Experience Paper: The Tension between Software Sustainability and Research Design: Visualising Data across Multiple Clinical Trials.

Alaa Alahmadi & Caroline Jay*
School of Computer Science, University of Manchester, UK

Abstract

Whilst it is usual for the requirements for software to come from the domain, in computational science we are moving towards a situation where requirements for the domain—for example, how data is formatted—are starting to come from software. Standardisation of operating procedures and improved data curation are a positive development, allowing the synthesis of multiple datasets and significantly advancing the potential of research. Applied in the general case, however, there is a risk that standardisation might constrain the research process. Here we describe our experience developing software to support meta-analysis in clinical domains, a process made very challenging by the diversity in data formats and methods. We reflect on the tension between standardising operating procedures to assist with data synthesis, and the constraints this may place on study design, and consider how to manage this process, such that software is sustainable, yet researchers retain autonomy in conducting their research.

1 Introduction

In recent years great progress has been made towards enhancing the reproducibility and reusability of computational research through initiatives to deal with diversity in data, and the development of principles to support reuse of data in the future. Examples of these include the FAIRDOM[1] and Open PHACTS[2] projects, which bring together data from systems biology and pharmacology respectively, in integrated, interoperable infrastructures. Recommendations for organising data have also been formalised in the FAIR Guiding Principles for scientific data management and stewardship [8], designed to enhance the reusability of datasets. A particular focus of the FAIR Principles—which aim to ensure data is Findable, Accessible, Interoperable and Reusable—is on enhancing the ability of machines to automatically find and use data, in addition to supporting its reuse by individuals.

Thus far, the focus of such initiatives has been on specific scientific domains, within which operating procedures and outputs may be expected to have some consistency. The presence of interoperable datasets has the potential to add value to research in any domain, so there is an argument for standardisation of operating procedures and data formats in the general case. With any form of standardisation, however, there is a curtailing of freedom, which places limits on the way in which research is conducted. Here we discuss our work developing software to support visualisation of data from multiple clinical trials—an aim of which is to aid meta-analysis—and

*caroline.jay@manchester.ac.uk
[1] http://fair-dom.org
[2] http://www.openphacts.org

reflect in particular on the extent to which the consistency in data formatting that would be required to make the software sustainable (or even viable) would limit researchers in terms of how they design their experiments.

2 Visualising data to support meta-analysis

Evidence-based medicine is the process of integrating individual clinical expertise with external evidence from systematic research [5]. An important tool in gathering evidence is meta-analysis, the process of comparing multiple data sets from different studies, described by Glass (1976) as, 'the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings' [2]. Meta-analysis offers a framework for the integration of multiple clinical trials, in order to systematically review the outcomes from a quantitative perspective, with the ultimate aim of determining the effectiveness of a drug or treatment through consideration of all the available evidence. Whilst single clinical trials might be too limited in scope to come to unequivocal findings or generalisable outcomes about drug efficiency [1], examining data from multiple trials may provide a more accurate estimate of the effect of a drug or a risk factor for disease, than can be given by any individual clinical trial contributing to the pooled assessment [3].

In order to perform meta-analysis effectively, study procedures and results must be comparable. At present the process of determining whether it is appropriate to compare the results of trials is performed manually, and involves a painstaking trawl through the study protocols and data. Even when studies purport to examine the same issue, in many cases meta-analysis is not possible, due to the heterogeneity of the data and unexplained inconsistencies between trials [7]. As a result, research communities are interested in tools that will help to assess heterogeneity across clinical trials, in order to understand whether it is appropriate to include them in a meta-analysis.

A current collaboration between the University of Manchester and the pharmaceutical company AstraZeneca is developing software to visualise equivalent variables in openly available clinical trial datasets from Project Data Sphere[3], to help determine the extent to which it would be appropriate to compare datasets systematically. The primary focus of the software is on comparison of treatments for metastatic (stage IV, or secondary) breast cancer, a particularly challenging area of clinical research. One of the most serious problems is the low overall survival (OS) rate for patients with metastatic breast cancer. Another serious issue is caused by the use of chemotherapy, which is more effective than hormonal therapy, but has a higher toxicity. Data visualisation would be particularly useful for exploring discontinuity across clinical trials, as a result of adverse events caused by such treatment.

Thus far, there has been very little work looking at the problem of cross-trial visualisation. The closest progress towards this has been made with the CTeXplorer tool, which was designed to visualise design heterogeneity in trials of mother-to-child HIV transmission in terms of eligibility criteria, sample size, intervention details, and study outcomes and results [4]. Manual curation of data files was required before it was possible to display the results visually, a phenomenon that we also experienced, as described below. This tool was constructed as a prototype, and is not presently accessible.

[3] `https://www.projectdatasphere.org/projectdatasphere`

3 A tool to visualise data from multiple clinical trials

A prototype tool to visualise data from multiple clinical trials is currently under development. In this paper, we focus specifically on the issues caused by the diversity in data recording and study design across the trials. The source code and further details of the tool's implementation can be found in our online repository[4].

3.1 Data transformation

Project Data Sphere contains SAS files for each trial, alongside information describing the study. The first step was to read the data dictionary file if it existed. This file provides the variable types and definitions, and the names of the files where variables can be found. Data dictionary files are not always provided, in which case it was necessary to consult other files such as the trial protocol and case report forms. Following this, all data files were examined manually to uncover any additional variables, and determine the relationship between them, as well to clarify the variable values and units of measurement.

After a human understanding of the clinical features and the data formats was obtained, three further steps followed to convert the data to a form suitable for visualisation. During the *data wrangling* phase, the files were converted from SAS to a CSV format, and variables were recoded and transformed, often manually, so they were in an equivalent format across trials. In the *data integration* phase, the cleaned data were written to a single CSV file using the opencsv Java library, which was then loaded into a SQL database in the *data querying* phase. Query results were converted to CSV to work with the D3.js library, which was used for visualisation. The process is illustrated in Figure 1.

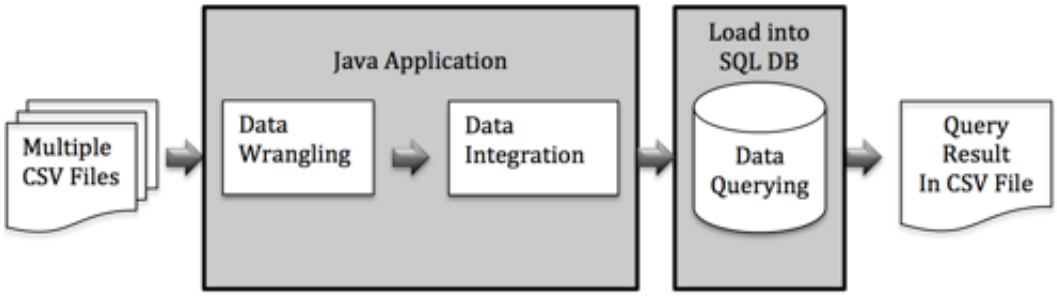

Figure 1: The data processing pipeline.

Figure 2 shows the results of visualising age from five different trials, after following the steps described above. Age was represented in disparate ways across the trials (see Section 3.2), and data required cleaning and recoding prior to visualisation.

3.2 Difficulties dealing with heterogenous data

During the development process, many difficulties were encountered when trying to determine equivalence between datasets sufficiently to visualise them simultaneously. Some of the key problems, along with examples, are listed below:

1. The data in each file are often in numbers or codes, so it is necessary to read a lengthy protocol, and often a number of other files, to understand what they mean. For example, ethnicity

[4] https://github.com/Alaa26Alahmadi/ClinicalTrialsVis

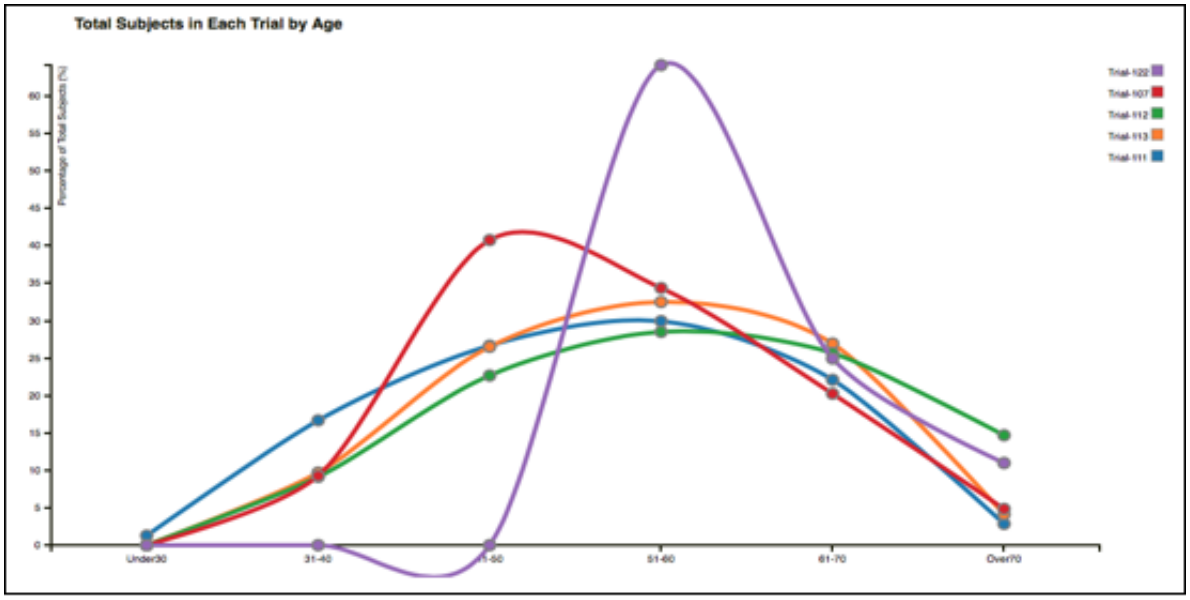

Figure 2: The percentage of patients in each age range across five clinical trials.

may be recorded using the numbers 0-6 in one trial, but with strings such as “asian”, “white” in another, with no standard use of or meaning for these values.

2. Data is organised in a different way in every trial. One trial may keep demographic data in three different files. Another may combine demographic data with an outcome variable such as “adverse events” in a single file.

3. The same feature may have a different name in different trials. For example, to identify age, one trial uses “AGE”, another uses “demo-004”, and a third uses “AGY”. Thus, there are different headers in different files across trials.

4. Both units of measurement and data type for a given variable may vary across trials. For example, age is measured in years (as an integer) in one trial, years and months (as a float) in another file, or defined as a certain range (using a string, e.g. “30-40”) in another.

5. It can be difficult to map data across studies. In one trial, the patient status variable is described using 10 different values (e.g. “adverse event”, “refused to continue”, “lost to follow up” etc.), while in another trial patient status is described as “alive”, “dead”, or “withdrawn”. Combining values in these cases made lead to a loss of accuracy in the data.

6. Values that may appear on immediate inspection to be equivalent, may in fact be qualitatively quite different. For example, one trial may record patient status each time a patient visits the clinic, leading to 11 records for the same patient with different or missing values. Another may include it only once, and treat it as an outcome variable. In this situation, comparing the variable across trials may not be appropriate.

7. Assumptions had to be made about the data, leaving open the possibility that these assumptions were not correct. Whilst the protocol generally described the contents of the files, it was not always clear, and it was sometimes necessary to ‘piece together’ what a variable name or value meant.

4 Discussion

If we consider sustainable software to be software that can be reused, in whole or in part, in future projects, a basic characteristic of a tool to visualise multiple datasets would be that it is possible to compare equivalent variables (for example, age of participants, or survival rate following treatment) automatically. Our experience developing a tool to perform this function for treatments for metastatic breast cancer has shown that, in reality, datasets require significant manual manipulation and the goal of being able to automate the process of loading and comparing files is some way off. Although the clinical studies examined followed the same randomised controlled trial protocol, which is the well-established and tightly defined gold standard for evaluating treatments, subtle variations within the study design and data format made automatic comparison impossible. Although the requirements for this software represent only a subset of the much bigger problem of determining automatically whether meta-analysis is appropriate (and ultimately automating the performance of such analysis), the process of trying to synthesize data from just a small number of datasets has already produced some significant challenges.

If it is not possible to automate the mining of clinical trial data, creating software that is able to deal with this data in a sustainable fashion becomes extremely difficult. An approach to facilitating this is to be proscriptive in terms of how data are formatted, by defining standards for interoperability [8]. Promoting the machine readability of data is very attractive from the perspective of software development, but it may carry some disadvantages from the perspective of the researcher. One is the significant burden associated with data curation [6]. Another consideration, which has received less attention in the literature, is the constraint that standardisation may place on the study design. Our own experience of developing software to visualise clinical trial data, has shown that there are instances where it is reasonable to expect data to conform to standards, but also that this has the potential to interfere with the research process.

A key issue identified with trying to integrate data from across the studies, was the lack of a standard way of describing it. The most significant difficulties arose when variables were not really defined at all: there was no data dictionary to describe the content, format or structure of the files, and the meaning of a column header or value had to be inferred from reading the protocol and cross-comparing files. This type of inference introduces the possibility of error, so comprehensive metadata are important to ensure results are not misinterpreted. To aid software development, the form of that metadata should be as consistent as possible.

There was also significant diversity in the representation of data that may be considered notionally equivalent. Age and ethnicity, for example, were categorised in different ways, and represented using a wide variety of labels and data types. Certain recommendations, such as not representing a quantitative value such as age with a string, could be considered best practice, and would not appear to interfere with a study design. Whether age should be interval or continuous data is less clear. Multiple measurements may be taken over the course of a trial, and age may take a different value in each. Consistently taking age at the start of the trial, and then date-stamping every measurement, would make this value easier to interpret across trials, and would still provide the information required for the research, but it may introduce an overhead, in terms of having to transform the variable for analysis. There is also, at present, no obvious way to define best practice: what may work for nine trials may not work in the tenth. The way we define variables also changes over time; in recent years gender, for example, has gone from having two values, to having four or more.

Producing sustainable software that can deal with data from multiple trials requires consistency in the way that variables are defined across those trials. There are certain basic standards, such as the provision of comprehensive metadata, that it would seem reasonable to expect in a clinical study: if data files are not described adequately, this harms not only their machine readability, but also their human readability, and therefore means they may be interpreted or used incorrectly. Beyond this starting point, the picture becomes more complicated. Whilst it may be possible to introduce practices that would support cross-comparison of trial data that also allow freedom in research design, the form that these practices should take is not clear at present, and determining them will require iterative refinement. A crucial part of this process is dialogue between researchers and software engineers, to ensure that sustainable software tools can enable transformative science, without constraining it in the process.

Acknowledgements

With thanks to James Weatherall at AstraZeneca for his input into this project.

References

[1] R. DerSimonian and N. Laird. Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials*, 45:139–145, nov 2015.

[2] G. Glass. Primary, secondary and meta-analysis of research. *Educational Researcher*, 5:3–8, 1976.

[3] A. B. Haidich. Meta-analysis in medical research. *Hippokratia*, 14(Suppl 1):29–37, 12 2010.

[4] M.-E. Hernandez, S. Carini, M.-A. Storey, and I. Sim. An interactive tool for visualizing design heterogeneity in clinical trials. *AMIA Annual Symposium Proceedings*, 2008:298–302, 2008.

[5] D. L. Sackett, W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1996.

[6] C. Tenopir, S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame. Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6(6):1–21, 06 2011.

[7] A. Whitehead. *Meta-Analysis Of Controlled Clinical Trials*. John Wiley & Sons, Ltd, jul 2002.

[8] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018 EP –, 03 2016.