

DEBATE

Open Access



# The challenges in data integration – heterogeneity and complexity in clinical trials and patient registries of Systemic Lupus Erythematosus

Helen Le Sueur<sup>1</sup>, Ian N. Bruce<sup>2,3</sup>, Nophar Geifman<sup>1,4\*</sup>  and on behalf of the MASTERPLANS Consortium

## Abstract

**Background:** Individual clinical trials and cohort studies are a useful source of data, often under-utilised once a study has ended. Pooling data from multiple sources could increase sample sizes and allow for further investigation of treatment effects; even if the original trial did not meet its primary goals. Through the MASTERPLANS (MAximizing Sle ThERapeutic Potential by Application of Novel and Stratified approaches) national consortium, focused on Systemic Lupus Erythematosus (SLE), we have gained valuable real-world experiences in aligning, harmonising and combining data from multiple studies and trials, specifically where standards for data capture, representation and documentation, were not used or were unavailable. This was not without challenges arising both from the inherent complexity of the disease and from differences in the way data were captured and represented across different studies.

**Main body:** Data were, unavoidably, aligned by hand, matching up equivalent or similar patient variables across the different studies. Heterogeneity-related issues were tackled and data were cleaned, organised and combined, resulting in a single large dataset ready for analysis. Overcoming these hurdles, often seen in large-scale data harmonization and integration endeavours of legacy datasets, was made possible within a realistic timescale and limited resource by focusing on specific research questions driven by the aims of MASTERPLANS. Here we describe our experiences tackling the complexities in the integration of large, diverse datasets, and the lessons learned.

**Conclusions:** Harmonising data across studies can be complex, and time and resource consuming. The work carried out here highlights the importance of using standards for data capture, recording, and representation, to facilitate both the integration of large datasets and comparison between studies. Where standards are not implemented at the source harmonisation is still possible by taking a flexible approach, with systematic preparation, and a focus on specific research questions.

**Keywords:** Data integration, Data harmonisation, Clinical trials, Lupus, Pooled analysis

\* Correspondence: [nophar.geifman@manchester.ac.uk](mailto:nophar.geifman@manchester.ac.uk)

<sup>1</sup>Centre for Health Informatics, Vaughan Housue, Portsmouth St., The University of Manchester, Manchester M13 9GB, UK

<sup>4</sup>The Manchester Molecular Pathology Innovation Centre, The University of Manchester, Manchester, UK

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

SLE is a chronic autoimmune disease, affecting different body organs, which presents with a range of symptoms and clinical manifestations. Whilst several therapies with differing modes of action are currently in use, individual response varies and overall, response rates are only 40–60% to any given treatment. MASTERPLANS, a national consortium, endeavours to improve on the current trial and error approach employed in tailoring of care, by taking a precision medicine approach; improving care for patients by identifying groups who respond well to particular therapies, and the factors that may predict response to therapy.

MASTERPLANS has gained access to a wealth of data, from a range of past clinical trials and patient cohort studies; several of which have failed to demonstrate efficacy of the drug investigated, when compared to standard of care or placebo [1, 2]. Despite the lack of detected treatment effects, data from these studies could still provide a valuable source of information. Pooling of patient-level data can increase sample sizes and provide new opportunities for analysis, while various statistical approaches can be employed to handle potential study-specific effects. Framing new research questions around groups of patients combined across studies could enable both better understanding of treatment effects, and of patient characteristics that differ between these groups.

## Main text

### Accomplishing data harmonisation

Legacy data can be extremely useful, however harmonising and combining large amounts of data from disparate datasets is not always straightforward [3–5]. Data integration would be made easier if data standards were consistently applied at the source. The Clinical Data Interchange Standards Consortium provides standards for data collection, capture, and representation, to improve accessibility, interoperability, and reusability of data for better clarity in clinical research [6]. Several different frameworks and guidelines have also been developed to assist with tackling issues related to data integration across different studies [7–11]. Other initiatives have focused on developing therapeutic-area specific data standards that could better enable data integration; such as in Polycystic Kidney Disease [12], as well as for over 30 other disease areas [6]. While using these is clearly the way forward for improving integrative research that relies on several data resources, many datasets, particularly legacy data, have not followed such guidelines or applied standards. Projects that want to integrate and use these data are then faced with the burden of aligning to standards; this is not always planned for in advance or resources may be limited. Prospective alignment of datasets, without the availability of

standards, is labour-intensive and often impossible to achieve perfectly. In such cases, little practical guidance, taken from real case experiences, is available to those undertaking this process, with few, if any details given where similar work has been carried out. To begin with, understanding the content of large datasets and then aligning variables to a common data model, based on similarity is extremely time-consuming. This process of alignment is undertaken manually due to nuanced complexities that require human interpretation and knowledge with regards to differences in the way data is captured and identified within the different studies being integrated. The successful process requires good documentation and comprehensive data dictionaries, explaining the content of table fields and outlining calculations that were made. Furthermore, the complexity of the integration problem only becomes apparent once alignment has taken place. In some medical fields, for example: cancer, outcome measures are well-defined clinically and recorded in a standard universal way. However this is not the case in many more complex and multi-faceted diseases, such as lupus. Differences in the response measures taken across studies can affect how response (e.g. remission and low disease activity) can be defined in the integrated set; responders may be defined using a response measure in one study that was not collected in another study. Projects such as MASTERPLANS must therefore tackle this additional level of complexity that is inherent in the study of conditions such as SLE.

We propose that prospective alignments of unstandardised data is achievable with limited resource when specific research questions are used to direct which data are to be integrated across studies.

Here we describe our own experience applying this approach in the integration of data from four lupus studies. These studies included i) The Aspreva Lupus Management Study (ALMS) study, a prospective randomized trial aimed at assessing the efficacy and safety of long term mycophenolate mofetil (MMF) compared to azathioprine and cyclophosphamide in patients with SLE [13]; ii) The LUNAR randomized, placebo-controlled trial evaluating the efficacy and safety of rituximab [2]; iii) EXPLORER, a second randomized, placebo-controlled of rituximab with background treatment distributed among azathioprine, mycophenolate mofetil and methotrexate [1]; and iv) British Isles Lupus Assessment Group Biologics Register (BILAG BR) a national patient registry looking at the safety and effectiveness of biologic and bio-similar treatment for SLE [14].

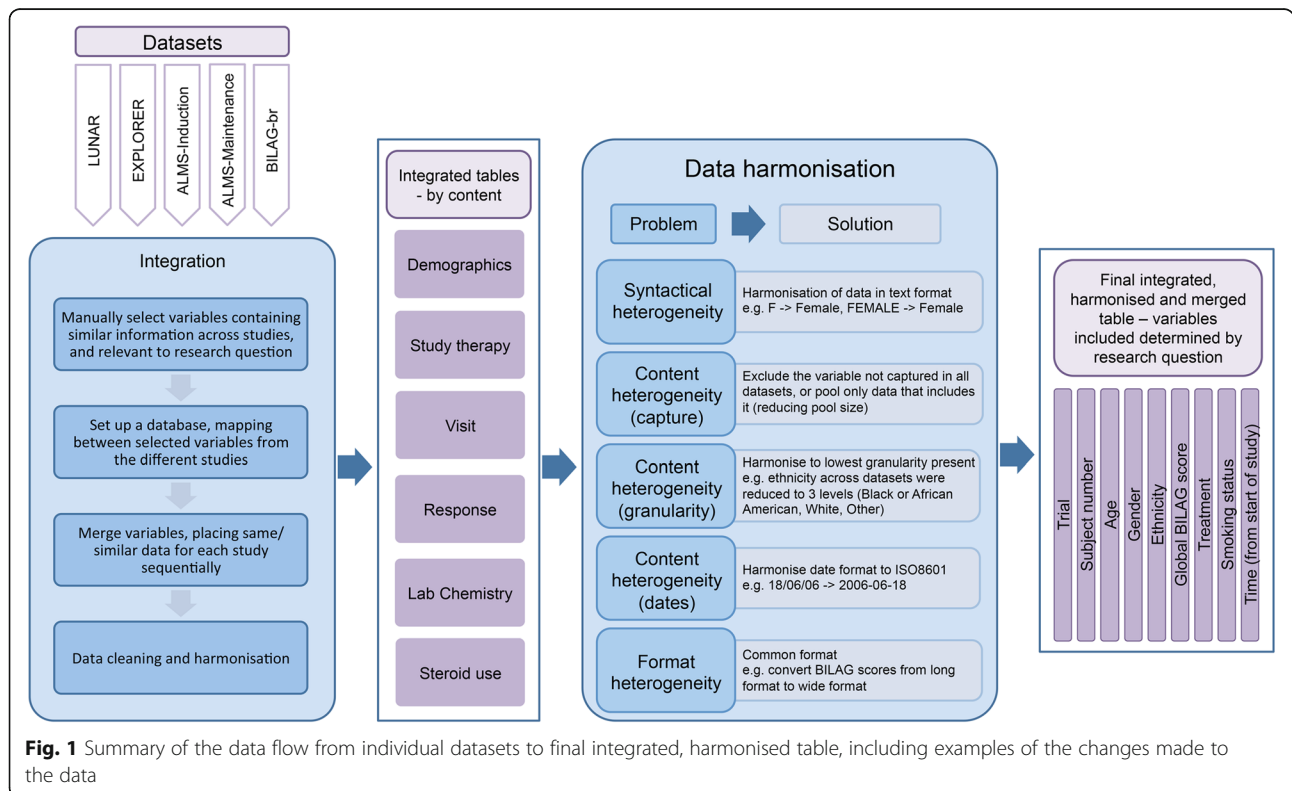
In order to carry out analyses on groups of patients across more than one study, relevant and overlapping variables were extracted, harmonised and combined into common tables, taking a common data model approach.

To simplify and make this laborious process more efficient, only variables likely to be relevant to answering specific, predefined research questions regarding treatment effects and patient characteristics were extracted, so that the amount of data to be handled was kept to a minimum. Variable choice was driven by research questions related to the MASTERPLANS project plan.

Existing documentation detailing the data tables and available variables was used to identify relevant and similar variables from each study and to ascertain which tables from each dataset should be extracted. Unfortunately, the different studies did not use similar data coding standards, making data extraction more complex. These raw data tables, containing variables of interest, underwent initial cleaning, removal of duplicate rows and superfluous variables. Each of these files were read into Matlab and organised to reflect the content of the data and structure of the tables. Separate structures were thereby created for demographics, visit variables, study drug variables, use of steroids, response variables and lab chemistry variables. To integrate across studies, the variables that contained the same or similar information from each study were identified manually and assigned to the same column in merged data-frame. Each variable was then extracted from the raw data tables for each study and assigned to the correct column, placing the data for each subsequent study sequentially

in the same column. These data were then harmonised and cleaned further.

Lengthy processes were carried out to integrate and harmonise the data, so that each trial contained comparable information that could be analysed together (Fig. 1). The general procedure outlined here, can be used as a more generalised guide for how to carry out similar work in the future. We encountered and tackled several issues around data heterogeneity. These could be categorised into types: (i) syntactical heterogeneity, where the meaning of the data captured is the same across sources, but the words used to capture the information are different between different datasets; (ii) content heterogeneity (capture), where a whole variable is captured in one study and not in another; (iii) content heterogeneity (granularity); where, for example in ethnicity, some datasets include more categories and subsets than others, or where visit time is captured as sequential visit numbers (i.e. visit 1, visit 2 etc) in one study but as time from baseline (in days, weeks, or months) in another; (iv) format heterogeneity (variable) where the same information, for example dates, is captured in different formats; and (v) format heterogeneity (dataset/table), where, for example, a variable is captured in a single row per patient across many columns in one dataset (wide format), but in many rows per patient and in one column in another dataset (long format). In addition,



further steps were taken to keep only those patients who had complete and accurate information. For example, rows with missing or mistyped dates were removed, patients were excluded if they were missing key information (e.g. treatment, follow up visits, British Isles Lupus Activity Group - BILAG response measures) and patients were excluded if they were not on their first biologic treatment. One limitation of this approach is the inevitable loss of information either due to differences in granularity or data capture; another is that the final patient group may suffer from selection biases. For example, patients on a second biologic may have more severe and/or more refractory disease and so excluding this group changes the attributes of the patients who are being analysed. Furthermore, excluding patients and rows of data based on missingness reduces the sample sizes and again, introduces potential biases.

## Conclusions

Our integration and harmonisation across SLE studies took a considerable length of time and is very specific to the data provided to MASTERPLANS. A more flexible, automated approach would significantly benefit future similar projects. Many of the steps carried out to generate the harmonised data resource required manual evaluation and examination of the data by an experienced data analyst; this is a tremendous waste of resource that is not uncommon in such projects. The work carried out here highlights the importance of standardisation, particularly regarding validated measures of disease response across both clinical trials and patient registries, at the clinical level. Application of standards for recording data at the source, measures that need including, and the format of variables, will facilitate both the integration of large datasets and comparison between studies. Significant efforts are being made to implement standards, such as the Fast Healthcare Interoperability Resources (FHIR) [8], across healthcare and medical research; however many data resources, especially legacy datasets, remain unstandardised. Where standards are not implemented at the source, reality dictates having to make a compromise in setting the approach; we argue that this can be made easier when specific research questions are used to direct which data are to be integrated across studies.

Similar integration work would benefit from the input of a data management specialist at the earliest stages in the conception of a project or trial. This would also allow for standardisation of the resulting integrated dataset, benefiting future investigations.

Despite the limitations, this work provides a useful, generalised procedure to address the known complexities in the integration of large datasets. Three key approaches were vital to the success of this work.

Systematic preparation regarding the data alignment before starting to integrate was essential. A flexible approach, enabling the addition of new variables and new datasets, meant that the resulting output could be updated easily to answer new research questions if required. Finally, a focus on specific, well-defined, research questions meant that the dataset size remained manageable and was tailored by design for its intended use.

## Abbreviations

SLE: Systemic lupus erythematosus; MASTERPLANS: MAXimizing Sle Therapeutic Potential by application of novel and stratified approaches

## Acknowledgements

The authors would like to thank Prof Niels Peek and Dr. Hilary Dexter for useful discussion and helpful suggestions; and Dr. Patrick Doherty for assistance with data integration. The authors would also like to thank the BILAG-br, EXPLORER, LUNAR and ALMS studies for making the data available for integration and subsequent research.

## Authors' contributions

HLS carried out the work presented in this article. HLS, INB and NG wrote the manuscript and contributed to valuable discussion. All authors read and approved the final manuscript.

## Funding

This work was supported by the Medical Research Council grant MR/M01665X/1 "Maximizing SLE therapeutic Potential by Application of Novel and Systemic Approaches and the Engineering (MASTERPLANS)", and by the Medical Research Council and the Physical Sciences Research Council grant MR/N00583X/1 "Manchester Molecular Pathology Innovation Centre (MMPaThC): bridging the gap between biomarker discovery and health and wealth"; and supported by researchers at the NIHR Manchester Biomedical Research Centre. The funders had no role in the design of the study, the collection, analysis, and interpretation of data, or in writing of the manuscript.

## Availability of data and materials

To access data available to the MASTERPLANS project (<http://www.lupusmasterplans.org/home.html>), please contact Thomas Schindler at Roche ([thomas.schindler@roche.com](mailto:thomas.schindler@roche.com)) for LUNAR and EXPLORER, Neil Solomons at Vifor/Aurinia ([nsolomons@auriniapharma.com](mailto:nsolomons@auriniapharma.com)) for ALMS, and Emily Sutton ([Emily.Sutton@manchester.ac.uk](mailto:Emily.Sutton@manchester.ac.uk)) for BILAG-BR. To access the documentation detailing the harmonisation process, please contact Dr. Patrick Doherty ([Patrick.A.Doherty@manchester.ac.uk](mailto:Patrick.A.Doherty@manchester.ac.uk)).

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

None.

## Author details

<sup>1</sup>Centre for Health Informatics, Vaughan Housue, Portsmouth St., The University of Manchester, Manchester M13 9GB, UK. <sup>2</sup>Arthritis Research UK Centre for Epidemiology, The University of Manchester, Manchester, UK. <sup>3</sup>NIHR Manchester Biomedical Research Centre, Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK. <sup>4</sup>The Manchester Molecular Pathology Innovation Centre, The University of Manchester, Manchester, UK.

Received: 23 September 2019 Accepted: 19 June 2020

Published online: 24 June 2020

## References

1. Merrill JT, Newwelt CM, Wallace DJ, Shanahan JC, Latinis KM, Oates JC, Utset OT, Gordon C, Isenberg DA, Hsieh HJ, et al. Efficacy and safety of rituximab in moderately-to-severely active systemic lupus erythematosus: the randomized, double-blind, phase II/III systemic lupus erythematosus evaluation of rituximab trial. *Arthritis Rheum.* 2010;62(1):222–33.
2. Rovin BH, Furie R, Latinis K, Looney RJ, Fervenza FC, Sanchez-Guerrero J, Maciuga R, Zhang D, Garg JP, Brunetta P, et al. Efficacy and safety of rituximab in patients with active proliferative lupus nephritis: the lupus nephritis assessment with rituximab study. *Arthritis Rheum.* 2012;64(4):1215–26.
3. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials.* 1986;7(3):177–88.
4. Doiron D, Burton P, Marcon Y, Gaye A, Wolffenbuttel BHR, Perola M, Stolk RP, Foco L, Minelli C, Waldenberger M, et al. Data harmonization and federated analysis of population-based studies: the BioSHARe project. *Emerg Themes Epidemiol.* 2013;10(1):12.
5. Fortier I, Burton PR, Robson PJ, Ferretti V, Little J, L'Heureux F, Deschenes M, Knoppers BM, Doiron D, Keers JC, et al. Quality, quantity and harmony: the DataSHAper approach to integrating data across bioclinical studies. *Int J Epidemiol.* 2010;39(5):1383–93.
6. Clinical Data Interchange Standards Consortium. <https://www.cdisc.org/>. Accessed 3 June 2020.
7. Ohmann C, Banzi R, Canham S, Battaglia S, Matei M, Ariyo C, Becnel L, Bierer B, Bowers S, Clivio L, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. *BMJ Open.* 2017;7(12):e018647.
8. Fast Healthcare Interoperability Resources [<https://www.hl7.org/fhir/index.html>].
9. Deshpande P, Rasin A, Brown E, Furst J, Raicu DS, Montner SM, Armato SG. Big Data Integration Case Study for Radiology Data Sources. Montreal, QC: IEEE Life Sciences Conference (LSC); 2018. p. 195–8.
10. Yu HQ, Zhao X, Deng Z, Dong F: Healthcare-Related Data Integration Framework and Knowledge Reasoning Process In: International Conference on Knowledge Management in Organizations: 2017; 2017: 386–396.
11. Teodoro D, Choquet R, Schober D, Mels G, Pasche E, Ruch P, Lovis C. Interoperability driven integration of biomedical data sources. *Stud Health Technol Inform.* 2011;169:185–9.
12. Perrone RD, Neville J, Chapman AB, Gitomer BY, Miskulin DC, Torres VE, Czerwiec FS, Dennis E, Kislner B, Kopko S, et al. Therapeutic area data standards for autosomal dominant polycystic kidney disease: a report from the polycystic kidney disease outcomes consortium (PKDOC). *Am J Kidney Dis.* 2015;66(4):583–90.
13. Dooley MA, Jayne D, Ginzler EM, Isenberg D, Olsen NJ, Wofsy D, Eitner F, Appel GB, Contreras G, Lisk L, et al. Mycophenolate versus azathioprine as maintenance therapy for lupus nephritis. *N Engl J Med.* 2011;365(20):1886–95.
14. McCarthy EM, Sutton E, Nesbit S, White J, Parker B, Jayne D, Griffiths B, Isenberg DA, Rahman A, Gordon C, et al. Short-term efficacy and safety of rituximab therapy in refractory systemic lupus erythematosus: results from the British Isles lupus assessment group biologics register. *Rheumatology (Oxford).* 2018;57(3):470–9.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

