Drug Saf (2014) 37:893–896
DOI 10.1007/s40264-014-0221-4

COMMENTARY

Converting to a Common Data Model: What is Lost in Translation?

Commentary on "Fidelity Assessment of a Clinical Practice Research Datalink Conversion to the OMOP Common Data Model"

Peter R. Rijnbeek

Published online: 4 September 2014

Observational drug safety and effectiveness studies provide important evidence for clinical practice. However, meta-analysis of these studies requires extensive efforts and quality control. Although observational studies at first glance may seem to address the same underlying question, differences in study design, definitions, and settings often limit our ability to compare and synthesize. Study designs may vary, sometimes with clear differences, but often with subtle nuances. Different observational data sets may have different biases. Data collected in routine care often reflect different healthcare delivery systems. As a result, an individual researcher trying to synthesize available evidence truly faces a challenge.

In an ideal world, a more harmonized method would be available by which data and results from different databases could be combined to answer a specific question. An overarching, generic protocol should drive pooling of results from multiple healthcare databases. Standard extraction and transformation tools together with common analytical tools should be part and parcel of each study. All these steps should be fully transparent, fast and reproducible.

In a recent review, Trifiro et al. [1] discuss the international projects and initiatives, funded either on a project basis or by governments, that aim to stimulate research with multiple observational databases in order to provide rapid responses to questions regarding benefits and risks of drugs. All aim to develop methods to combine databases with different underlying data models, different types of information collected, and different coding systems. The primary motivation for this work is not only to increase statistical power, but to also exploit heterogeneity of exposure in the different data sources. In addition, there is a clear regulatory incentive. The European Medicine Agency (EMA), for example, often requires post-authorization safety studies (PASS) to be conducted in multiple countries. However, Europe is characterized by heterogeneity regarding healthcare delivery systems, language barriers, and different coding systems.

To perform drug safety studies on multiple databases, different approaches have been developed [1]. In all projects, a distributed network design is chosen. In such a design, local involvement of experts is needed to convert data. One approach is a study-specific transformation of local data. That is, the needs of a specific study provide the context and boundaries for the transformation of data. Typically, a limited set of data is prepared locally in a common format. That set then is the basis for analysis. Although experience for previous studies could be re-used, the objective is not to transform as much data as possible, but only the required set. For example, in the EU-ADR project, a custom built tool (Jerboa) was applied on common input files stemming from various data sources [2]. Aggregated output generated by Jerboa allowed for pooling of results and subsequent analysis.

In this issue of *Drug Safety*, Matcho et al. [3] follow another approach: their objective is to convert all CPRD data to the OMOP common data model independent of the requirements originating from a specific study. On this generic common data model, all subsequent extractions and analysis are performed. A common data model aims to achieve both syntactic and semantic interoperability [4]. Syntactic interoperability refers to the common underlying data structure which enables exchange of data between different sites. Syntactic interoperability focuses on the

P. R. Rijnbeek (✉)
Department of Medical Informatics, Erasmus Medical Center, Rotterdam, The Netherlands
e-mail: p.rijnbeek@erasmusmc.nl

grammar in which data is described. Although the same grammar might be used, the meaning might differ. The notion of semantic interoperability emphasizes that meaning also needs to be aligned. Semantic interoperability refers to a common understanding that is required to interchange information, i.e., all the data sources are mapped to a standardized terminology system.

The standardization to a common data model has advantages. A common data model allows for fast assessment and utilization of standard analytical tools. As a result, with a common data model, studies could be readily and transparently replicated in different databases. However, a potential limitation is the risk of information loss due to incomplete mapping to the common data model or the vocabulary. Therefore, a thorough evaluation of the extraction, transformation, and loading (ETL) process that specifies how the local data are mapped to a common model is essential to estimate the amount of information lost and the impact of that lost information on study results.

Matcho et al. [3] address this critical issue of transformation of a data source to the OMOP common data model. The authors assess the transformation by judging the completeness of the mappings of the conditions, demographics, and lifestyle data, and exposure data. Furthermore, in an initial evaluation, a published case-control study performed on the original data was replicated in the transformed database. Matcho et al. [3] report that the transformation was of high quality and resulted in a minimum amount of information loss. Moreover, the already published case-control study could be replicated in the transformed database.

A first question is how generalizable these results are to other drug safety use cases. The authors answer this question in their paper to some extent. The authors argue that the ability to generalize their results predominantly depends on the quality of the transformation of both the conditions and the drug exposure. I agree that the replication of an already performed study is important, but the comparison of the prevalence of all the drugs and conditions in the original data with those in the transformed data is the more important contribution of this paper. Matcho et al. [3] show that the transformation from the source data to the common data model results in only minimal differences in the prevalence of both the conditions and the drug exposures. One might therefore argue that any study involving another drug class and health outcome of interest will probably also replicate.

Nevertheless, the devil is always in the details. In particular situations, the chosen ETL process might have an impact on the results of a study. For example, the authors state that, by convention, data outside the patient's valid observation period are not converted to the common data model. One could argue that probably the observation period is called valid for good reasons and that therefore data outside the valid observation period should indeed not be converted to the common data model. This convention, however, limits the ability to exclude patients that have a history of a particular condition in that "not valid" observation period. Another example is the impact of the imputation of the length of exposure using the algorithm described in the paper and the construction of the drug eras with the predefined persistence window of 30 days. It is important to note that these constructed drug eras are not a direct translation from the source data but are derived. Sensitivity analyses may be necessary to judge the effect derivation choices may have on the study result. A researcher who relies on the mapping to the common data model therefore needs to be aware of the applied ETL process and has to reflect on the potential impact on a study.

A second important question is whether the results of this study can be generalized to other data sources. Obviously, the ETL process developed for this particular database (CPRD) cannot be completely re-used by another data source. The underlying data model and data elements in the source database are often so specific that the ETL procedure needs to be tuned considerably when applied to another data source. A paper of Zhou et al. [5] describes the transformation of the THIN database to the OMOP common data model and the impact of this transformation on the study results. They assessed the ETL by implementing a proportional reporting ratio, univariate self-case control series, and a high-dimensional propensity score. The conclusion of the study of Zhou et al. [5] is that incomplete mapping of medical and drug codes limited the use of the transformed database for epidemiological evaluation studies. Both CPRD and THIN, however, are at least in part overlapping (that is, individual general practitioners can contribute data to both THIN and CPRD) and contain similar data. Why do the results of mapping THIN and CPRD to the same OMOP common data model differ? These different results could be due to different underlying data models and data elements, as suggested in the paper of Zhou et al. [5]. Another possible reason could be that the transformation to the common data model in the study of Matcho et al. [3] was performed in close collaboration with the OMOP researchers and thus ensured a more optimal balance between expertise concerning both the source data and the common data model. The papers of Zhou et al. [5] and Matcho et al. [3] underscore the importance of a detailed understanding of the ETL processes that convert a data source to a common data model.

The proof of the pudding will be in the eating. Possibly the best way of assessing the generalizability of the strategy described in the paper of Matcho et al. [3] is to transform more and more databases into the OMOP

common data model, optimize the common data model in iterations if deemed necessary, and repeat already published studies on the transformed database. As different types of databases are mapped to a common data model, the particular characteristics of these data sources will challenge the structure and content of the common data model. Especially, the European databases, with their diverse set of coding systems, languages, and healthcare delivery systems, might be challenging.

As researchers begin to explore the role of common data models to exploit diverse observational data sources and replicate already performed studies, a number of issues must be addressed.

1. It is important to observe that the OMOP common data model anticipates that study-specific conversions might be required. It stores the source data as well as the converted data and allows the creation of "derived variables." These derived variables can be defined in the context of a specific study. That is, inherent to the design of the OMOP common data model is the ability to bypass the conversion through the ETL process and introduce modifications or additions on a per-study basis. Monitoring and analyzing individual researchers' use of this ability to define study-specific variables will be a valuable learning opportunity. The need to create a derived variable can, of course, be the consequence of the specific requirements of a particular study. But the need to create a derived variable could also indicate shortcomings in the original ETL process or limitations of the OMOP common data model.
2. Typically, common data models focus on coded data. Often, however, data are recorded in free text. In our own setting, we work with the Dutch database Integrated Primary Care Information (IPCI), which is based on medical records from general practitioners [6]. In IPCI, free text is available in the form of clinical notes and specialist letters, and is extensively used in drug benefits and risks studies. Natural language processing (NLP) techniques need to be further developed and standardized to leverage textual information and facilitate replication of studies [7].
3. We expect that replication of already published studies will be an important method to assess the quality of a conversion to a common data model. It is important to note, however, that replication of a study relies not only on the quality of the conversion to a common data model, but also depends on the amount of information that can be obtained from the published study regarding the details of the original analysis. A detailed understanding will be needed to distinguish between shortcomings in the mapping of data to the common data model versus differences in the study design and execution.
4. Extraction, transformation, and loading processes are complex and often difficult to understand in detail. In addition, ETL processes require maintenance—both because the structure of the source data may change over time and the common data model will evolve. Ensuring and maintaining adequate quality of the ETL process will need to be supported by adequate software tools. An interesting line of research, for example, could be the development of data visualization tools for this purpose.
5. As the importance of a common data model is recognized, we see that different common data models are proposed by different researchers. As alternative common data models are proposed, their impact on the study results must be assessed. Interestingly, the recent IMEDS research agenda (http://imeds.reaganudall.org/) promises comparisons of different common data models. A study that maps the same raw data set as used by Matcho et al. [3] to the Mini-Sentinel common data model [8] would be an interesting paper.

Finally, as also argued by Matcho et al. [3], a common data model cannot be a substitute for a detailed understanding of source data. At the end of the day, a detailed understanding of underlying source data will be required to ensure that appropriate conclusions are inferred.

Acknowledgments No sources of funding were used to assist in the preparation of this commentary. Peter R. Rijnbeek works in a group that has received funding from European Commission and several pharmaceutical companies. Dr Rijnbeek has no other conflicts of interest that are directly related to the content of this commentary.

References

1. Trifiro G, Coloma PM, Rijnbeek PR, Romio S, Mosseveld B, Weibel D, et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? J Intern Med. 2014;275(6):551–61. doi:10.1111/joim.12159.
2. Coloma PM, Schuemie MJ, Trifiro G, Gini R, Herings R, Hippisley-Cox J, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol Drug Saf. 2011;20(1):1–11. doi:10.1002/pds.2053.
3. Matcho A, Ryan PB, Fife D, Reich C. Fidelity assessment of a Clinical Practice Research Datalink conversion to the OMOP common data model. Drug Saf 2014. doi:10.1007/s40264-014-0214-3.
4. Ogunyemi OI, Meeker D, Kim HE, Ashish N, Farzaneh S, Boxwala A. Identifying appropriate reference data models for comparative effectiveness research (CER) studies based on data from clinical information systems. Med Care. 2013;51(8 Suppl 3):S45–52. doi:10.1097/MLR.0b013e31829b1e0b.
5. Zhou X, Murugesan S, Bhullar H, Liu Q, Cai B, Wentworth C, et al. An evaluation of the THIN database in the OMOP common data model for active drug safety surveillance. Drug Saf. 2013;36(2):119–34. doi:10.1007/s40264-012-0009-3.

6. Vlug AE, van der Lei J, Mosseveld BM, van Wijk MA, van der Linden PD, Sturkenboom MC, et al. Postmarketing surveillance based on electronic patient records: the IPCI project. Methods Inf Med. 1999;38(4–5):339–44. doi:10.1267/METH99040339.
7. Afzal Z, Schuemie MJ, van Blijderveen JC, Sen EF, Sturkenboom MC, Kors JA. Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. BMC Med Inform Decis Mak. 2013;13:30. doi:10.1186/1472-6947-13-30.
8. Platt R, Wilson M, Chan KA, Benner JS, Marchibroda J, McClellan M. The new Sentinel Network—improving the evidence of medical-product safety. N Engl J Med. 2009;361(7):645–7. doi:10.1056/NEJMp0905338.