MULTIPLE SCLEROSIS JOURNAL | MSJ

Personal Viewpoint

Fair data for next-generation management of multiple sclerosis

Liesbet M Peeters

Multiple Sclerosis Journal
2018, Vol. 24(9) 1151–1156
DOI: 10.1177/1352458517748475

Correspondence to:
LM Peeters
Biomedical Research Institute, Hasselt University and School of Life Sciences, Transnationale Universiteit Limburg, Diepenbeek, 3590, Belgium.
liesbet.peeters@uhasselt.be

Liesbet M Peeters
Biomedical Research Institute, Hasselt University and School of Life Sciences, Transnationale Universiteit Limburg, Diepenbeek, Belgium

Abstract: Multiple sclerosis (MS) is a progressive demyelinating and degenerative disease of the central nervous system with symptoms depending on the disease type and the site of lesions and is featured by heterogeneity of clinical expressions and responses to treatment strategies. An individualized clinical follow-up and multidisciplinary treatment is required. Transforming the population-based management of today into an individualized, personalized and precision-level management is a major goal in research. Indeed, a complex and unique interplay between genetic background and environmental exposure in each case likely determines clinical heterogeneity. To reach insights at the individual level, extensive amount of data are required. Many databases have been developed over the last few decades, but access to them is limited, and data are acquired in different ways and differences in definitions and indexing and software platforms preclude direct integration. Most existing (inter)national registers and IT platforms are strictly observational or focus on disease epidemiology or access to new disease modifying drugs. Here, a method to revolutionize management of MS to a personalized, individualized and precision level is outlined. The key to achieve this next level is FAIR data.

Keywords: Individualized medicine, data management, multidisciplinary treatment, FAIR data, next-generation management, multiple sclerosis

Date received: 9 September 2017; revised: 17 November 2017; accepted: 21 November 2017

Multiple sclerosis (MS) is a progressive demyelinating and degenerative disease of the central nervous system with symptoms depending on the disease type and the site of lesions and is featured by heterogeneity of clinical expressions and responses to treatment strategies. An individualized clinical follow-up and multidisciplinary treatment is required. Transforming the population-based management of today into an individualized, personalized and precision-level management is a major goal in research. However, a complex and unique interplay between genetic background and environmental exposure in each case likely determines clinical heterogeneity. To reach insights at the individual level, extensive amount of data are required. Here, a method to revolutionize management of MS to a personalized, individualized and precision level is outlined. The key to reach this next level is FAIR data. FAIR is a fairly recent concept that stands for Findable, Accessible, Interoperable and Reusable.[1]

Imagine any type of data being 'Findable, Accessible, Interoperable and Reusable' by both humans and machines. 'Findable' does not mean 'for everyone to find', 'accessible' does not mean 'open access', 'interoperable' does not mean 'for everyone to operate on' and 're-usable' 'for everyone to use'. However, it creates the possibility to find, access, interoperate and re-use data when necessary. In other words, it gives data the opportunity to have maximal impact. The possibilities to discover new insights multiplies manifold. But before we get there, many hurdles have to be overcome. Here, a 4C plan is proposed to reach this goal (Collect-Connect-Complete-Construct). An intuitive representation of this plan is represented in Figure 1.

Data are collected all over the world by different stakeholders resulting in many datasets, represented by puzzles of a face (step 1: COLLECT). Every dataset has its own weaknesses and strengths. For example, existing and emerging MS-specific data initiatives resulted in international pooling of observational clinical data (e.g. MSBase Registry,[2] Big MS Data Group, European register for MS (EuReMS[3])), clinical trial data (e.g. Sylvia Lawry Centre for Multiple Sclerosis Research (SLCMSR[4])), magnetic resonance

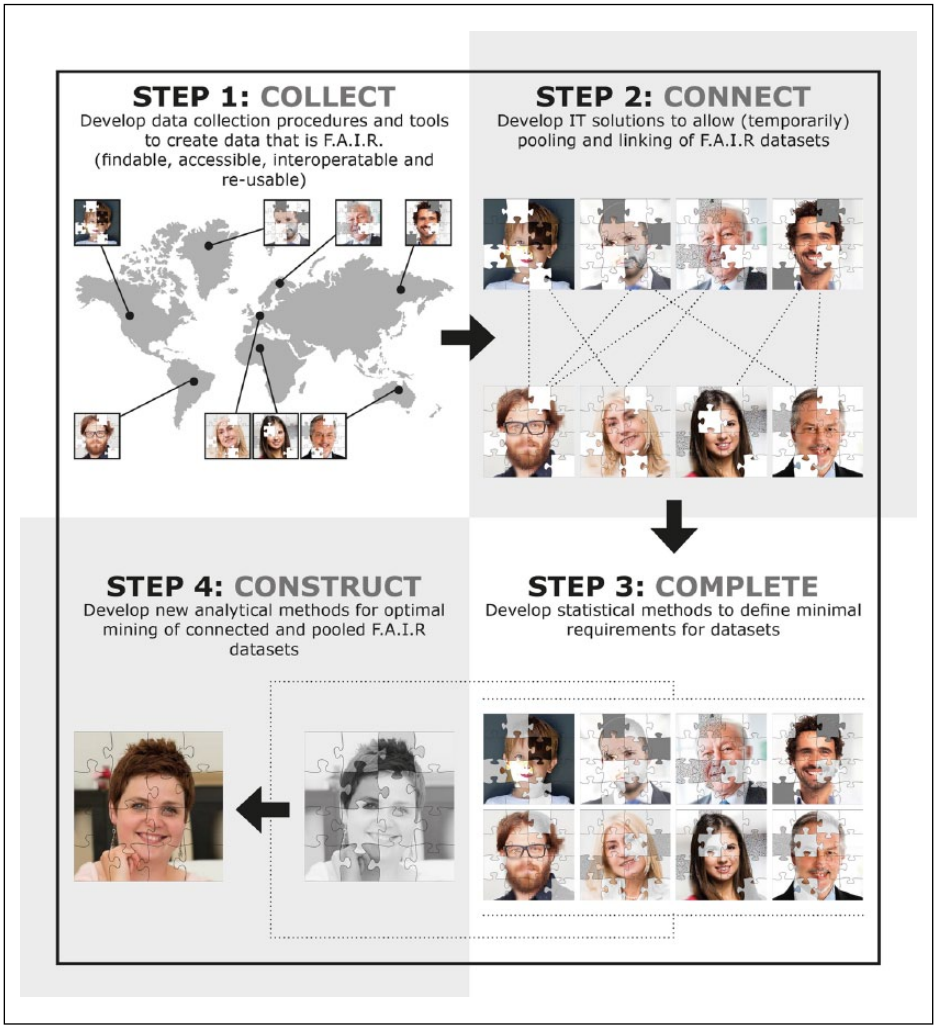

Figure 1. An intuitive representation of a 4C plan towards next-generation management.
Data are collected all over the world by different stakeholders resulting in many datasets, represented by puzzles of a face (step 1: COLLECT). Every dataset has its own weaknesses and strengths. Although none of these datasets are perfect (nor will they ever be), many insights could be discovered when these datasets could be pooled and connected (step 2: CONNECT). Sometimes, the existing data are insufficient to investigate a certain question and additional data are required. Because collecting data collection is expensive and time-consuming, efforts should be as focused as possible and methods to identify the minimal requirements for common datasets are required (step 3: COMPLETE). When sufficient overlap between the databases involved is secured and powerful analytical methods are developed to cope with the imperfections of datasets featured by different layers of missing data, these datasets can be optimally mined to create new insights for MS management (step 4: CONSTRUCT).

images (e.g. magnetic resonance imaging in multiple sclerosis (MAGNIMS[5])), genetic data (e.g. International Multiple Sclerosis Genetic Consortium (IMSGC[6])), functional and patient-reported outcomes (e.g. iConquerMS, North American Research Committee on Multiple Sclerosis (NARCOMS[7])) and patient-centred outcomes (e.g. iPCO initiative of the Multiple Sclerosis International Federation). In addition, there is international interest in rehabilitation repositories (MSREHABREP[8]). Imperfect and inaccessible 'data silos' are also present on a local level, since data on the same patients are collected by different data collectors. Indeed, every variable is ideally collected by the expert involved. For example, neurologists are the best candidate to collect diagnosis date, medication strategy and so on. However, when it comes to patient-reported outcomes, patients should be included directly in the data collection process. The same is true for nurses (e.g. weight and blood pressure), rehabilitation specialists (e.g. physical and daily functioning), psychologists (e.g. tests for cognition and depression), speech therapists (e.g. swallowing tests and speech recognition), researchers (e.g. genetic-, immunological- and lipid metabolism) and so on. Although none of these datasets are perfect (nor will they ever be), many insights could be discovered

when these datasets could be pooled and connected (step 2: CONNECT). Indeed, in a recently approved Horizon2020 project 'MultipleMS' (www.multiplems.eu), universities and companies across Europe and the United States are connecting their datasets to tailor the development and application of therapies to the individual MS patient. Still, sometimes the existing data are insufficient to investigate a certain question and additional data are required. Because collecting data is expensive and time-consuming, efforts should be as focused as possible and methods to identify the minimal requirements for common datasets are required (step 3: COMPLETE). For example, when it comes to personalized prediction of disease progression, information on different levels is required. Collecting all these data for every patient is not feasible. But what if we could mine existing datasets to formulate guidelines for a minimal core dataset? When sufficient overlap between the databases involved is secured and powerful analytical methods are developed to cope with the imperfections of datasets featured by different layers of missing data, these datasets can be optimally mined to create new insights for MS management (step 4: CONSTRUCT).

More specifically, the following steps are formulated:

1. *COLLECT: develop data collection procedures and tools to create FAIR data*. Ideally, data collection procedures should be using open-source IT codes, permit visit-entry or automation. The term 'visit-entry' implies the direct entry of variables by the experts involved in gathering the data, and automation refers to the fact that automatic import of data that is digitally collected should be possible. Working with open-source IT codes enables a low-priced implementation of IT platforms and the possibility to meeting local needs. Next to this, procedures should be legal, ethical and practical consequences of data sharing and re-use. Currently, there are a lot of insecurities around extensive data sharing initiatives (e.g. what about informed consents? how should we handle pseudo-anonymization? and how can data be shared respecting security and privacy?).
2. *CONNECT: develop IT solutions to allow (temporarily) pooling and linking of FAIR datasets*. IT solutions that allow local collection, storage and management of data, and enable (research) question-based pooling are necessary. Today, when data need to be pooled, a lot of time, efforts and costs are necessary for data processing before pooling is possible. Some concrete solutions to simplify pooling are (1) using unique patient identification strategies (e.g. use of national identification number), (2) using database catalogues clearly defining the variables involved and (3) standardization where possible (CDISC labels[9] and Human Phenotype Ontology (HPO) classifications[10]).
3. *COMPLETE: develop statistical methods to define minimal requirements for datasets*. Many insights can be reached using existing data. However, focussed prospective of retrospective data collection will often be required. Therefore, there is a need to develop statistical methods to define minimal requirements for datasets. Being able to objectively define these minimal requirements will make it much easier to motivate people involved in data collection to collect data that are not necessary for their initial intent. Motivating or giving incentives for the collection and storage of biological samples could already be one step forward because this makes retrospective retrieval much easier and less time-consuming.
4. *CONSTRUCT: develop analytical methods for optimal data mining*. We lack proper analytical tools for optimal mining of datasets that are featured by different layers of missing data. New insights in building decision-support systems using a combination of imperfect datasets are required. Different research questions should be investigated here (e.g. what is the power of machine learning techniques for these applications? and how can we better handle missing data?).

Table 1 summarizes some concrete recommendations on how we could implement this 4C plan in the MS data arena. Our research project MS DataConnect (www.msdataconnect.com) aims at providing proof-of-concept of this 4C plan to enhance MS research. The MS DataConnect Consortium and the Belgian healthdata.be platform will collaborate to set up a multidisciplinary MS register connecting information collected by care givers, patients and researchers. User-friendly, sustainable FAIR data collection tools and procedures are developed for MS-relevant data striving towards 'the only once' – principle referring to (1) data capture from primary (operation) sources of health care actors and (2) re-use of previously collected data.

Data collection is extremely expensive and time-consuming. Enabling data to achieve maximal impact is our duty on a social and ethical level, but also greatly decreases financial costs associated with

Table 1. Recommendations for MS-specific implementation of the 4C plan.

	Where are we now?	Where do we want to be?	How can we get there?
COLLECT	Many MS-specific IT software platforms (e.g. Imed, OPTIMISE, MSBase DES, MS Bioscreen[11] and MSDS,[12])	Methods for IT-independent data capture (=international collaborations are possible independent from the IT platform used)	Catalogues with unambiguous definitions of variables with internationally accepted labels
	Limited availability and implementation of data collection tools for functional and patient-reported outcome	Standardized and widely implemented data collection tools for functional and patient-reported outcomes	Development and evaluation of mobile health application to measure functional and patient-reported outcomes
	Limited interoperability and re-use of data because of ethical and legal challenges	Informed consents and governance structures allowing maximal interoperability and re-use of data	Formulate guidelines for informed consents and repository governance structures that are GDPR compliant and respect ethical restrictions
	Excessive manual data re-entry	Get to an 'only-once' principle in which data should only be collected one time	Use of primary systems for data entry and automated data extraction for re-use
CONNECT	Successful meta-data initiatives (e.g. MSBase,[2] MAGNIMS,[5] IMSGC,[6] SLCMSR[4] and EuReMS[3])	Sustainable meta-data initiatives including as many patient records as possible	Sustainable financial support for data collection
	Limited connectivity between data silos (e.g. difficult to connect clinical data to genetic data or MRI data), mainly because the patient identifier is lost in meta-data	Patient connectivity can be ensured while guaranteeing privacy	Development of standard operating procedures approved for privacy restrictions
	Request-based pooling is time-consuming	IT solutions allowing request-based data pooling and moving towards a federated meta-database approach	Development of IT solutions to allow request-based data pooling
COMPLETE	No consensus towards core minimal datasets and limited knowledge on the relative importance of variables (e.g. is whole genome sequence necessary or are 1 or 2 SNPs enough?)	Core minimal datasets that are widely implemented	Guidelines for core minimal datasets based on relevant statistical analysis
	Retrospective retrieval is difficult (e.g. new genetic-, MRI, CSF of serum biomarker are constantly being identified)	Fast and cheap retrospective data retrieval when necessary	Sustainable collection and storage of MRI images and biological samples (CSF, serum and DNA) allowing longitudinal retrospective retrieval of biomarkers
	Current statistical methods to investigate the relative importance of variables require extensive and complete datasets	New statistical methods to identify minimal core dataset requirements using existing and imperfect datasets	Development and evaluation of statistical methods starting from imperfect datasets
CONSTRUCT	Lack of implementation of complex statistical methodology and an urgent need for new statistical methods handling missing data on different levels	Use of state-of-the-art analysis strategy in MS research	Educate researchers and encourage collaborations with statistical experts to develop and evaluate innovative methods handling data imperfections

IT: information technology; MS: multiple sclerosis; DES: data entry system; MSDS: multiple sclerosis documentation system; MAGNIMS: magnetic resonance imaging in multiple sclerosis; IMSGC: international multiple sclerosis genetic consortium; SLCMSR: Sylvia Lawry Centre for multiple sclerosis research; MRI: magnetic resonance imaging; SNP: single nucleotide polymorphism; CSF: cerebrospinal fluid; GDPR: global data protection regulation.

data collection and management. Making data available to peers incentivizes researchers to better manage their data and ensure their data are of high quality. This is recognized by several authorities, resulting in new legislations, guidelines and international calls for proposals pushing and supporting the implementation of these FAIR principles (e.g. global data protection regulation (GDPR), European Commission guidelines on FAIR data management in Horizon2020 projects, Innovative Medicine Initiatives calls for the establishment of European Health Data Networks, the rise of organizations and projects that solely focus on improving data management in life sciences, for example, the ELIXIR project (www.elixir-europe.org/)). But implementing the FAIR principles is not only about generating value for the community. It benefits the initial researcher, research sponsor, data repositories, the scientific community and the public. In a time of reduced monetary investment for science and research, data sharing is more efficient because it allows researchers to share resources. Collaboration between scientists is facilitated, enabled and encouraged, resulting in larger and more expansive datasets. This results not only in new insights and better results for the community (e.g. enhanced clinical decision-making/best practice and increased efficiency for identification of research gaps) but also benefit the researcher in many other personal ways as well (e.g. networking, increased number and impact of publication). This will lead to more motivation to contribute to the data collection and quality of the data as well, a win-win situation.[8]

The future perspectives of this 4C plan are endless and depend on the stakeholders involved. Indeed, regulators need data for life-cycle assessment of medicinal products, health technology assessment bodies want to incorporate data from clinical practice into the drug development process and researchers want to build personalized decision-support systems. To truly capture the potential of this '4C plan', please reflect on the following question: '*what would YOU investigate, if you had all the data in the world to your disposal and the analysis tools to optimally mine this data?*' This can only be achieved when efforts towards this ultimate common goal are combined and synchronized.

Acknowledgements

The author would like to thank all the members of the MS DataConnect Consortium (www.msdataconnect.com). MS DATACONNECT operates in a very strong national and international interdisciplinary network. This network connects partners involved in MS care, rehabilitation and research with partners involved in IT development, database management, data sharing procedures, statistics, machine learning and prediction modelling, and this network is expanding very fast. Currently, the following partners are involved: (1) partners involved in MS care, rehabilitation and research: the Biomedical Research Institute (BIOMED) and Rehabilitation Research Center of BIOMED (REVAL), PXL University College Hasselt, the Rehabilitation and MS Center Overpelt (RMSC), MS liga and University Biobank Limburg (UbiLim); (2) partners ensuring the technical expertise: the Center of Statistics (CENSTAT), the IT department of PXL University College Hasselt and the Data Science Institute of Imperial College London. In particular, the author thanks his supervisor Prof. Niels Hellings for his support, guidance and textual suggestions. In addition, the author would like to thank Johan Van Bussel, coordinator of the Belgian healthdata.be platform, and his team for his technical expertise and insights on e-health. Finally, the author would like to thank Hans Constandt and Filip Pattyn (Ontoforce, Belgium) for inspiring her towards using FAIR data in health care.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: L.P. was supported by Interuniversity Attraction Pole (IUAP, no IAP VII/39), the MS network Limburg, the Biomedical Research Institute of Hasselt University and the ELIXIR project.

ORCID iD

Liesbet M. Peeters http://orcid.org/0000-0002-6066-3899

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016; 3: 160018.
2. Butzkueven H, Chapman J, Cristiano E, et al. MSBase: An international, online registry and platform for collaborative outcomes research in multiple sclerosis. *Mult Scler* 2006; 12: 769–774.

3. Pugliatti M, Eskic D, Mikolcic T, et al. Assess, compare and enhance the status of persons with multiple sclerosis (MS) in Europe: A European register for MS. *Acta Neurol Scand Suppl* 2012; 195: 24–30.
4. Daumer M, Neuhaus A, Lederer C, et al. Prognosis of the individual course of disease—Steps in developing a decision support tool for multiple sclerosis. *BMC Med Inform Decis Mak* 2007; 7: 11.
5. Filippi M, Rocca MA, Ciccarelli O, et al. MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *Lancet Neurol* 2016; 15: 292–303.
6. International Multiple Sclerosis Genetics Consortium, Beecham AH, Patsopoulos NA, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet* 2013; 45: 1353–1360.
7. Fox RJ, Bacon TE, Chamot E, et al. Prevalence of multiple sclerosis symptoms across lifespan: Data from the NARCOMS registry. *Neurodegener Dis Manag* 2015; 5: 3–10.
8. Held Bradford E, Baert I, Finlayson M, et al. Feasibility of an international multiple sclerosis rehabilitation data repository: Perceived challenges and motivators for sharing data. *Int J MS Care* 2017. Epub ahead of print. DOI: 10.7224/1537-2073 .2016-009
9. Kuchinke W, Aerts J, Semler SC, et al. CDISC standard-based electronic archiving of clinical trials. *Methods Inf Med* 2009; 48: 408–413.
10. Kohler S, Vasilevsky NA, Engelstad M, et al. The human phenotype ontology in 2017. *Nucleic Acids Res* 2017; 45: D865–D876.
11. Gourraud PA, Henry RG, Cree BA, et al. Precision medicine in chronic disease management: The multiple sclerosis BioScreen. *Ann Neurol* 2014; 76: 633–642.
12. Kern R, Haase R, Eisele JC, et al. Designing an electronic patient management system for multiple sclerosis: Building a next generation multiple sclerosis documentation system. *Interact J Med Res* 2016; 5: e2.