Translational Medicine

Data Standards are needed to move Translational Medicine forward

Ann Marie Martin[1]*, Nathalie Seigneuret[1], Ferran Sanz[2,3] and Michel Goldman[1]

[1]Innovative Medicines Initiative, Belgium

[2]Research Programme on Biomedical Informatics (GRIB), IMIM, DCEXS, Pompeu Fabra University, Barcelona, Spain

[3]Spanish Institute of Bioinformatics (INB), Spain

Abstract

Data standards and data management procedures help to overcome the data and information management challenges in translational medicine for the purpose of deriving new results and knowledge. In regulated medicine, the reason for using data standards is to facilitate the review, verification and replication of the research results as part of a drug or device approval process. Additionally, data standards are indispensable in enabling ongoing safety and quality surveillance of marketed products. Data management procedures must be adhered to and documented to ensure the integrity and verifiability of the research results. Lack of data standards, appropriate data management procedures and absence of good documentation make it hard and labour intensive to combine data in order to answer new research questions. Fortunately, many data standards are available which are well-recognized and certified and which can be readily applied. We argue that it does not matter which data standards are chosen as long as well-recognized data standards are used and the data management procedures are documented and referenced in a citable way in the write-up of the study results and publications. Data could be used in far more powerful and efficient ways when one observes these principles. For this reason, the Innovative Medicines Initiative (IMI) is actively encouraging its projects to both use and contribute to data standards.

Keywords: Data standards; Data management procedures; Translational medicine; Regulatory medicine

Why Everyone Should be Interested in Data Standards

In translational research, the challenge is to combine data derived from clinical records and "omics" technologies for the purpose of new research [1]. The data are typically held in different databases using different data models and conventions leading to difficulty integrating the data, both at the level of the research unit (e.g. at the patient-level) and/or at the level of the data fields and coding schemes utilized when adding similar data from e.g. another region.

Computers simply cannot deal with even small discrepancies in the identifiers of the patient, such as an identification number, name, birthdate and address, without requiring a person intervention confirming whether or not the data belong to the same individual. Apart from this obvious problem, we have an additional problem in that often data within the same field, whether it is genetics, clinical research, or healthcare data, may be held in databases using a multitude of different data models and no standards at all. Computers cannot determine whether fields and coding schemes are the same if they are named differently, nor can they distinguish whether two identically named data fields are truly the same or not. The problems are worsened due to differences in the patient's data when similar information is held in multiple databases due to potential discrepancies. Again it requires human intervention to identify the correct data. Further compounding the problem, an individual standard may be implemented somewhat differently from organization to organization. Worst of all is the situation where the database is undocumented whether in terms of content or quality. The result is that scientists often need resource intensive, expensive one-off computer programs to help them to get through the work of combining the databases in a reasonable manner. Even when data standards are used it turns out that they are often badly documented, forcing the data scientist to carefully review the data model versus the standard before proceeding to the task of combining the data.

This leads to the conclusion that computers are pretty useless unless standard data models and coding schemes are rigorously used and documented in order to combine the datasets. This is why everyone should be interested in Data Standards.

Of course this problem of interoperability of data is not unique to the medical sciences. All data-intensive sciences have to deal with the data management and interoperability problem. In 2009, the electronic Infrastructure Reflection Group (e-IRG) Data Management Task Force came up with a list of recommendations for appropriate data management [2]. The recommendations broadly come under 3 headings: 1) the need for descriptive metadata; 2) the need to describe the quality of the data; and 3) the use of data standards to ensure technical and content interoperability.

The good news is that many data standards are available even if there are too many to choose from [3]. We argue that it does not matter which one is used, as long as the standard used is well recognized and the data management procedures are documented and referenced in the write-up of the study. It is better to use a citable data standard than to use no data standard at all or to create a new one. Once it is clear which standards have been used, it is usually possible to generate standardized mapping programs from one data standard to another, which can be reused for different data poolings.

In regulatory medicine, the applicant for a product is required to provide a summary of safety and efficacy based on all data combined. The regulators worldwide require that for verification and reproducibility purposes, individual patient data be reported and submitted according to the International Conference of Harmonization Guideline E3 [4]. There is guidance as to the technical (SAS® version 5 transport files or XML) and content format (mentioning the CDISC SEND, SDTM and ADAM standards) [5-7] for the data, and some guidance as to

*Corresponding author: Ann Marie Martin, Principal Scientific Manager Knowledge Management, Innovative Medicines Initiative, Postal mail: IMI JU, TO 56, 6th floor, B-1049 Brussels, Spain, E-mail: ann.martin@imi.europa.eu

how to ensure that the data standards are used correctly. To this end, software tools are used (e.g. WEBSDM, OpenCDISC, SAS® macros) [8]. Furthermore, data management procedures must be fully documented.

The Innovative Medicines Initiative (IMI) is a public private partnership between the European Union and the European Federation of Pharmaceutical Industry Associations that supports collaborative projects to speed up the development of better and safer drugs for patients. The IMI concluded in 2011 a memorandum of understanding with and membership of CDISC (Clinical Data Interchange Consortium) [9], a standards development organization, to facilitate the data management tasks of the collaborative projects it supports. At the same time, IMI projects are encouraged to contribute to emerging standards by contributing new standards to CDISC. Of course, CDISC is not the only data standards development organization and focuses only on the domain of clinical research. Yet by emphasizing the need for using data standards, documenting the data management procedures and making data management work referenceable, IMI encourages best practise in the projects. The authors believe that these principles are key to all collaborative translational research.

In short, data standards are essential components of translational research. For this reason, we suggest that careful attention be paid to data management procedures, both in the design of research studies and in the ensuing reports and publications, and from basic research studies to healthcare implementation studies.

References

1. M. Cases, L. I. Furlong, J. Albanell, R. B. Altman, R. Bellazzi, et al. (2013) Improving data and knowledge management to better integrate health care and research. Journal of internal medicine 274: 321–328.

2. Data Management Task Force (2009) "e-IRG Report on Data Management".

3. I. N. Sarkar (2010) "Biomedical informatics and translational medicine" Journal of translational medicine.

4. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (1995) "Sructure and Content of Clinical Study Reports E3".

5. US Food and Drug Administration (2013) "Study Data Standards for Regulatory Submissions".

6. US Food and Drug Administration (2011) "Study Data Standards Resources".

7. US Food and Drug Administration (2012) "Study Data Specifications".

8. US Food and Drug Administration (2010) "Study Data Submissions to CBER".

9. CDISC (2013) "Clinical Data Interchange Standards Consortium: Standards & Implementations".