicability of the PTC. During the 1-day meeting, the preparatory work was presented and discussed, the target audience of the PTC was defined, then the PTC were formulated and extensively discussed. The PTC were finalised over the subsequent 2 weeks by online discussions, taking into account the publication the same week of an EMA consensus document on big data.[18]

During the meeting and through online discussions, based on the gaps in evidence and the issues raised among the task force, a research agenda was also formulated. After the PTC were finalised, the level of evidence and strength of each PTC were ascertained according to the Oxford system.[27] Finally, each task force member voted anonymously on their level of agreement with each PTC via email (numeric rating scale ranging from 0=do not agree to 10=fully agree). The mean and SD of the level of agreement of task force members were calculated.

The final manuscript was reviewed and approved by all task force members and approved by the EULAR Executive Committee.

RESULTS

Target audience

The target audience of these PTC includes researchers in the field of big data in RMDs and researchers outside the field of RMDs; data collection organisations and/or groups collecting data (eg, registries, hospitals, telecommunications operators, search engines, genetic sequencing teams, institutions which collect images etc); data analysts and organisations; people with RMDs, people at risk of developing RMDs and patient associations; clinicians involved in the management of people with RMDs; and other stakeholders such as research organisations and funding agencies, policy-makers, authorities, governments and medical societies outside of RMDs.

Overarching principles and PTC were formulated, which are shown in table 1 and are discussed in detail below.

Box 1 Some definitions of the terms 'big data' in the literature

- Extremely large sets of information which require specialised computational tools to enable their analysis and exploitation. These data might come from electronic health records from millions of patients, genomics, social media, clinical trials or spontaneous adverse reaction reports.[18]
- Data sets that are too large or complex for traditional data-processing application software to adequately deal with.[73]
- Defined by volume, if log(n∗p) is superior or equal to 7, where n is number of rows and p is number of columns.[74]
- Data sets that are large or complex (multidimensional and/or dynamic) enough to apply complex methods, eg, artificial intelligence.[75]
- Information assets characterised by such high velocity, variety and volume that specific data mining methods and technology are required for its transformation into value.[76]
- A generic and comprehensive definition of big data is based on the five V paradigm, ie, volume of data, variety of data, velocity of processing, veracity and value.[77]
- The term big data refers to the emerging use of rapidly collected, complex data in such unprecedented quantities that terabytes (10^{12} bytes), petabytes (10^{15} bytes) or even zettabytes (10^{21} bytes) of storage may be required.[78]

Definitions of terms

This first point in table 1 proposes a definition of terms relating to big data. Although the term big data is widely used, there is not one commonly accepted definition. When performing the literature review, several definitions were found (box 1).[6 21] The first overarching principle defines the term big data, largely based on the EMA definition.[18] Big data is defined by its size and diversity—it is diverse, heterogeneous and large and incorporates multiple data types and forms; but also by the specific complexity and challenges of integrating the data to enable a combined analysis.[18] The second half of the definition refers to artificial intelligence (AI). AI is defined as the ability of a machine to mimic 'cognitive' functions that humans associate with human minds, such as 'learning' and 'problem solving'.[6] New computational techniques, such as AI (which includes machine learning and deep learning) are often (but not necessarily) applied to big data.[18]

This next sentence is informative and aims to present the diversity of data sources leading to big data; we listed in a non-exhaustive way some of the sources of big data. The most common sources of big healthcare data found in the SLR were clinical; these include electronic health records, studies and registries, billing and healthcare system claims databases.[21 28 29] A more recent source of clinical big data currently underused in RMDs is the Internet of Things (eg, wearables, apps, medical devices and sensors), but also social media, behavioural and environmental data.[18 30 31] Imaging is also a growing field of application of big data.[10 32 33] Regarding basic and translational research results, -omics such as genomics and bioanalytical omics are an important and rapidly growing field for big data.[18 34]

Overarching principles

Overarching principle A: ethical aspects

This overarching principle addresses ethical issues with big data. The collection, analysis and implementation of big data in RMDs must adhere to all applicable regulations. This covers

privacy, confidentiality and security, ownership of data, data minimalisation, and flow of data within the EU and with third countries.[22 35] This is both a regulatory and legal requirement, and an ethical one.[12] In terms of legal requirements, the General Data Protection Regulation (GDPR) has set standards which apply across Europe, but for health-related data, national rules could also apply on top of these.[12]

In this overarching principle, we also raise the question of the role of the patient and/or carer in big data. Big data enables active participation of patients, but this is not always the case. Participation of patients and patient research partners can be helpful in data interpretation; for big data, the active participation of patients is still a field to be explored.[36] This principle highlights issues around information, consent and responsibilities, and also patient rights and participation.[35]

B: Potential of big data

Big data provides unprecedented opportunities which we wished to highlight in this overarching principle. Maybe even more than other types of data, big data benefits from transversal thinking, by both original 'outside the box' approaches and cross-fertilisation approaches taking into account other medical fields and aspects such as comorbidities, psychological, sociological and environmental findings.[18] In this regard, collaboration both within the RMD field and in particular with patients, and outside of RMDs, is key, as will be addressed later in these PTC.[15 24 37]

C: Ultimate goal

This overarching principle states that the ultimate goal is to be of benefit to people with RMDs. This is always a key priority of EULAR and is in keeping with the EULAR Strategic Objectives and Roadmap.[19 38]

Points to consider

Table 1 provides the level of evidence, strength of recommendation and level of agreement for each of the 10 PTC.[20 27]

PTC 1: data collection—use of standards

As the amount of big data increases, the need for data harmonisation becomes more apparent, with the possibility for using different data sources through application of global standards. It is essential to ensure that existing and future datasets can be used and, in particular, pooled for big data approaches. To this end, they must be harmonised/aligned to facilitate interoperability of data.[18] Where possible, minimising the number of standards and using global data standards would be helpful; as stated by the EMA, standards should be transparent, open to promote widespread uptake and globally applicable.[18]

In that regard, international consensus efforts such as data standards, developed by groups such as the International Consortium for Health Outcomes Measures, International Council for Harmonisation, Health Level Seven International, International Organization for Standardization and Clinical Data Interchange Standards (to name a few) are useful.[39–42] Some of these groups have developed standards for rheumatology.[40] The EULAR dataset for rheumatoid arthritis registries, or other core sets, are also helpful in this regard.[43 44] While these standards regulate the way in which the data are recorded and stored, they do not control how efficient the data collection is at the care team level.

Table 1 EULAR-endorsed overarching principles and points to consider for the use of big data in RMDs, with levels of agreement and for the specific points, levels of evidence and strength

Definitions			
The term 'big data' refers to extremely large datasets which may be complex, multidimensional, unstructured and from heterogeneous sources, and which accumulate rapidly. Computational technologies, including artificial intelligence (eg, machine learning), are often applied to big data. Big data may arise from multiple data sources including clinical, biological, social and environmental data sources			
Overarching principles	**LoA, mean (SD)**		
A. For all big data use, ethical issues related to privacy, confidentiality, identity and transparency are key principles to consider	9.6 (0.7)		
B. Big data provides unprecedented opportunities to deliver transformative discoveries in RMD research and practice	9.5 (1.2)		
C. The ultimate goal of using big data in RMDs is to improve the health, lives and care of people including health promotion and assessment, prevention, diagnosis, treatment and monitoring of disease	9.6 (0.5)		
Points to consider	**LoA, mean (SD)**	**LoE**	**SoR**
1. The use of global, harmonised and comprehensive standards should be promoted to facilitate interoperability of big data	9.7 (0.6)	4	C
2. Big data should be Findable, Accessible, Interoperable and Reusable (FAIR principle)	9.6 (0.9)	5	D
3. Open data platforms should be preferred for big data related to RMDs	8.7 (1.2)	5	D
4. Privacy by design must be applied to the collection, processing, storage, analysis and interpretation of big data	9.6 (0.5)	4	C
5. The collection, processing, storage, analysis and interpretation of big data should be underpinned by interdisciplinary collaboration, including biomedical/health/life scientists, computational and/or data scientists, relevant clinicians/health professionals and patients	9.7 (0.6)	4	C
6. The methods used to analyse big data must be reported explicitly and transparently in scientific publications	10 (0)	4	C
7. Benchmarking of computational methods for big data used in RMD research should be encouraged	9.4 (1.2)	5	D
8. Before implementation, conclusions and/or models drawn from big data should be independently validated	9.1 (0.7)	4	C
9. Researchers using big data should proactively consider the implementation of findings in clinical practice	9.3 (0.8)	5	D
10. Interdisciplinary training on big data methods in RMDs for clinicians/health professionals/health and life scientists and data scientists must be encouraged	9.7 (0.6)	5	D

Numbers in the column 'LoA' indicate the mean and SD (in parentheses) of the LoA, as well as the mean agreement of the 14 task force members on a 0–10 scale. LoE and strength based on the Oxford Centre for Evidence-Based Medicine classification, with 'Level 1' corresponding to meta-analysis or randomised controlled trials (RCTs) or high-quality RCTs; 'Level 2' to lesser quality RCT or prospective comparative studies; 'Level 3' to case–control studies or retrospective studies; 'Level 4' to case series without the use of comparison or control groups; 'Level 5' to case reports or expert opinion.[27]
LoA, level of agreement; LoE, level of evidence; RMD, rheumatic and musculoskeletal disorder; SoR, strength of recommendation.

PTC 2: data collection and storage—FAIR principle

The FAIR (Findable, Accessible, Interoperable and Reusable) data principles are a measurable set of principles intended to act as a guideline to enhance the reusability of their data.[45] The FAIR principles are recognised by many actors, including the EMA and the EU Commission.[18 22 24 46] The FAIR principles are strongly linked to PTC 1 and 3, referring to standardisation, interoperability and data storage. Efforts are ongoing to promote the FAIR principles, such as those of the EU commission through the development of the EU eHealth Digital Service Infrastructure.[47]

PTC 3: data storage—data platforms

Several platforms have been developed to facilitate big data projects. These platforms are independent, standardised, collaborative and not at all limited to use for RMDs.[48–50] These platforms have been developed with financial support from the EU and therefore adhere to necessary standards. Hence, the use of such platforms should be promoted as recently stated by the EMA.[18] In these PTC, we refer to the use of such platforms for RMD big data, but of course this would also apply to other groups of big data.

Public access to data is an important point, which raised much debate within the task force. Internationally, several groups emphasised the principle that big data should be made publicly available to promote open and reproducible research; in particular, when the data are publicly funded.[18 26 51 52] On the contrary, downsides of public access to data are the potential loss of momentum to secure intellectual property and scientific publications from the researchers who initially generated the data.[53] Given this controversy, data sharing should be achieved in a way that is sustainable for all parties involved.[53] How to make data but also algorithms openly available is very complex.[54 55] The task force consensus was in favour of accessible data, but in the current situation, with limited and supervised access; we also felt that pilot projects to assess the impact of data sharing are needed and that such data sharing should be evidence based.[56] This consensus will need to be revised as the situation evolves. The topic of data sharing was also added to the research agenda.

PTC 4: privacy by design

Privacy by design is an important approach which should be followed when managing big data projects. This point insists on the importance of privacy by design at the different levels of big data use, including the collection, processing, storage, analysis and interpretation of big data.[17 57] Privacy by design is directly quoted in EU law about personal data.[12] This approach prompts thinking on the reasons you collect/gather, process, store and protect data, from inception to final deletion. Privacy by design also prompts individuals to self-assess the potential risks or weaknesses relating to data, and how best to manage such risks. This PTC is a major challenge for researchers in big data, but it appeared to the task force to be a legal requirement or an ethical one, and also an educational one since this practice is not widely understood. For big data projects, the data source is key: either the data are collected for the purpose of the project or data are re-used from existing sources. In the first case, obtaining consent is mandatory and must involve a data officer and follow a transparent and effective process in terms of data governance.[35] When data are re-used, the national laws on consent, data sharing and governance must be applied. In this context, the development of common principles for data anonymisation would facilitate data sharing, including regulations for sharing, de-identifying, securely storing, transmitting and handling personal health information.[18]

The European regulatory framework around data is currently undergoing change: from May 2019, the circulation of non-identifying data will be facilitated.[47] The implications of this change will have to be assessed.

PTC 5: collaboration

While interdisciplinary collaboration is beneficial and required for all research projects, it is even more important in big data projects where expertise is dispersed among different stakeholders. The task force insisted on the importance of collaboration between appropriate stakeholders at the analysis stage, for example, where AI methods require appropriate expertise, and at all phases of a big data project.[25] Interdisciplinary collaborations should intervene at different times across a project, to enable the most appropriate design to be chosen, while ensuring that data collection and the type of analysis are fit for purpose. Of note, the statistical methods may be based on AI or may include more traditional statistics and/or computational methodologies, as appropriate. Further knowledge is needed on the comparison of statistical methods, which is discussed in more detail in PTC 7.[21 58] The appropriate individuals to collaborate include clinical/biological scientists, computational/data scientists, health professionals and patients; proposals for respective roles are shown in table 2.

PTC 6: data analyses reporting

The methods, parameters and tools used in big data processing must be reported explicitly in any scientific paper. This is pivotal to allow comparison and interpretation of findings. Our SLR found that 8% of papers using AI did not report in any way what artificial intelligence methods were being used.[21] Proper reporting is important for all research, but even more so when innovative methods such as artificial intelligence are used, to avoid confusion and to promote reproducibility.[14 18 30 59]

PTC 7: benchmarking of data analyses

AI encompasses several techniques which are intended to solve the most difficult problems in computer science: search and optimisation (heuristics), logic (fuzzy logic), uncertain reasoning and learning (machine learning).[60] In our SLR, machine-learning methods were the most used AI techniques in RMDs and in other medical fields (98% and 100% of AI papers, respectively). The most used machine-learning algorithms were artificial neural networks (with deep learning as the most advanced version), representing 48% of AI articles.[21 61]

In addition, comparison of artificial intelligence methods within RMDs should be promoted.[17 18 24 62] This is particularly needed because AI is a rapidly growing field; there is an ongoing and unsolved debate as to which methods within AI perform best.[63 64] The comparison of AI methods was also added to the research agenda since it was felt that this particular topic was difficult to perform at this moment in time and was more aspirational.

PTC 8: validation of big data findings

Although there may be a perception that big data are more valid or less subject to bias than traditional studies, model overfitting, inappropriate generalisation of the results and/or bias can in fact lead to inappropriate conclusions.[14 18 28] Thus, it is important both to assess and benchmark the quality of the generated data and the methods used to avoid overinterpretation of results,

Table 2 Stakeholders involved in big data research: proposal of potential roles

Stakeholder	Characteristics	Potential role in big data research
Clinicians/health professionals, biomedical/health/life scientists	Knowledge of the diseases, prognosis and treatments	Clinically relevant question, study protocol, data collection, interpretation and implementation of findings
Data scientist	To analyse and interpret complex digital data, should be proficient in a broad spectrum of analytical methodologies that encompass traditional (biostatistics, epidemiology, discrete-event simulation and causal modelling) as well as emerging methods [67]	Provide early guidance on the best tools or algorithm to analyse the data Analyses of data and interpretation
Computational biologist	Involved in the development and application of data-analytical and theoretical methods, mathematical modelling and computational simulation techniques to the study of biological, ecological, behavioural and social systems. Has domain knowledge in biology	Provide early guidance on the best tools or algorithm to analyse the data Analyses of data and interpretation
Data protection officer	Expert on data privacy	Orient the project team in privacy by design practice
Patients, carers, patient research partners and patient associations	People living with RMDs who have knowledge of day-to-day life with RMDs, from diagnosis to treatment and long-term management	Participation in all stages of the study, from the protocol to the interpretation of the findings
Database expert	Expert of the data in a database	Help the project team to understand the real 'value' of data in a database, and provide guidance on data selection
Computer sciences expert	Expert in computer sciences solutions	Provide guidance on the best technical solution to manage the big data, from its collection to massive calculation solutions

RMD, rheumatic and musculoskeletal disorder.

overfitting of the models and generalisation of the results when using big data. The task force also felt that it was important to validate results in independent datasets.[24 28] Overall, the task force agreed that conclusions drawn from big data need independent validation (in other datasets) to overcome current limitations and to assure scientific soundness. However, a specific challenge for big datasets and the validation of results is the need for other (similar) big datasets—thus, feasibility of validation is a key issue which was discussed at length within the task force.

PTC 9: implementation of findings

The clinical implementation of big data findings should be considered at the earliest opportunity. The SLR and from literature showed that this implementation is currently mostly lacking.[21 65] The task force consensus was that researchers using big data should consider implementation of their results in clinical practice; this would include, for example, discussing implementation of findings in clinical practice in the original papers. The task force is well aware that this is a difficult task; such implementation being both complex to set up, costly, and potentially not within the scope of the primary study.[66] In this regard, the EMA states that regulatory guidance is required on the acceptability of evidence derived from big data sources.[18 67] However, taking all these limitations into account, the task force consensus was that implementation of findings should be proactively considered early on.

PTC 10: training

Interdisciplinary training for clinical, biological or imaging researchers, healthcare professionals and computational biologists/data scientists in the field of big data is important and links closely with the need for collaborations in the field of big data (table 2). Indeed, machine-learning methods are becoming ubiquitous and have major implications for scientific discovery[26]; however, healthcare professionals are not perfectly aware of the correct use of these methods, whereas data scientists may lack the clinical knowledge to design studies and interpret the findings (table 2). Given the current relative lack of expertise related to big data in the field of RMDs, and given the rapid changes in this field, certain organisations should set up or facilitate training sessions.[18 37] This may include academic institutes, public research bodies and international organisations, such as EULAR. The training is needed for both sides: the healthcare professionals needing to learn about the basics of big data, and the data scientists needing to better understand the clinical questions and context within which big data have been collected, and/or is being applied.[68] The training can be performed separately for the different stakeholders, but in some instances, it will require an interdisciplinary educational setting in order to engage multidisciplinary teams and their unique dynamics (eg, the need to set a common vocabulary). The training process should detect skills gaps, identify individuals with bioinformatics/biostatistics/analytics/data science expertise within or outside the field of RMDs and implement appropriate training. The training should also aim for different levels of education provision, ranging from academic taught modules (undergraduate and postgraduate), academic research modules (PhD) and continuous professional development opportunities (eg, through seminars and workshops). Similar efforts can be observed in Systems Biology and Systems Medicine.[18 68–70]

Research agenda

Based on the discussions among the task force and the areas of uncertainty identified within the SLR and discussions among expert stakeholders, a research agenda has been proposed, depicted in table 3. This research agenda covers issues related to data collection, data analyses, training, interpretation of findings and implementation of findings.

DISCUSSION

These are the first EULAR-endorsed PTC for the use of big data within the field of RMDs, which could well be applied by other medical disciplines. These PTC address the core aspects of big data, namely data sources and storage, including ethical aspects, data analyses, data interpretation and implementation. Legal aspects are not clearly mentioned, but these PTC were meant

Table 3 Research agenda

Theme	Research point
Data sources	Leverage EULAR legacy initiatives around core datasets that should be collected in research (and usual care) as foundations for successful big data projects in the field of RMDs
	Determine the optimal use of eHealth data through digital traces and patient-generated/patient-reported data
	Determine the potential use of database linkages, such as healthcare system claims databases
Data access	Identify the mechanisms supporting and implications following open access to, and sharing of, big data
	Assess positive and negative aspects of data sharing in terms of article impact (academic/social) and translational success
	Identify the challenges, opportunities and solutions for international data sharing
	Develop a repository of privacy rules in different European countries
	Identify public platforms for data and how the public can access their own data within big data sets for knowledge/education/self-management purposes
Analyses	Evaluate and compare statistical methods and benchmarking of big data
	Develop methods of assessment and minimisation of bias and of generalisation/reproducibility
	Determine the most appropriate open source tools to improve reproducibility of the results
	Perform a critical assessment of statistical significance vs clinical relevance of the results obtained from medical big data
Reporting	Stimulate consistent reporting of big data studies using validated reporting guidelines
	Stimulate and facilitate open sharing of codes/scripts
Implementation	Determine the value of algorithms and big data findings in terms of quality of care and cost effectiveness
	Assess levels of evidence in evidence-based medicine when based on big-data studies
	Manage the potential rapid and frequent changes of outcomes when implementing big data findings
Training	Identify opportunities for training via the EULAR School of Rheumatology and other relevant organisations
	Assess the importance of inter and cross-disciplinarity
	Assess the place of multidisciplinary training at specific stages of individual careers and/or at specific stages of specific projects
	Consider introducing a basic big data/systems biology/bioinformatic course at bachelors' levels for healthcare professionals
Collaborations	Stimulate national and international interest among the data scientist community in relation to RMDs
	Promote the integration of RMD fluent 'ethical experts' in collaborative teams working on big data
Ethics and roles	Stimulate ethical and moral discussions with patients and 'data donors' specifically in the context of big data, addressing topics such as informed consent/assent, confidentiality, anonymity and privacy concerns, particularly with regards to the re-use of the data
	Discuss the roles and responsibilities of healthcare professionals, scientists/researchers and patients in relation to big data
	Assess issues pertaining to commercial use of big data, particularly involving public–private consortiums and the use of multiple datasets
	Assess the effects of big data results on use of drugs including in unauthorised/compassionate use cases
	Define the role, modalities and rules of patient engagement in the generation and exploitation of big data

RMD, rheumatic and musculoskeletal disorder.

to cover principles and practical aspects of big data; however, the law, and in particular GDPR, applies first.[12] For the update of these points to consider in a few years, participants with legal and ethical expertise should be considered.

This consensus effort is original and should help to promote growth and alignment in the field of big data. However, we are aware that this is a rapidly moving field and that the present PTC may quickly become outdated. It is reassuring that our proposals were not in contradiction to other recent recommendations, such as those of the EMA or the National Health Service in the UK.[17 18]

To our knowledge, no other non-governmental organisation representing patients, healthcare professional and scientific societies to date has developed recommendations for big data. While the American College of Rheumatology has not published specific guidance relating to big data, it has developed an online patient registry from electronic health records which could potentially be used as a big data source.[71]

The use of big data is rapidly expending as witnessed by the increasing number of organisations, companies and publications/books dealing with this topic. Undoubtedly, the exploration, use and implementation of big data provide opportunities to improve healthcare, but it is also clear that this field is in need for guidelines and criteria. These PTC are a first tool to set those guidelines. With the growth of big data in RMDs, we expect that these PTC inspire governmental and research organisations, healthcare providers, researchers and patients to increase relevant training of the stakeholders, promotes research on interpretation and clinical applications of big data results, and develop benchmarks/guidelines for reproducible research.

Points 8 and 9 referring to validation and implementation raised much debate within the task force since we felt it was important to both insist on the importance of these steps and at the same time aim for applicability/feasibility of the points to consider. The final formulation of the points was thought to encourage progress without being too directive, to allow researchers to move forward as needed. Such elements will have to be updated as more data become available.

The grading of the evidence was a challenge in the present work as the Oxford level of evidence[27] which is used in EULAR task forces is better adapted to therapeutic evidence than to observational or prognostic evidence as is often obtained in big data work. However, according to EULAR Standardized Operating Procedures,[20] levels of evidence and strength of recommendations should be rated by the Oxford Levels of Evidence. Moreover, in the case where there is little data-driver evidence, EULAR Standardized Operating Procedures recommend to downgrade the recommendations to the level of 'points to consider', which is what was performed here.

This work has several limitations: the main one is that the present PTC are not specific to RMDs. However, they are not specific because the aspects of big data that they address are

universal, and at present, there is no specific issue related to big data in RMDs, as is also the case in any other medical specialty. Moreover, the experts we consulted consider big data as an opportunity to go beyond the traditional division of medical specialties and allow multidisciplinary approaches. The other main limitation was the extremely low level of evidence for all the PTC, raising the question of the interest of evidence in this specific field where the PTC were expert driven. This is often the case on subjects where recommendations are formulated before supportive data are produced.[72] It is linked to the novelty of the subject.

In conclusion, it is anticipated that new data in this rapidly moving field will emerge over the next few years and that some of the questions formulated in the research agenda will be answered. Therefore, we will consider an update of these PTC as needed in a few years.

Author affiliations
[1]Institut Pierre Louis d'Epidémiologie et de Santé Publique, INSERM, Sorbonne Universite, Paris, France
[2]APHP, Rheumatology Department, Pitie Salpetriere Hospital, Paris, France
[3]Sanoïa, e-Health services, Gardanne, France
[4]Dept of Rheumatology, Clinical Immunology and Laboratory of Translational Immunology, Universitair Medisch Centrum Utrecht, Utrecht, The Netherlands
[5]School of Healthcare, University of Leeds, Leeds, UK
[6]Rheumatology, St Antoine Hospital, Sorbonne Université, INSERM, Paris, France
[7]Division of Rheumatology, University of Geneva, Geneva, Switzerland
[8]Rheumazentrum Ruhrgebiet Sankt Josefs-Krankenhaus, Herne, Germany
[9]Ruhr-Universitat Bochum, Bochum, Germany
[10]Section for Outcomes Research, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria
[11]Translational Bioinformatics Unit, Navarra Biomed, Departamento de Salud-Universidad Públicade Navarra, Pamplona, Navarra, Spain
[12]Ospedale San Filippo Neri, Rome, Italy
[13]Orange Healthcare, INSERM U1142, Paris, France
[14]Rheumatology and Clinical Immunology, Charité University Hospital, Berlin, Germany

Correction notice This article has been corrected since it published Online First. The equal contribution statement has been added.

Contributors All authors have contributed to this work and have approved the final version.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests LG has published a study for which Orange IMT (telecommunications company) performed machine-learning analyses, without charge to the author. HS is an employee of Sanoïa, Digital CRO providing clinical research services including data science. RC is an employee of Orange Healthcare. There are no competing interests for the other authors.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Laure Gossec http://orcid.org/0000-0002-4528-310X
Joanna Kedra http://orcid.org/0000-0003-3535-3183
Aridaman Pandit http://orcid.org/0000-0003-2057-9737
Francis Berenbaum http://orcid.org/0000-0001-8252-7815
Tanja A Stamm http://orcid.org/0000-0003-3073-7284
David Gomez-Cabrero http://orcid.org/0000-0003-4186-3788

REFERENCES

1 Saria S, Butte A, Sheikh A. Better medicine through machine learning: what's real, and what's artificial? *PLoS Med* 2018;15.
2 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.
3 Dixon WG, Michaud K. Using technology to support clinical care and research in rheumatoid arthritis. *Curr Opin Rheumatol* 2018;30:276–81.
4 Auffray C, Sagner M, Abdelhak S, *et al*. VIVA Europa, a land of excellence in research and innovation for health and wellbeing. *Progr Prev Med* 2017;2:e006.
5 Sagner M, McNeil A, Puska P, *et al*. The P4 Health Spectrum—A Predictive, Preventive, Personalized and Participatory Continuum for Promoting Healthspan. *Prog Cardiovasc Dis* 2017;59:506–21.
6 Russell SJ, Norvig P. Upper Saddle River. In: *Artificial intelligence: a modern approach*. 3rd ed. Prentice Hall: NJ, 2009.
7 Koza JR, Bennett FH, Andre D, *et al*. 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.
8 Safran T, Viezel-Mathieu A, Corban J, *et al*. Machine learning and melanoma: the future of screening. *J Am Acad Dermatol* 2018;78:620–1.
9 Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
10 Sun R, Limkin EJ, Vakalopoulou M, *et al*. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *Lancet Oncol* 2018;19:1180–91.
11 Khojasteh P, Aliahmad B, Kumar DK. Fundus images analysis using deep features for detection of exudates, hemorrhages and microaneurysms. *BMC Ophthalmol* 2018;18.
12 GDPR key changes with the general data protection regulation—EUGDPR. Available: https://eugdpr.org/the-regulation/ [Accessed 2 Dec 2018].
13 Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol* 2016;13:350–9.
14 Price WN. Big data and black-box medical algorithms. *Sci Transl Med* 2018;10. doi:10.1126/scitranslmed.aao5333. [Epub ahead of print: 12 Dec 2018].
15 Swan AL, Stekel DJ, Hodgman C, *et al*. A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. *BMC Genomics* 2015;16(Suppl 1).
16 Banjar H, Adelson D, Brown F, *et al*. Intelligent techniques using molecular data analysis in leukaemia: an opportunity for personalized medicine support system. *BioMed Research International* 2017;2017:1–21.
17 Code of conduct for data driven health and care technology—NHS. Available: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology [Accessed 28 Feb 2019].
18 HMA-EMA Joint Big Data Task Force: summary report. Available: https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf [Accessed 16 Feb 2019].
19 EULAR strategy. Available: https://www.eular.org/eular_strategy_2018.cfm [Accessed 16 Feb 2019].
20 van der Heijde D, Aletaha D, Carmona L, *et al*. Update of the EULAR standardised operating procedures for EULAR-endorsed recommendations. *Ann Rheum Dis* 2014;2015:8–13.
21 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. submitted.
22 Aegle legal—how does your country processes health data after GDPR? Available: http://www.aegle-uhealth.eu/en/aegle-in-your-country/united-kingdom-report.html [Accessed 16 Feb 2019].
23 European Association of Systems Medicine. EASYM Europe. Available: https://easym.eu/ [Accessed 16 Feb 2019].
24 ICPerMed International Consortium. Available: https://www.icpermed.eu/ [Accessed 16 Feb 2019].
25 NIH funds additional medical centers to expand national precision medicine research program. Available: https://allofus.nih.gov/news-events-and-media/announcements/nih-funds-additional-medical-centers-expand-national-precision [Accessed 16 Feb 2019].
26 Open Data in a Big Data World—The World Academy of Science Website. Available: https://twas.org/sites/default/files/open-data-in-big-data-world_short_en.pdf [Accessed 16 Feb 2019].
27 Oxford Centre for Evidence-Based Medicine—levels of evidence. Available: https://www.cebm.net/2009/06/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/ [Accessed Feb 16, 2019].
28 Aphinyanaphongs Y. Big data analyses in health and opportunities for research in radiology. *Semin Musculoskelet Radiol* 2017;21:032–6.
29 Claerhout B, Kalra D, Mueller C, *et al*. Federated electronic health records research technology to support clinical trial protocol optimization: evidence from EHR4CR and the InSite platform. *J Biomed Inform* 2019;90.
30 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.
31 Ramos-Casals M, Brito-Zerón P, Kostov B, *et al*. Google-driven search for big data in autoimmune geoepidemiology: analysis of 394,827 patients with systemic autoimmune diseases. *Autoimmun Rev* 2015;14:670–9.
32 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.
33 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.
34 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.

35 Wyber R, Vaillancourt S, Perry W, *et al*. Big data in global health: improving health in low- and middle-income countries. *Bull World Health Organ* 2015;93:203–8.
36 de Wit MPT, Berlo SE, Aanerud GJ, *et al*. European League Against Rheumatism recommendations for the inclusion of patient representatives in scientific projects. *Ann Rheum Dis* 2011;70:722–6.
37 Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff* 2014;33:1163–70.
38 Roadmap E. Available: https://www.eular.org/public_affairs_research_roadmap.cfm [Accessed 16 Feb 2019].
39 Guidelines ICH. Available: https://www.ich.org/products/guidelines.html [Accessed 16 Feb 2019].
40 Data collection reference guide—ICHOM inflammatory arthritis website. Available: https://ichom.org/files/medical-conditions/inflammatory-arthritis/inflammatory-arthritis-reference-guide.pdf [Accessed 16 Feb 2019].
41 Available: https://www.iso.org/en/deliverables-all.html [Accessed 16 Feb 2019].
42 CDISC standards in the clinical research process—CDISC website. Available: https://www.cdisc.org/standards [Accessed 16 Feb 2019].
43 Radner H, Chatzidionysiou K, Nikiphorou E, *et al*. EULAR recommendations for a core data set to support observational research and clinical care in rheumatoid arthritis. *Ann Rheum Dis* 2017;2018:476–9.
44 Boers M, Kirwan JR, Wells G, *et al*. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. *J Clin Epidemiol* 2014;67:745–53.
45 Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al*. The fair guiding principles for scientific data management and stewardship. *Sci Data* 2016;3.
46 Townend D. Conclusion: harmonisation in genomic and health data sharing for research: an impossible dream? *Hum Genet* 2018;137:657–64.
47 Free flow on non-personal data—European Commission Website. Available: https://ec.europa.eu/digital-single-market/en/free-flow-non-personal-data [Accessed 16 Feb 2019].
48 Available: https://www.etriks.org [Accessed 16 Feb 2019].
49 Available: https://transmartfoundation.org/ [Accessed 16 Feb 2019].
50 Available: https://flowrepository.org/ [Accessed 16 Feb 2019].
51 Taichman DB, Sahni P, Pinborg A, *et al*. Data sharing statements for clinical trials: a requirement of the International Committee of Medical Journal Editors. *Lancet* 2017;389:e12–14.
52 Data sharing—the New England Journal of Medicine website. Available: https://www.nejm.org/data-sharing [Accessed 16 Feb 2019].
53 Callaway E. Zika-microcephaly paper sparks data-sharing confusion. *Nature* 2016.
54 Wallach JD, Boyack KW, Ioannidis JPA. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. *PLoS Biol* 2018;16:e2006930.
55 Iqbal SA, Wallach JD, Khoury MJ, *et al*. Reproducible research practices and transparency across the biomedical literature. *PLoS Biol* 2016;14:e1002333.
56 Available: https://ega-archive.org/ [Accessed 16 Feb 2019].
57 Bender JL, Cyr AB, Arbuckle L, *et al*. Ethics and privacy implications of using the Internet and social media to recruit participants for health research: a privacy-by-design framework for online recruitment. *J Med Internet Res* 2017;19:e104.
58 Cichosz SL Johansen MD, Hejlesen O. Toward big data analytics: review of predictive models in management of diabetes and its complications. *J Diabetes Sci Technol* 2015;10:27–34.
59 Perry DC, Parsons N, Costa ML. 'Big data' reporting guidelines: how to answer big questions, yet avoid big problems. *Bone Joint J* 2014;96-B:1575–7. B:.
60 Russell SJ, Norvig P. Upper Saddle River. In: *Artificial intelligence: a modern approach*. 2nd ed. Prentice Hall: NJ, 2015.
61 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.
62 Available: http://dreamchallenges.org/ [Accessed 16 Feb 2019].
63 Jin X, Wah BW, Cheng X, *et al*. Significance and challenges of big data research. *Big Data Research* 2015;2:59–64.
64 Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 2016;375:1216–9.
65 Ermann J, Rao DA, Teslovich NC, *et al*. Immune cell profiling to guide therapeutic decisions in rheumatic diseases. *Nat Rev Rheumatol* 2015;11:541–51.
66 Lee CH, Yoon H-J. Medical big data: promise and challenges. *Kidney Res Clin Pract* 2017;36:3–11.
67 Suresh S. Big data and predictive analytics: applications in the care of children. *Pediatr Clin North Am* 2016;63:357–66.
68 Cvijovic M, Höfer T, Aćimović J, *et al*. Strategies for structuring interdisciplinary education in Systems Biology: an European perspective. *npj Syst Biol Appl* 2016;2.
69 Cascante M, de Atauri P, Gomez-Cabrero D, *et al*. Workforce preparation: the Biohealth computing model for master and PhD students. *J Transl Med* 2014;12(Suppl 2).
70 Gomez-Cabrero D, Marabita F, Tarazona S, *et al*. Guidelines for developing successful short advanced courses in systems medicine and systems biology. *Cell Syst* 2017;5:168–75.
71 Rise Registry – ACR. Available: https://www.rheumatology.org/I-Am-A/Rheumatologist/RISE-Registry [Accessed 2 Dec 2018].
72 Najm A, Nikiphorou E, Gossec L, *et al*. EULAR points to consider for the development process of mobile health applications for self-management in patients with rheumatic and musculoskeletal diseases. submitted..
73 Cox M, Ellsworth D. Managing big data for scientific visualization. ACM SIGGRAPH '97 course #4, exploring gigabyte datasets in real-time: algorithms, data management, and time-critical design. *Anaheim, CA: ACM Digital Library* 1997:5–17.
74 Baro E, Degoul S, Beuscart R, *et al*. Toward a literature-driven definition of big data in healthcare. *Biomed Res Int* 2015;2015:1–9.
75 A machine learning revolution—PhysicsWorld website. Available: https://physicsworld.com/a/a-machine-learning-revolution [Accessed 2 Dec 2018].
76 Fei Y, Liu X-Q, Gao K, *et al*. Analysis of influencing factors of severity in acute pancreatitis using big data mining. *Rev Assoc Med Bras* 2018;64:454–61.
77 Moscatelli M, Manconi A, Pessina M, *et al*. An infrastructure for precision medicine through analysis of big data. *BMC Bioinformatics* 2018;19.
78 Groves P, Kayyali B, Knott D, *et al*. The 'big data' revolution in healthcare. Accelerating value and innovation.. Available: https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care [Accessed 16 Feb 2019].