



**HAL**  
open science

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

Laure Gossec, Joanna Kedra, Hervé Servy, Aridaman Pandit, Simon Stones,  
Francis Berenbaum, Axel Finckh, Xenofon Baraliakos, Tanja Stamm, David  
Gomez-Cabrero, et al.

► **To cite this version:**

Laure Gossec, Joanna Kedra, Hervé Servy, Aridaman Pandit, Simon Stones, et al.. EULAR points to consider for the use of big data in rheumatic and musculoskeletal diseases. *Annals of the Rheumatic Diseases*, 2019, pp.annrheumdis-2019-215694. 10.1136/annrheumdis-2019-215694 . hal-02409516

**HAL Id: hal-02409516**

**<https://hal.sorbonne-universite.fr/hal-02409516>**

Submitted on 13 Dec 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **European League Against Rheumatism points to consider for the use of big**  
 2 **data in rheumatic and musculoskeletal diseases**

3 Laure Gossec\*, Joanna Kedra\*, Hervé Servy, Simon Stones, Aridaman Pandit,  
 4 Francis Berenbaum, Axel Finckh, Xenofon Baraliakos, Tanja Stamm, David Gomez-  
 5 Cabrero, Christian Pristipino, Rémy Choquet, Gerd Burmester, Timothy R.D.J.  
 6 Radstake

7 \* The first 2 authors have contributed equally to the study (joint authors).

8

| Name         | Highest degree | email  | ORCID               | Affiliation  | Disclosures of interest  |
|--------------|----------------|--|---------------------|--|--|
| Laure Gossec | MD, PhD        | <a href="mailto:laure.gossec@gmail.com">laure.gossec@gmail.com</a> | 0000-0002-4528-310X | Sorbonne Université, Institut Pierre Louis d'Epidémiologie et de Santé Publique, INSERM UMR S1136, Paris France ; AP-HP, Pitié Salpêtrière Hospital, Department of rheumatology, Paris, France | L Gossec has published a study for which Orange IMT (telecommunications company) performed machine learning analyses, without charge to the author |
| Joanna Kedra | MD, MSc        | <a href="mailto:jkedra.pro@gmail.com">jkedra.pro@gmail.com</a>     | 0000-0003-3535-3183 | Sorbonne Université, Institut Pierre Louis d'Epidémiologie et de Santé Publique, INSERM UMR S1136, Paris France  | None   |
| Hervé        | MSc            | <a href="mailto:hservy@sanoia.com">hservy@sanoia.com</a>           | 0000-               | Sanoia e-health  | Employee of  |

|                   |               |  |                     |   |  |
|-------------------|---------------|--|---------------------|---|--|
| Servy             |               | <a href="http://noia.com">noia.com</a>                                   | 0002-8476-4619      | services, France  | Sanoia, Digital CRO providing clinical research services including data science. |
| Aridaman Pandit   | PhD           | <a href="mailto:A.Pandit@umcutrecht.nl">A.Pandit@umcutrecht.nl</a>       | 0000-0003-2057-9737 | Dept. of Rheumatology, Clinical Immunology and Laboratory for Translational Immunology, University Medical Center Utrecht, The Netherlands. | None   |
| Simon R. Stones   | BSc (Honours) | <a href="mailto:simon@simonstones.com">simon@simonstones.com</a>         | 0000-0002-5943-1310 | School of Healthcare, University of Leeds, Leeds, UK; Young PARE, Zurich, Switzerland   | None   |
| Francis Berenbaum | MD, PhD       | <a href="mailto:Francis.berenbaum@aphp.fr">Francis.berenbaum@aphp.fr</a> | 0000-0001-8252-7815 | Sorbonne Université, INSERM CRSA, APHP Saint Antoine Hospital, Paris, France  | none   |
| Axel Finckh       | MD, PD        | <a href="mailto:Axel.finckh@hcuge.ch">Axel.finckh@hcuge.ch</a>           | 0000-0002-1210-4347 | Division of Rheumatology, University Hospital of Geneva, Switzerland  | None   |
| Xenofon           | MD,           | <a href="mailto:Xenofon.Ba">Xenofon.Ba</a>                               | 0000-               | Rheumazentrum   | None   |

|                      |                                    |                                  |                            |   |                               |
|----------------------|------------------------------------|----------------------------------|----------------------------|---|-------------------------------|
| Baraliakos           | PhD                                | raliakos@elisabethgruppe.de      | 0002-9475-9362             | Ruhrgebiet Herne, Ruhr-University Bochum, Germany   |                               |
| Tanja Stamm          | MSc, MBA, Dr. rer. biol. hum., PhD | tanja.stamm@meduniwien.ac.at     | 0000-0003-3073-7284        | Section for Outcomes Research, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Austria | None                          |
| David Gomez-Cabrero  | PhD                                | david.gomezcabrero@khi.se        | 0000-0003-4186-3788        | Translational Bioinformatics Unit, Navarra Biomed, Departamento de Salud-Universidad Pública de Navarra, Pamplona, Navarra, Spain         | None                          |
| Christina Pristipino | MD, PhD                            | pristipino.c@gmail.com           | <u>0000-0002-2246-0652</u> | <u>Ospedale San Filippo Neri, Rome, Italy.</u>  | None                          |
| Rémy Choquet         | PhD                                | remy.choquet@orange.com          | 0000-0002-1843-489X        | INSERM U1142, Orange Healthcare, France   | Employee of Orange Healthcare |
| Gerd Burmester       | MD, PhD                            | <u>gerd.burmester@charite.de</u> | 0000-0001-7518-1131        | Department of Rheumatology and Clinical Immunology,   | None                          |

|                 |            |  |                    |  |      |
|-----------------|------------|--|--------------------|--|------|
|                 |            |  |                    | Charité<br>University<br>Medicine Berlin,<br>Germany   |      |
| Tim<br>Radstake | MD,<br>PhD | <a href="mailto:T.R.D.J.Radstake@umcutrecht.nl">T.R.D.J.Radstake@umcutrecht.nl</a> | 0000-0003-1241-691 | Dept. of<br>Rheumatology,<br>Clinical<br>Immunology and<br>Laboratory for<br>Translational<br>Immunology,<br>University Medical<br>Center Utrecht,<br>The Netherlands. | None |

9

10 **Grants:** supported by the European League Against Rheumatism, EULAR (grant  
11 SCI018).

12

13 **CORRESPONDING AUTHOR**

14 Pr. Laure Gossec, Hôpital Pitié-Salpêtrière, Service de Rhumatologie, 47-83,  
15 boulevard de l'Hôpital - 75013 Paris France

16 [laure.gossec@aphp.fr](mailto:laure.gossec@aphp.fr) Tel: [+33 1 42 17 84 21](tel:+33142178421) Fax: +33 1 42 17 79 59

17

18 **RUNNING TITLE**

19 EULAR points to consider for the use of big data in RMDs

20

21 **Word count:** 3817 words, 4 tables, 78 references

22 **Key words for journal submission:** recommendations, big data, artificial  
23 intelligence, machine learning, biostatistics, data management, EULAR,  
24 epidemiology, health services research, outcomes research.

25 **Contributorship statement** All authors have contributed to this work and have  
26 approved the final version.

27

28 **Key messages**

29

30 ***What is already known about this subject?***

31 The use of big data by artificial intelligence, computational modelling and machine  
32 learning is a rapidly evolving field with the potential to profoundly modify RMD  
33 research and patient care.

34 ***What does this study add?***

35 These are the first European League Against Rheumatism (EULAR)-endorsed ‘points  
36 to consider’ for the use of big data in RMDs. These points address key issues  
37 including: ethics, data sources, data storage, data analyses, artificial intelligence  
38 (e.g., computational modelling, machine learning), the need for benchmarking,  
39 adequate reporting of methods, and implementation of findings into clinical practice.

40 ***How might this impact on clinical practice or future developments?***

41 These points to consider will promote advances and homogeneity in the field of big  
42 data in RMDs, and may be useful as guidance in other medical fields.

43

44

45 **Abstract** (238 words)

46

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

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

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

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

69

70

## 71 INTRODUCTION

72

73 The recent expansion of big datasets and advanced computational techniques lead  
74 to tremendous opportunities for health research.[1] As elegantly elaborated by E.  
75 Topol the use of big data in medicine is going to disrupt the medical system as we  
76 know it.[2] Big data include both clinical data (e.g. originating from electronic health  
77 records, healthcare system claims data or patient-generated data such as from  
78 Apps), biological data issued from the development of molecular research leading to  
79 multi-omics complex molecular data,[3] social data (e.g. originating from social  
80 networks, Internet Of Things, physical social connections or economic data  
81 repositories), imaging data, and environmental data (e.g. urbanistic data, pollution or  
82 atmospheric conditions).[4, 5] In parallel, artificial intelligence-based methodologies  
83 allowing computer systems to "learn" from data (i.e., progressively improve  
84 performance on a specific task without being explicitly programmed) are more and  
85 more accessible.[6, 7] The collection of big data combined with such information  
86 processing techniques (computational modelling, machine learning) lead to an  
87 opportunity for progress in medical research, which should ultimately modify patient  
88 care and clinical decision making.

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

99 To our knowledge, no academic societies have developed consensus guidelines  
100 dealing with big data.[17] Very recently, the European Medicines Agency (EMA)  
101 released recommendations focused on the acceptability of evidence derived from big  
102 data in support of the evaluation and supervision of medicines by regulators;[18]  
103 however, these recommendations deal mainly with the interpretation of drug-related  
104 big data. The European League Against Rheumatism (EULAR) has recently



105 formulated as one of its key strategic objectives, the advancement of high-quality  
106 collaborative research and comprehensive quality of care for people living with  
107 rheumatic and musculoskeletal disorders (RMDs).[19] Thus, EULAR naturally takes  
108 an interest in big data and its applications.

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

111

112

113

## 114 **METHODS**

115

116 After approval by the EULAR Executive Committee, the convenors (LG, TR) and the  
117 project fellow (JK) led a multidisciplinary taskforce guided by the 2014 updated  
118 EULAR Standardised Operating Procedures, [20] which were modified for this  
119 specific taskforce. In October 2018, the main questions to be addressed in the  
120 preparatory work for the taskforce were defined as: 1) data sources and collection; 2)  
121 data analyses; and 3) data interpretation and implementation of findings. These  
122 questions were addressed in subsequent months leading up to the face-to-face  
123 meeting by the project fellow and the convenors. A systematic literature review (SLR)  
124 was performed between November 2018 and February 2019, regarding publications  
125 employing big data in RMDs, with a comparison in other medical fields.[21]  
126 Additionally, a narrative review of unpublished data on websites on big data and  
127 artificial intelligence was performed to inform the taskforce [12, 17,18, 22-26] and  
128 expert opinions were obtained from four selected persons through individual  
129 telephone interviews.

130 In February 2019, during a one-day face-to-face task force meeting, overarching  
131 principles and PTC were developed. The process was both evidence-based and  
132 consensus-based, through discussions of the international task force of experts from  
133 a range of disciplines including computer science and artificial intelligence. The task  
134 force consisted of 14 individuals from 8 European countries: 6 rheumatologists, 4  
135 data scientists/big data experts, 1 cardiologist specialized in systems medicine, 1  
136 patient research partner, 1 health professional with expertise in outcomes research  
137 and 1 fellow in rheumatology. Furthermore, feedback was obtained after the meeting  
138 from 2 additional experts. This inclusive approach aimed to obtain broad consensus

139 and applicability of the PTC. During the one-day meeting, the preparatory work was  
140 presented and discussed, the target audience of the PTC was defined, then the PTC  
141 were formulated and extensively discussed. The PTC were finalised over the  
142 subsequent 2 weeks by online discussions, taking into account the publication the  
143 same week of an EMA consensus document on big data.[18]  
144 During the meeting and through online discussions, based on the gaps in evidence  
145 and the issues raised among the task force, a research agenda was also formulated.  
146 After the PTC were finalised, the level of evidence and strength of each PTC were  
147 ascertained according to the Oxford system.[27] Finally, each task force member  
148 voted anonymously on their level of agreement with each PTC via email (numeric  
149 rating scale ranging from 0=do not agree to 10=fully agree). The mean and standard  
150 deviation of the level of agreement of taskforce members were calculated.  
151 The final manuscript was reviewed and approved by all task force members and  
152 approved by the EULAR Executive Committee.

153

154

155

## 156 **RESULTS**

157

### 158 **Target audience**

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

167

168 Overarching principles and PTC were formulated, which are shown in **Table 1** and  
169 are discussed in detail below.

170

### 171 **Definitions of terms**

172 This first point in **Table 1** proposes a definition of terms relating to big data. Although  
173 the term big data is widely utilised, there is not one commonly accepted definition.  
174 When performing the literature review, several definitions were found (**Table 2**).[6,  
175 21] The first overarching principle defines the term big data, largely based on the  
176 EMA definition.[18] Big data is defined by its size and diversity– it is diverse,  
177 heterogeneous and large and incorporates multiple data types and forms; but also by  
178 the specific complexity and challenges of integrating the data to enable a combined  
179 analysis.[18] The second half of the definition refers to artificial intelligence (AI). AI is  
180 defined as the ability of a machine to mimic "cognitive" functions that humans  
181 associate with human minds, such as "learning" and "problem solving".[6] New  
182 computational techniques, such as AI (which includes machine-learning and deep  
183 learning) are often (but not necessarily) applied to big data.[18]  
184 This next sentence is informative and aims to present the diversity of data sources  
185 leading to big data; we listed in a non-exhaustive way some of the sources of big  
186 data. The most common sources of big healthcare data found in the SLR were  
187 clinical; these include electronic health records, studies and registries, billing and  
188 healthcare system claims databases.[21, 28, 29] A more recent source of clinical big  
189 data, currently underused in RMDs is the Internet of Things (e.g. wearables, apps,  
190 medical devices and sensors), but also social media, behavioural and environmental  
191 data.[18, 30, 31] Imaging is also a growing field of application of big data.[10, 32, 33]  
192 Regarding basic and translational research results, -omics such as genomics and  
193 bioanalytical omics are an important and rapidly growing field for big data.[18, 34]

194

## 195 **Overarching principles**

### 196 ***Overarching principle A – Ethical aspects***

197 This overarching principle addresses ethical issues with big data. The collection,  
198 analysis and implementation of big data in RMDs must adhere to all applicable  
199 regulations. This covers privacy, confidentiality and security, ownership of data, data  
200 minimalization, and flow of data within the EU and with third countries.[22, 35] This is  
201 both a regulatory and legal requirement, and an ethical one.[12] In terms of legal  
202 requirements, the General Data Protection Regulation (GDPR) has set standards  
203 which apply across Europe but for health-related data, national rules could also apply  
204 on top of these.[12]

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

211

### 212 ***B – Potential of big data***

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

220

### 221 ***C – Ultimate goal***

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

225

### 226 **Points to consider**

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

229

### 230 ***PTC 1: Data collection - use of standards***

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

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

248

### 249 ***PTC 2: Data collection and storage - FAIR principle***

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

257

### 258 ***PTC 3: Data storage - data platforms***

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

266 Public access to data is an important point, which raised much debate within the task  
267 force. Internationally, several groups emphasised the principle that big data should  
268 be made publicly available to promote open and reproducible research; in particular  
269 when the data is publicly funded.[18, 26, 51, 52] On the contrary, downsides of public  
270 access to data are the potential loss of momentum to secure intellectual property and  
271 scientific publications from the researchers who initially generated the data [53] Given  
272 this controversy, data sharing should be achieved in a way that is sustainable for all

273 parties involved.[53] How to make data but also algorithms openly available is very  
274 complex.[54, 55] The task force consensus was in favour of accessible data, but in  
275 the current situation, with limited and supervised access; we also felt that pilot  
276 projects to assess the impact of data sharing are needed and that such data sharing  
277 should be evidence-based. [56] This consensus will need to be revised as the  
278 situation evolves. The topic of data sharing was also added to the research agenda.

279

#### 280 ***PTC 4: Privacy by design***

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

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

303

#### 304 ***PTC 5: Collaboration***

305 While interdisciplinary collaboration is beneficial and required for all research  
306 projects, it is even more important in big data projects where expertise is dispersed

307 among different stakeholders. The task force insisted on the importance of  
308 collaboration between appropriate stakeholders, not only at the analysis stage, for  
309 example, where AI methods require appropriate expertise, but at all phases of a big  
310 data project.[25] Interdisciplinary collaborations should intervene at different times  
311 across a project, to enable the most appropriate design to be chosen, while ensuring  
312 that data collection and the type of analysis are fit for purpose. Of note, the statistical  
313 methods may be based on AI or may include more traditional statistics and/or  
314 computational methodologies, as appropriate. Further knowledge is needed on the  
315 comparison of statistical methods, which discussed in more detail in PTC 7.[21, 58]  
316 The appropriate individuals to collaborate include clinical/biological scientists,  
317 computational/data scientists, health professionals and patients: proposals for  
318 respective roles are shown in **Table 3**.

319

#### 320 ***PTC 6: Data analyses reporting***

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

327

#### 328 ***PTC 7: Benchmarking of data analyses***

329 AI encompasses several techniques which are intended to solve the most difficult  
330 problems in computer science: search and optimisation (heuristics), logic (fuzzy  
331 logic), uncertain reasoning and learning (machine learning).[60]. In our SLR, machine  
332 learning methods were the most used artificial intelligence techniques, in RMDs and  
333 in other medical fields (98% and 100% of artificial intelligence papers, respectively).  
334 The most used machine learning algorithms were artificial neural networks (with deep  
335 learning as the most advanced version), representing 48% of AI articles.[21 , 61]  
336 In addition, comparison of artificial intelligence methods within RMDs should be  
337 promoted. [17, 18, 24, 62] This is particularly needed because AI is a rapidly growing  
338 field; there is an ongoing and unsolved debate, as to which methods within artificial  
339 intelligence perform best.[63, 64] The comparison of AI methods was also added to

340 the research agenda, since it was felt that this particularly topic was difficult to  
341 perform at this moment in time, and was more aspirational.

342

### 343 ***PTC 8: Validation of big data findings***

344 Although there may be a perception that big data are more valid or less subject to  
345 bias than traditional studies, model overfitting, inappropriate generalisation of the  
346 results and/or bias can in fact lead to inappropriate conclusions. [14, 18, 28]. Thus, it  
347 is important both to assess and benchmark the quality of the generated data and the  
348 methods used to avoid over-interpretation of results, overfitting of the models, and  
349 generalisation of the results when using big data. The task force also felt that it was  
350 important to validate results in independent datasets. [24, 28] Overall, the task force  
351 agreed that conclusions drawn from big data need independent validation (in other  
352 datasets) to overcome current limitations and to assure scientific soundness.  
353 However, a specific challenge for big datasets and the validation of results is the  
354 need for other (similar) big datasets – thus, feasibility of validation is a key issue  
355 which was discussed at length within the taskforce

356

### 357 ***PTC 9: Implementation of findings***

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

369

### 370 ***PTC 10: Training***

371 Interdisciplinary training for clinical, biological or imaging researchers, healthcare  
372 professionals and computational biologists/data scientists in the field of big data is  
373 important and links closely with the need for collaborations in the field of big data



374 **(Table 3)**. Indeed, machine learning methods are becoming ubiquitous, and have  
375 major implications for scientific discovery;[26] however, healthcare professionals are  
376 not perfectly aware of the correct use of these methods, whereas data scientists may  
377 lack the clinical knowledge to design studies and interpret the findings **(Table 3)**.  
378 Given the current relative lack of expertise related to big data in the field of RMDs,  
379 and given the rapid changes in this field, certain organisations should set up or  
380 facilitate training sessions.[18, 37] This may include academic institutes, public  
381 research bodies and international organisations, such as EULAR. The training is  
382 needed for both sides: the healthcare professionals needing to learn about the basics  
383 of big data, and the data scientists needing to better understand the clinical questions  
384 and context within which big data has been collected, and/or is being applied.[68]  
385 The training can be performed separately for the different stakeholders but, in some  
386 instances, it will require an interdisciplinary educational setting in order to engage  
387 multidisciplinary teams and their unique dynamics (e.g. the need to set a common  
388 vocabulary). The training process should detect skills gaps, identify individuals with  
389 bioinformatics/biostatistics/analytics/data science expertise within or outside the field  
390 of RMDs, and implement appropriate training. The training should also aim for  
391 different levels of education provision, ranging from academic taught modules  
392 (undergraduate and postgraduate), academic research modules (PhD) and  
393 continuous professional development opportunities (for example, through seminars  
394 and workshops). Similar efforts can be observed in Systems Biology and Systems  
395 Medicine.[18, 68-70]

396

397

### 398 **Research agenda**

399

400 Based on the discussions among the task force and the areas of uncertainty  
401 identified within the SLR and discussions among expert stakeholders, a research  
402 agenda has been proposed, depicted in **Table 4**. This research agenda covers  
403 issues related to data collection, data analyses, training, interpretation of findings and  
404 implementation of findings.

405

406

407

408 **DISCUSSION**

409

410 These are the first EULAR-endorsed PTC for the use of big data within the field of  
411 RMDs, which could well be applied by other medical disciplines. These PTC address  
412 the core aspects of big data, namely data sources and storage, including ethical  
413 aspects, data analyses, data interpretation and implementation. Legal aspects are  
414 not clearly mentioned, but these points to consider were meant to cover principles  
415 and practical aspects of big data; however, the law, and in particular GDPR, applies  
416 first. (12) For the update of these points to consider in a few years, participants with  
417 legal and ethical expertise should be considered.

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

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

429

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

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

446 The grading of the evidence was a challenge in the present work as the Oxford level  
447 of evidence (27) which is used in EULAR taskforces is better adapted to therapeutic  
448 evidence than to observational or prognostic evidence as is often obtained in big data  
449 work. However, according to EULAR Standardized Operating Procedures (20), levels  
450 of evidence and strength of recommendations should be rated by the Oxford Levels  
451 of Evidence. Moreover, in the case where there is little data-driven evidence, EULAR  
452 Standardized Operating Procedures recommend to downgrade the recommendations  
453 to the level of “points to consider”, which is what was performed here.

454 This work has several limitations: the main one is that the present PTC are not  
455 specific to RMDs. However, they are not specific because the aspects of big data that  
456 they address are universal, and at present, there is no specific issue related to big  
457 data in RMDs, as is also the case in any other medical speciality. Moreover, the  
458 experts we consulted consider big data as an opportunity to go beyond the traditional  
459 division of medical specialties and allow multidisciplinary approaches. The other main  
460 limitation was the extremely low level of evidence for all the PTC, raising the question  
461 of the interest of evidence in this specific field where the PTC were expert-driven.  
462 This is often the case on subjects where recommendations are formulated before  
463 supportive data are produced.[72] It is linked to the novelty of the subject.

464  
465 In conclusion, it is anticipated that new data in this rapidly moving field will emerge  
466 over the next few years and that some of the questions formulated in the research  
467 agenda will be answered. Therefore, we will consider an update of these PTC as  
468 needed in a few years.

469

470 **ACKNOWLEDGMENTS**

471 The authors would like to thank the experts who kindly shared their opinions during  
472 individual interviews as part of the preparation for this taskforce: William Dixon (UK),  
473 Iain McInnes (UK), Harald Schmidt (Netherlands) and Jan Baumbach (Germany). We  
474 also wish to thank the experts who kindly advised us on the manuscript: Iain McInnes  
475 (UK) and Neil Betteridge (UK).

476

477

478  
479  
480

**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

| <b>Definitions</b>  |  |                       |                       |            |
|---|--|-----------------------|-----------------------|------------|
| The term ‘big data’ refers to extremely large datasets which may be complex, multi-dimensional, unstructured and from heterogeneous sources, and which accumulate rapidly. Computational technologies, including artificial intelligence (e.g. machine learning), are often applied to big data. Big data may arise from multiple data sources including clinical, biological, social and environmental data sources. |  |                       |                       |            |
| <b>Overarching principles</b>   |  |                       | <b>LoA, mean (SD)</b> |            |
| 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)             |            |
| <b>Points to consider</b>   |  | <b>LoA, mean (SD)</b> | <b>LoE</b>            | <b>SoR</b> |
| 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          |

481 LoA, level of agreement; LoE, level of evidence, SoR: strength of recommendation  
482 Numbers in the column 'LoA' indicate the mean and SD (in parentheses) of the LoA, as well as the  
483 mean agreement of the 14 task force members on a 0-10 scale. LoE and strength based on the  
484 Oxford Centre for Evidence-Based Medicine classification, with 'Level 1' corresponding to meta-  
485 analysis or randomized controlled trials (RCT) or high quality RCTs; 'Level 2' to lesser quality RCT or  
486 prospective comparative studies; 'Level 3' to case-control studies or retrospective studies; 'Level 4' to  
487 case series without the use of comparison or control groups; 'Level 5' to case reports or expert  
488 opinion[27]  
489  
490

491 **Table 2.** 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 $\text{Log}(n \cdot p)$ 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 e.g. Artificial intelligence [75]  |
| Information assets characterized 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 i.e., 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]  |

492

493

**Table 3.** Stakeholders involved in big data research: Proposal of potential roles

| <b>Stakeholder</b>   | <b>Characteristics</b>   | <b>Potential role in big data research</b>   |
|--|--|--|
| 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 modeling) as well as emerging methods (67). | Provide early guidance on the best tools or algorithm to analyze the data.<br><br>Analyses of data and interpretation        |
| Computational biologist  | Involved in the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, ecological, behavioral, and social systems. Has domain knowledge in biology. | Provide early guidance on the best tools or algorithm to analyze the data.<br><br>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 |



496 **Table 4.** 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 minimization 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 |
|                  | 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 unauthorized/ compassionate use cases  |
|                  | Define the role, modalities and rules of patient engagement in the generation and exploitation of big data.  |

## REFERENCES

1. Saria S, Butte A, Sheikh A. Better medicine through Machine Learning: what's real, and what's artificial? *PloS Med.* 2018;15:E1002721.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25:44-56.
3. Dixon W, Michaud K. Using technology to support clinical care and research in rheumatoid arthritis. *Curr Opin Rheumatol* 2018, 30:276–281.
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, Auffray C, Price ND, Hood L, et al. The P4 Health Spectrum – A Predictive, Preventive, Personalized and Participatory Continuum for Promoting Healthspan. *Progr Cardiovasc Dis* 2017; 59, 506–521.
6. Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach (3rd ed.). Upper Saddle River, NJ: Prentice Hall 2009.
7. Koza JR, Bennett FH, Andre D, Keane MA. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In: Gero JS, Sudweeks F, eds. Artificial Intelligence in Design. Dordrecht (NL): Elsevier Academic Publishers 1996.
8. Safran T, Viezel-Mathieu A, Corban J, Kanevsky A, Thibaudeau S, Kanevsky J. Machine learning and melanoma: The future of screening. *J Am Acad Dermatol.* 2018;78:620-621.
9. Esteva A, Kuprel B, Novoa RA., Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017, 542: 115-118.
10. Sun R, Limkin EJ, Vakalopoulou M, Dercle L, Champiat S, Han SR, 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;19:1180-1191.
11. Khojasteh P, Aliahmad B, Kumar DK. Fundus images analysis using deep features for detection of exudates, hemorrhages and microaneurysms. *BMC Ophthalmol.* 2018;18:288.
12. GDPR Key Changes with the General Data Protection Regulation – EUGDPR <https://eugdpr.org/the-regulation/> [accessed Dec 2 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;10:471.
15. Swan AL, Stekel DJ, Hodgman C, Allaway D, Alqahtani MH, Mobasheri A et al. A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. *BMC Genomics.* 2015;16(Suppl 1) :S2.
16. Banjar H, Adelson D, Brown F, Chaudhri N. Intelligent Techniques Using Molecular Data Analysis in Leukaemia: An Opportunity for Personalized Medicine Support System. *Biomed Res Int.* 2017;2017:1-21.
17. Code of conduct for data driven health and care technology – NHS <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 Feb 28 2019].
18. HMA-EMA Joint Big Data Taskforce: summary report. [https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report\\_en.pdf](https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf) [accessed Feb 16, 2019].

19. EULAR Strategy. [https://www.eular.org/eular\\_strategy\\_2018.cfm](https://www.eular.org/eular_strategy_2018.cfm) [accessed Feb 16, 2019].
20. van der Heijde D, Aletaha D, Carmona L, Edwards CJ, Kvien TK, Kouloumas M et al. 2014 Update of the EULAR standardised operating procedures for EULAR-endorsed recommendations. *Ann Rheum Dis*. 2015;74:8-13.
21. Kedra J, Radstake T, Pandit A, Baraliakos X, Berenbaum F, Finckh A et al. Current status of the use of Big Data and Artificial Intelligence in RMDs: a systematic literature review informing EULAR recommendations. *RMD Open* 2019 (submitted)
22. AEGLE legal – how does your country processes Health Data after GDPR? <http://www.aegle-uhealth.eu/en/aegle-in-your-country/united-kingdom-report.html> [accessed Feb 16, 2019].
23. <https://easym.eu/> [accessed Feb 16, 2019].
24. <https://www.icpermed.eu/> [accessed Feb 16, 2019].
25. NIH funds additional medical centers to expand national precision medicine research program <https://allofus.nih.gov/news-events-and-media/announcements/nih-funds-additional-medical-centers-expand-national-precision> [accessed Feb 16, 2019].
26. Open Data in a Big Data World - The World Academy of Science Website [https://twas.org/sites/default/files/open-data-in-big-data-world\\_short\\_en.pdf](https://twas.org/sites/default/files/open-data-in-big-data-world_short_en.pdf) [accessed Feb 16, 2019].
27. Oxford Centre for Evidence-based Medicine – Levels of Evidence <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;21:32-36.
29. Claerhout B, Kalra D, Mueller C, Singh G, Ammour N, Meloni L, 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;90:103090.
30. Gossec L, Guyard F, Leroy D, Lafargue T, Seiler M, Jacquemin C et al. Detection of flares by decrease in physical activity, collected using wearable activity trackers, in rheumatoid arthritis or axial spondyloarthritis: an application of Machine-Learning analyses in rheumatology. *Arthritis Care Res (Hoboken)*. 2018. doi: 10.1002/acr.23768. [Epub ahead of print]
31. Ramos-Casals M, Brito-Zeron P, Kostov B, Siso-Almirall A, Bosch X, Buss D 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, Gao J, Siegel EL. 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;44:307-315.
34. Suwinski P, Ong C, Ling MHT, Poh YM, Khan AM, Ong HS. Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. *Front Genet*. 2019;10:49.
35. Wyber R, Vaillancourt S, Perry W, Mannava P, Falranmi T, Celi LA. Big data in global health: improving health in low- and middle-income countries. *Bull World Health Organ*. 2015;93:203-8.

36. de Wit MP, Berlo SE, Aanerud GJ, Aletaha D, Bijlsma JW, Croucher L 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 (Millwood)*. 2014;33:1163-1170.
38. EULAR Roadmap. [https://www.eular.org/public\\_affairs\\_research\\_roadmap.cfm](https://www.eular.org/public_affairs_research_roadmap.cfm) [accessed Feb 16, 2019]
39. ICH Guidelines. <https://www.ich.org/products/guidelines.html> [accessed Feb 16, 2019].
40. Data collection reference guide - ICHOM inflammatory arthritis website. <https://ichom.org/files/medical-conditions/inflammatory-arthritis/inflammatory-arthritis-reference-guide.pdf> [accessed Feb 16, 2019].
41. <https://www.iso.org/en/deliverables-all.html> [accessed Feb 16, 2019].
42. CDISC Standards in the Clinical Research Process – CDISC website. <https://www.cdisc.org/standards> [accessed Feb 16, 2019].
43. Radner H, Chatzidionysiou K, Nikiphorou E, Gossec L, Hyrich KL, Zabalán C et al. 2017 EULAR recommendations for a core data set to support observational research and clinical care in rheumatoid arthritis. *Ann Rheum Dis*. 2018;77:476-479.
44. Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d’Agostino MA 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 IJ, Appleton G, Axton M, Baak A et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
46. Townend D. Conclusion: harmonisation in genomic and health data sharing for research: an impossible dream? *Hum Genet*. 2018;137:657-664.
47. Free flow on non-personal data – European Commission Website. <https://ec.europa.eu/digital-single-market/en/free-flow-non-personal-data> [accessed Feb 16, 2019].
48. <https://www.etriks.org> [accessed Feb 16, 2019].
49. <https://transmartfoundation.org/> [accessed Feb 16, 2019].
50. <https://flowrepository.org/> [accessed Feb 16, 2019].
51. Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A et al. Data sharing statements for clinical trials: a requirement of the International Committee of medical Journals Editors. *Lancet*. 2017;389:e12-e14.
52. Data sharing – the New England Journal of Medicine Website. <https://www.nejm.org/data-sharing> [accessed Feb 16, 2019].
53. Callaway E. Zika-microcephaly paper sparks data-sharing confusion. *Nature*. 2016 Feb 12. <https://www.nature.com/news/zika-microcephaly-paper-sparks-data-sharing-confusion-1.19367> [accessed Feb 16, 2019]
54. Wallach JD, Boyack KW, Ionnadis JPA. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. *Plos Biology*. 2018;16:e2006930.
55. Iqbal SA, Wallash JD, Houry MJ, Schully SD, Ioannidis JPA. Reproducible Research Practices and Transparency across the Biomedical Literature. *PLoS Biol*. 2016; 14: e1002333.

56. <https://ega-archive.org/> [accessed Feb 16, 2019]. Available on:
57. Bender JL, Cyr AB, Arbuckle L, Ferris LE. 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 NL, Costa ML. 'Big data' reporting guidelines: how to answer big questions, yet avoid big problems. *Bone Joint J*. 2014;96-B:1575–7.
60. Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach (2nd ed.), Upper Saddle River, NJ: Prentice Hall 2015.
61. LeCunn Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436-444.
62. <http://dreamchallenges.org/> [accessed Feb 16, 2019].
63. Jin X, Wah BW, Cheng X, Wang Y. 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, Brenner MB, Raychaudhuri S. Immune cell profiling to guide therapeutic decisions in rheumatic diseases. *Nat Rev Rheumatol*. 2015;11:541-51.
66. Lee CH, Yoon HJ. 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, Acimovic J, Alberghina L, Almaas E, Besozzi D et al. Strategies for structuring interdisciplinary education in Systems Biology: an European perspective. *NPJ Syst Biol Appl*. 2016; 2: 16011.
69. Cascante M, de Atauri P, Gomez-Cabrero D, Wagner P, Centelles JJ, Marin S et al. Workforce preparation: the Biohealth computing model for Master and PhD students. *J Transl Med*. 2014;12(Suppl2) :S11.
70. Gomez-Cabrero D, Marabita F, Tarazona S, Cano I, Roca J, Conesa A et al. Guidelines for Developing Successful Short Advanced Courses in Systems Medicine and Systems Biology. *Cell Syst*. 2017;5:168-175.
71. Rise Registry – ACR <https://www.rheumatology.org/I-Am-A/Rheumatologist/RISE-Registry> [accessed Dec 2 2018].
72. Najm A, Nikiphorou E, Gossec L, Berenbaum F. 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, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. *Biomed Res Int*. 2015;2015:639021.
75. A machine learning revolution – PhysicsWorld website. <https://physicsworld.com/a/a-machine-learning-revolution> [accessed Dec 2 2018].
76. Fei Y, Liu XQ, Gao K, Xue CB, Tang L, Tu JF, et al. Analysis of influencing factors of severity in acute pancreatitis using big data mining. *Rev Assoc Med Bras (1992)*. 2018;64:454-461.

77. Moscatelli M, Manconi A, Pessina M, Fellegara G, Rampoldi S, Milanesi L, Casasco A, Gnocchi M. An infrastructure for precision medicine through analysis of big data. *BMC Bioinformatics*. 2018;19(Suppl 10):351.
78. Groves P, Kayyali B, Knott D, Van Kuiken S. The 'big data' revolution in healthcare. Accelerating value and innovation. <https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care> [accessed Feb 16, 2019].