Decision Support Systems and Education
J. Mantas et al. (Eds.)

doi:10.3233/978-1-61499-921-8-60

Common Data Models (CDMs) to Enhance International Big Data Analytics: A Diabetes Use Case to Compare Three CDMs

Harshana LIYANAGE[a], Siaw-Teng LIAW[b], Jitendra JONNAGADDALA[b], William HINTON[a] and Simon de LUSIGNAN[a,1]
[a] *Department of Clinical & Experimental Medicine, University of Surrey, UK*
[b] *School of Public Health & Community Medicine, UNSW Medicine Australia, Ingham Institute of Applied Medical Research, NSW, Australia*

Abstract. Common data models (CDM) have enabled the simultaneous analysis of disparate and large data sources. A literature review identified three relevant CDMs: The Observational Medical Outcomes Partnership (OMOP) was the most cited; next the Sentinel; and then the Patient Centered Outcomes Research Institute (PCORI). We tested these three CDMs with fifteen pre-defined criteria for a diabetes cohort study use case, assessing the benefit (good diabetes control), risk (hypoglycaemia) and cost effectiveness of recently licenced medications. We found all three CDMs have a useful role in planning collaborative research and enhance analysis of data cross jurisdiction. However, the number of pre-defined criteria achieved by these three CDMs varied. OMOP met 14/15, Sentinel 13/15, and PCORI 10/15. None met the privacy level we specified, and most of the other gaps were clinical and cost outcome related data.

Keywords. Common data models, data harmonisation, Medical record systems, computerized; Diabetes Mellitus; Costs and cost analysis; Organization and administration.

1. Introduction

Routine health databases, based on routine computerized medical record (CMR) systems differ in purpose, content and design. Common Data Models (CDM) can enable the simultaneous analysis of disparate data sources simultaneously.

The Royal College of General Practitioners Research and Surveillance Centre (RCGP-RSC) is one of the oldest sentinel networks in Europe. The electronic Practice Based Research Network (ePBRN) is a network of general practices in South Western Sydney, contributing observational data to the ePBRN data repository. Both research groups have been involved in developing integrated longitudinal data repositories to support diabetes research and informatics.

[1] Corresponding Author, Simon de Lusignan, Department of Clinical & Experimental Medicine, University of Surrey, GUILDFORD, Surrey, GU2 7XH, UK; Email: s.lusignan@surrey.ac.uk

There is a worldwide epidemic of Type 2 diabetes mellitus (T2DM), and research is required to understand whether newer medications that are less likely to cause hypoglycaemia in real world clinical practice. Hypoglycemia in T2DM is important because it may be unsafe for the patient and those around them, result in emergency care, and can precipitate complications including myocardial infarction.

We carried out this study to see which established CDM would provide most insight into conducting our use case study. Namely, the extent to which recently introduced classes of medication for T2DM might achieve good glucose control but less hypoglycemia.

2. Method

2.1. Identification of common data models

We searched Medline using the search term "Common Data Model (CDM)". There were 82 hits which we hand searched. There were 31 papers referring to the Observational Health Data Sciences and Informatics (OHDSI) collaborative's Observational Medical Outcomes Partnership (OMOP) CDM [1], 13 papers referring to the Sentinel CDM (http://www.mini-sentinel.org/); and 3 papers referring to the Patient Centered Outcomes Research Institute Network (PCORNet) CDM (http://www.pcornet.org/resource-center/pcornet-common-data-model/). (Table 1).

Table 1. Types of common data models

CDM Type	Purpose
OMOP CDM	Transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes), and then perform systematic analyses using a library of standard analytic routines that have been written based on the common format.
Sentinel CDM	Create evaluation or assessment protocol, centrally develop the analytic code (i.e., query) and distribute the code to each data partner to run against the data they have stored in a common format.
PCORNet CDM	Defines a standard organization and representation of data for the PCORnet Distributed Research Network.

We also identified other more specialized CDM (e.g. for hematology) and papers that used the term CDM for very specific circumstances and domains or used it incorrectly.

2.2. Development of a T2DM use case

Patient-level data across disparate data sources need to be anonymised (or pseudonymised), extracted and linked securely and confidentially to create analytic datasets to enable the research questions to be answered. The CDM must be able to support the definition of the case and cohort with exposure to the medication under study; identify relevant medications and calculate the "exposure", use standardized terminologies to take advantage of conceptual hierarchies and relationships among the observations, measurements and procedures undertaken. We drew on previous data model evaluation methodology reported by Garza et al [2] and Kahn et al [3,4]. The criteria used included:

1. Business processes criteria:

   a. Comply with privacy regulations;
   b. Comply with data governance, access and use regulations;
   c. Support operational requirements such as ease of ETL (Extract, Transform and Load), querying, ease of implementation, low cost, and stability of the model in terms of user base, maintenance, and frequency of updates.
2. Data-based criteria:
   a. Support pseudonymisation to enable secure re-identification in sources;
   b. Support data linkage within and across care settings e.g. primary & secondary care;
   c. Support provenance and traceability of data e.g. historical encounters from multiple sources of similar and related health information;
   d. Support common health record data domains such as demographics, social history, allergies and immunizations, diagnoses, procedures, problem list, encounters, hospitalizations, deaths, medications, labs, vital signs, and relationships as defined by our data models;
   e. Support controlled terminologies used in healthcare and research;
   f. Support data quality assessment (verification and validation);
   g. Support useful curation of the data and knowledge created via queries.
3. Specific study based criteria: for this study listed below.

To conduct our assessment, we developed a use case: a retrospective cohort study with the following characteristics:

- **Case:** Person with T2DM, for a least a year, with a diagnosis code or elevated HbA1c or fasting glucose diagnostic of diabetes and, treated with at least two medications for their T2DM[5].
- **Exposure:** To a new class of drug not associated with hypoglycemia. International guidance suggests three classes of medication meet these criteria.[6]
- **Outcome measures:** Glycemic control and number of hypoglycemic episodes one year after exposure; and cost.

We then identified in our three CDMs (OHDSI/OMOP, Sentinel, and PCORI) whether they had a field for our data type. As an example, for privacy, we considered whether there are any HIPAA identifiers stored in the CDMs. If a CDM allows to capture identifiable information, we then assessed if it's possible to store this information on the type of data (identifiable or de-identified) in the CDMs in any form including metadata tables of CDMs.

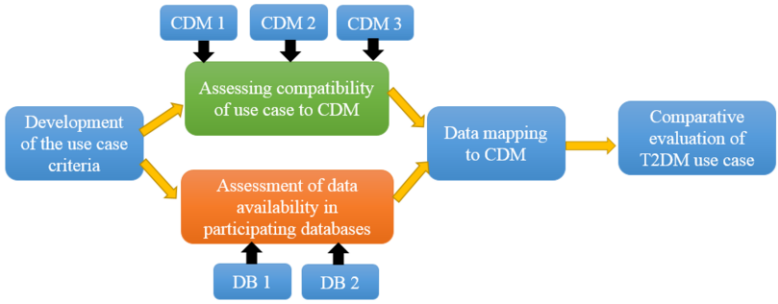

Figure 1. Process for comparing T2DM outcomes using the developed use case

2.3. Evaluation of CDMs and use case compatibility of candidate databases

During the initial exploration we compared the CDMs using the pre-defined criteria discussed above. This was followed by use case based readiness evaluation of RCGP-RSC and ePBRN databases to participate using our T2DM use case. The process we carried out is given in Figure 1. We created an instrument to capture the availability of data elements in the use case and invited to candidate databases to provide their response. In the next phase we plan to map the chosen CDM and conduct the evaluation exercise.

3. Results

From the specified criteria in the use case OMOP met 14/15, Sentinel 13/15, and PCORI 10/15 (Table 2).

Table 2. Ability of considered common data models to map to the use case criteria

	Common data models		
	ODHSI-OMOP v5.1	**PEDSnet v2.9**	**PCORNet v3.9**
Business processes criteria	**Assessment of SOPs and business processes**		
Privacy	No	No	No
Governance	Yes	Yes	Yes
Operational	Yes	Yes	Yes
Data-based criteria	**Assessment of data and data types**		
Pseudonymisation	Yes	Yes	Yes
Data linkage	Yes	Yes	No
Provenance/ traceability	Yes	Yes	Yes
EHR data types: Structured/Unstructured	Yes/Yes	Yes/Yes	Yes/Yes/
Controlled terminologies	Yes	Yes	Yes
DQ assessment	Yes	Yes	Yes
Curation and display	Yes	Yes	Yes
Study specific Criteria	**CDM has specifically identified study data elements**		
Case definition	Yes	Yes	No
Exposure	Yes	Yes	Yes
Clinical outcome measure	Yes	Yes	No
Cost outcome measure	Yes	No	No

The responses for the instrument indicated the following levels of compatibility to the T2DM use cases (Table 3).

Table 3. Availability of data to complete the T2DM use case criteria (A - Captures all data, P - Data captured partially, N - No data is captured).

				ePBRN database	RCGP RSC database
1)	Case	a)	Diagnosis code for T2DM	A	A
		b)	Test results: HbA1c/ fasting glucose	A	A
		c)	Treatment with T2DM medication	A	A
2)	Exposure	a)	Three classes of medication	A	A
3)	Outcome measure	a)	Glycemic control	A	A
		b)	Number of hypoglycemic episode one year after exposure	P	A
		c)	Cost	N	N

4. Conclusions

We found the CDMs compared in this study have different strengths and weaknesses. All lacked the level of privacy we pre-defined. PCORNet did not support data linkage, case definitions and clinical outcome measures whereas the others did. Cost outcome measures could only be mapped to OMOP. Notwithstanding any shortcomings CDMs have a role in large scale data analytics; it may be timely to develop a European CDM. Whilst there is merit in a generic CDM our conclusions based on this exercise is that there should be a CDM core, but allow domain specific extensions to take account of the clinical needs of different domains, such as diabetes – where the high levels of complication from the disease and risks of hypoglycaemia from treatment create a unique disease context.

References

[1] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard MA, R.W. Park, I.C. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G.N. Norén, Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers, *Stud Health Technol Inform* **216** (2015), 574-578.

[2] M. Garza, G. Del Fiol, J. Tenenbaum, A. Walden, M.N. Zozus, Evaluating common data models for use with a longitudinal community registry, *Journal of Biomedical Informatics* **64** (2016), 333-341.

[3] M.G. Kahn, D. Batson, L.M. Schilling, Data Model Considerations for Clinical Effectiveness Researchers, *Medical Care* **50** (2012), S60-S7.

[4] M. Kahn, M. Raebel, J. Glanz, K. Riedlinger, J. Steiner, A pragmatic framework for single site and multisite data quality assessment in EHR-based clinical research, *Med Care.* **50** (2012), S21-9.

[5] A. Rahimi, S. Liaw, J. Taggart, P. Ray, H. Yu, Validating an ontology-based algorithm to identify patients with Type 2 Diabetes Mellitus in electronic health records, *International Journal of Medical Informatics* **83** (2014), 768-778.

[6] S.E. Inzucchi, R.M. Bergenstal, J.B. Buse, M. Diamant, E. Ferrannini, M. Nauck, A.L. Peters, A. Tsapas, R. Wender, D.R. Matthews, American Diabetes Association (ADA); European Association for the Study of Diabetes (EASD). Management of hyperglycemia in type 2 diabetes: a patient-centered approach: position statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD), *Diabetes Care* **35** (6) (2012), 364-379.