MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190218

Transforming Two Decades of ePR Data to OMOP CDM for Clinical Research

Daniel M. Lima[a], Jose F. Rodrigues-Jr[a], Agma J. M. Traina[a], Fabio A. Pires[a], Marco A. Gutierrez[b]

[a] *Institute of Mathematical and Computer Sciences (ICMC), University of São Paulo, São Carlos, São Paulo, Brazil,*
[b] *Heart Institute (InCor), Clinics Hospital, Faculty of Medicine, University of São Paulo, São Paulo, São Paulo, Brazil*

Abstract

This paper presents the extract-transform-and-load (ETL) process from the Electronic Patient Records (ePR) at the Heart Institute (InCor) to the OMOP Common Data Model (CDM) format. We describe the initial database characterization, relational source mappings, selection filters, data transformations and patient de-identification using the open-source OHDSI tools and SQL scripts. We evaluate the resulting InCor-CDM database by recreating the same patient cohort from a previous reference study (over the original data source) and comparing the cohorts' descriptive statistics and inclusion reports. The results exhibit that up to 91% of the reference patients were retrieved by our method from the ePR through InCor-CDM, with AUC=0.938. The results indicate that the method that we employed was able to produce a new database that was both consistent with the original data and in accordance to the OMOP CDM standard.

Keywords:

Medical Informatics; Health Information Exchange; Data Curation

Introduction

In the last decades, the field of Informatics has unveiled the, so-called, Big Data phenomenon; an intense increase of data gathering, exchanging and storage in several human activities. This phenomenon is outlined by the so called five Vs: velocity, volume, variety, veracity and value of data; posed as the big challenges for data analysis and processing [1]. Such characteristics are also common in Medical and Health Information Systems, whose databases have grown into huge amounts of patient information and health-related activities, in diverse formats, always online, and easily accessible from a mobile screen. If properly interconnected and treated, these systems offer interesting data sources for evidence-based research, such as Precision Healthcare, Population Health, Clinical Research, and more. [11]

However, this data exchange is often a significant challenge. Most of the Electronic Patient Record (ePR) systems were not explicitly designed for research; rather, they are organized by standards and structures which are local to the institutions they primarily serve, e.g. hospital facilities, clinics, pharmacies, health insurance companies, etc. Thus, interchange methods, protocols and architectures were designed to cope with this challenge, such as the HL7 standards [12] and the OpenEHR platform [13]. These approaches mediate the communication of near-line and online transaction processing (OLTP) systems, specifying standard data elements and transformations from their internal data to a common messaging format. Furthermore, data analysis follows an approach akin to online analytical processing (OLAP), using denormalized, coalesced and preprocessed data in a standard common database format.

OHDSI and the Common Data Model

In this context, the Observational Health Data Sciences and Informatics (OHDSI – www.ohdsi.org) initiative grew out of the Observational Medical Outcomes Partnership (OMOP) developing a mature data standardization model, the OMOP Common Data Model (CDM) [6]. Having a ready-to-use database in a standard common model such as the OMOP CDM simplifies the exchange and integration of standardized methods, applications, information and tools between clinical researchers; a critical feature for distributed research networks using patient-centric clinical databases. [7]

The CDM is a strong information model; its conceptual elements and their relationships are explicitly specified in a formal language, and every piece of information is connected to a standard term from SNOMED-CT. The CDM's Standard Clinical Tables include Person, Visits, Observations, Conditions, Death, Procedure occurrences, Drug exposures, Measurements and more detailed information such as Drug ingredients, and Condition modifiers. OHDSI also provides open-source CDM applications for visualization and statistical analysis of patient-exposure-outcome cohorts. [10]

The InCor data integration challenges

The Heart Institute (InCor) of São Paulo, Brazil, is one of the six institutes of the Clinics Hospital complex, University of São Paulo Medical School. In the last two decades, InCor has increased its commitment in integrating all the relevant information of its patients, successfully developing an ePR named SI³. The first version of SI was deployed in year 2000; currently, it stores the clinical history, examinations, procedures, surgeries, notes, laboratory tests, medication, bills, and more for 1.3 million patients. Since then, the system has continuously evolved, overcoming several challenges related to the exchange of information among different healthcare institutions and remote installations. Furuie et al., describe details of the system architecture [8], while a number of studies involved cohort selection based on information collected by the SI³ system [3-5].

However, the workload involved in extracting the relevant patient information from SI³ has motivated the adoption of new strategies. Recently, we started the mapping from the SI³ data model to a standard data model that can simplify the observational retrospective studies related to Clinical Research. In this paper, we present the steps related to the mapping between SI³ and the CDM data models to prepare a new standardized database, named InCor-CDM, that can be used with the OHDSI toolset and a number of visual analytics tools. We measure the quality of the resulting InCor-CDM database using precision and recall statistics, when compared to the cohort generated by a previous study (gold standard).

Methods

Environment preparation

We prepared the InCor-CDM database environment by installing a PostgreSQL 10 DBMS, Java 10 JDK and Docker-compose in a Linux workstation. The database setup includes: (a) creating the required database users; (b) creating the OMOP CDM tables with the CommonDataModel/PostgreSQL scripts; and (c) importing the standard OMOP vocabularies from athena.ohdsi.org. All the OHDSI sources are available at github.org/OHDSI. Next, we installed the Achilles and Broadsea repositories required for the OHDSI web applications, configured the addresses and JDBC URLs and started their respective docker containers.

Database characterization

The InCor SI^3 database is stored in an Oracle 12c instance, accessed with the Oracle JDBC connector and DBeaver SQL client. The first step to manage the database was sampling its tables and columns with the OHDSI's WhiteRabbit software. This application generates a spreadsheet with the most frequent values of each selected column, so we can inspect them and make decisions about which columns to ignore, especially those with irrelevant or missing values. Then, we use RabbitInAHat to parse the output of WhiteRabbit and draw relational data-flow diagrams for documentation.

Some namespaces in SI^3 were ported from non-relational or older systems; such namespaces had no associated documentation, comments, constraints or foreign keys. For this reason, the database metadata had to be converted to a searchable JSON format with table and column names, types, comments, constraints and foreign keys. The column names were tokenized according to the naming scheme of the institution (e.g. abbreviations separated by underscore) and matched to similar columns in other tables; the goal was to find implicit relationships where the foreign keys were missing. Some attribute domains, such as internal record status codes, event sequence and timing diagrams, were documented from interviews with the support staff of InCor.

Patient de-identification

For de-identification purposes, personal information mapped to the CDM was limited to a minimum. Any key with a path to a patient primary key (and the PK itself) was **pseudonymized** [14], i.e., direct identification information such as citizenship document, phones numbers, addresses and names are not ported to the CDM, and the record primary key is exchanged to a *new id* (pseudonym), which is a random number drawn from an uniform distribution in the range 1×10^{10} and 9×10^{15} using Oracle's DBMS_RANDOM functions, addressing collisions with repeated sampling. This range was selected not to conflict with OMOP's standard concept ids (0 up to 2×10^9 are reserved) and to be within the limits of JSON numbers (53-bit precision). InCor holds the mapping from the new ids to the original keys in a private table; the mapping is to be used for notifying the patient, or her/his physician, in case the result of a study can improve a patient's condition.

Numerical variables were truncated in order to satisfy a baseline level of **k-anonymity** [15], i.e., guaranteeing that any patient variable value have at least *k* patients with the same information, so no patient is uniquely identifiable. For example, event dates were truncated to yearly, monthly, daily or hourly precision where original precision were not needed. Records with spurious attributes (e.g., dates in the future, outside any visit, invalid range, null required field, missing keys) were discarded. Also, we only loaded data from patients born before 2010 (aged 18+), with at least one valid visit.

si3.pac_paciente

paci_id	tp_sexo	...
01721	M	...
01722	M	...
01723	F	...

keys

table	src_id	new_id
pac_paciente	01721	7369111123
pac_paciente	01722	1257321234
pac_paciente	01723	3038618654

gender_map

src	id
F	8532
M	8507

omop.person

person_id	gender_concept_id	...
7369111123	8507	...
1257321234	8507	...
3038618654	8532	...

Figure 1– Example: gender mapping from SI^3 to the CDM.

Figure 1 shows sample data to illustrate this process. Each column from CDM person table (e.g., gender_concept_id) is extracted from the *source* table and column (e.g., pac_paciente. tp_sexo), and transformed with the appropriate domain map (e.g., value 'F' used in SI^3 for female gender is mapped to CDM concept 8532). The *keys* and *gender_map* tables are populated beforehand. The standard CDM concept ids were searched in OHDSI's Athena – a web-based CDM vocabulary explorer. Then, observe the following query, which loads data from the SI^3 PAC_PACIENTE table into the InCor-CDM omop.person table with remapped keys and concepts:

```
INSERT INTO omop.person
SELECT K.new_id          AS person_id,
       EXTRACT(YEAR FROM P.dt_nasc)
                         AS year_of_birth,
       COALESCE(G.id, 0) AS gender_concept_id
FROM si3.pac_paciente P
JOIN keys K ON  K.table=’pac_paciente’
            AND K.src_id=P.paci_id
LEFT JOIN gender_map G ON P.tp_sexo=G.src;
```

This query operates on the table samples in Figure 1, where it is assumed that the *keys* table holds the patient's random new_id, and that the *gender_map* table is a domain map table defined as (src char(1); id integer) corresponding to values of (tp_sexo, gender_concept_id). Unmapped values receive code 0, meaning "unknown concept" in the standard CDM vocabulary. Related tables were joined to the patient PK as usual, with their PK also remapped by keys.new_id.

Coding translation

InCor SI^3 uses the ICD 10 for diagnosis, a set of Brazilian vocabularies for coding clinical conditions, drugs, and procedures (TUSS is used by the Brazilian Health Care System for general terms and Brasindice for drugs – datasus.gov.br) and an internal coding system for generic billable items. Internal codes in use at InCor were inserted as new Concepts in the CDM (with ids mapped between 3×10^9 and 9×10^9) under the "InCor" vocabulary, with Concept Relationships to standard concepts whenever this information was available in SI^3. Initially, the records are inserted in the InCor-CDM with the original source codes, then the Concept Relationships from local to international codes are used to update the local InCor-CDM references to OMOP standardized terminologies, such as the SNOMED-CT, RxNorm, and LOINC [6], while the preserving the original source code stored in the patient record.

Quality assessment

After loading all the CDM tables, we execute the Achilles analysis, which will report data quality issues as errors, warnings or notifications. It will also preprocess the demographic characterization of the database for the visualizations and reports; after that, we start the Atlas WebAPI and front-end servers. Atlas offers both a RESTful API and a graphical web interface to schedule the execution of OHDSI methods, automatically generating the queries on a properly built and configured CDM database.

The most relevant errors initially found were related to: (a) the referential integrity, e.g., events without associated visit or events without a valid person_id, which were discarded; (b) lack of condition_eras, drug_eras and observation_periods, which were imputed from the ePR records by collapsing all the events of a patient not apart for more than a year into a single era – in the CDM standard, an “event” era refers to a time period of interest where “events” are recorded in the ePR; (c) events with invalid date, e.g. a condition_occurrence with start date in a future date; (d) too many patients without a diagnostic or prescription; InCor-CDM keeps such records because they can be used in the condition or drug_exposure dashboards, regardless of previous diagnosis.

Statistical evaluation

After initial corrections, we evaluate the quality of InCor-CDM by using software OHDSI’s Atlas to recreate a previous CVD patient cohort observed in a reference study that was executed over the InCor SI^3 database by Abrahao et al. [9]. That reference study prepared a clean de-identified database named Pauá, based on a 2016 snapshot of SI^3 records of patients, admissions, discharges, diagnoses, surgeries, PCI, medications, and laboratory tests. Within the Pauá database, Abrahao et al., verified the effect of statins on the survival rate of patients diagnosed with cardiovascular diseases. Our study will evaluate the InCor-CDM quality by defining a cohort with the same criteria and by verifying how many patients (with the same private InCor identifiers) were retrieved by each criteria given that reference study.

Results

In this section we evaluate InCor-CDM quality by selecting the same DCV cohort of a previous study in the Pauá database [9], by using Atlas over the InCor-CDM database. Our evaluation computes the Area Under the ROC Curve (AUC) based on the results of 12 executions of the DCV cohort at various settings. We compute predictive statistics for each cohort execution, using Pauá as a gold standard. The resulting AUC ranges from 0.5 (no different than random sampling) to 1.0 (reproduces exactly the same result as the gold standard).

Table 1– Databases cardinality (thousands of records).

Domain	SI^{3}2016	Pauá	SI^{3}2018	CDM
Person	1,116	323	1,346	946
Visit Occurrence	6,427	5,686	7,499	7,305
Condition Occur.	1,205	1,007	1,361	1,324
Procedure Occur.	45,024	144	53,945	51,479
Drug Exposure	83,283	2,775	100,052	38,962
Measurement	22,025	20,528	31,095	30,177
Death	17	21	18	18

In Table 1, we present the cardinality of the InCor-CDM in comparison to the SI^3 and Pauá database. Note that the Procedure and Drug domains in the Pauá database have substantially less records; this is because they are restricted to surgeries and particular classes of drugs, and because Pauá only uses patients with at least one admission and diagnosis. Additionally, InCor-CDM is based on a more recent snapshot, named SI^{3}2018 in the aforementioned table.

Cohort definition

To replicate the Pauá study, we created a cohort over the InCor-CDM database with the criteria below. Each list item directly corresponds to a HTML input field in the Atlas **cohort definition form**. These criteria are translated to SQL queries over our database in the CDM format (InCor-CDM).

- *Initial Event Cohort*: People having any of a **visit** occurrence of **Outpatient concept set** (with concept_id 9202, Outpatient Visit);
- *Additional Qualifying Inclusion Criteria*:
  - *Condition occurrence criteria*: with **at least 1** of **any condition**;
  - *Demographic criteria*: **age** greater than or equal **23** (in the censor window), matching the patients over 18 years old at the Pauá study start (1999);
  - *Demographic criteria*: with a **gender** of **MALE** (8507) or **FEMALE** (8532);
  - *Condition occurrence criteria*: of **CardioVascular Disease** (a concept set of concept ids from ICD-10 categories I20 to I25, I64 to I70 and G45, including descendants and mapped), with occurrence start between **2003-01-01** and **2013-12-31**, where event starts between **All** days before and **30** days after index end date (meaning the diagnosis was recorded around the time of the initial outpatient visit);
  - *Visit occurrence criteria*: with **at least 1** of **Outpatient concept set**, where event starts between **30** after and **All** days after index end date (a subsequent visit recording the outcome);
  - *Limit qualifying cohort to the* **earliest** *event*;
- *Era collapse gap size*: **1 day**;
- *Cohort censor window*: starting **2003-01-01** and ending **2013-12-31**.

After verifying the Concept Sets to have the correct concept ids, we generated the cohort on the InCor-CDM database and verified the cohort attrition report (the number of patients remaining in the selection after each filter) in Table 2:

Table 2– InCor-CDM CVD cohort attrition report in Atlas.

Criteria	n	%	Visualization
Initial	778,015	100.00	🟩
a) Dx	351,205	45.14	🟩
b) 18+	321,827	41.37	🟩
c) M/F	303,847	39.05	🟩
d) CVD	45,710	5.88	🟩
e) 2nd-V	39,910	5.13	🟩
f) People	39,498		

Table 2 summarizes the cohort definition criteria, where each abbreviation mean: (a) **Dx**: has a valid condition occurrence (diagnosis); (b) **18+**: over 18 years old; (c) **M/F**: Male or Female gender; (d) **CVD**: has any occurrence of ICD-10 I20 to I25, I64 to I70 or G45; (e) **2nd-V**: has a second outpatient visit occurrence more than 30 days after the index event (the initial visit); (f) **People**: lists how many people actually matched the events, because some patients have more than one episode. The baseline results achieved by the InCor-CDM CVD cohort indicate that 39,498 patients satisfy the criteria.

Table 3– Pauá reference CVD cohort attrition report.

Criteria	n	%	Visualization
Initial	313,894	100.00	
Dx	313,894	100.00	
18+	282,677	90.00	
M/F	263,339	83.87	
CVD	56,799	18.06	
2nd-V	27,698	8.80	
People	27,698		

In comparison, Table 3 displays an attrition report for the reference study (Pauá). The visualization was drawn with Python and GNU Gimp because Pauá's schema is not based on the CDM standard, and thus could not be used in Atlas. The reference cohort selected with the same criteria in the Pauá database has 27,698 patients (30% less than our first result), indicating the need of further refinement.

Initial Evaluation

All subject_ids in the cohort defined in Atlas were compared to the set of patient ids in the Pauá cohort. This comparison was executed by computing confusion matrices against the Pauá cohort result as a gold standard, then the derived scores: true and false positive ratios (TPR/FPR), positive and negative predictive power (PPV/NPV), accuracy (ACC) and F1-score. The confusion matrix the cohort #1 is given in Table 7, and derived scores make the first row of Table 8.

Table 7– Confusion matrix for cohort #1 (in Table 2).

Pauá InCor-CDM	P	N	Total
P	25,423	12,290	37,713
N	2,651	282,887	285,538
Total	28,074	295,177	323,251

Parameter refinement

Therefore, we noted cohort accuracy variations by creating additional cohorts with slightly adjusted parameters for the CVD criteria. Table 4 displays the attrition reports with increased periods for the qualifying criteria of CVD occurrence start date after index end date. Table 5 displays the attrition reports with increased periods for the qualifying criteria of Cardiopathy occurrence start date. We observe that increasing the collapse gap size to 7 and 14 days had no substantial effect on the results (Table 4).

Table 4– Varying condition start after index event (days).

Criteria \ days	7	14	21	30
Initial		778,015		
Dx, 18+, M/F		303,847		
CVD	44,967	45,255	45,484	45,710
People	39,055	39,203	39,342	39,498

Table 5– Varying condition start periods (years).

Criteria \ days	2003-2013	2000-2013	2000-2016
Initial		778,015	
Dx, 18+, M/F		303,847	
CVD	45,710	49,942	63,656
People	39,498	43,293	54,126

Table 6– Varying 2nd visit event start after index (days).

Criteria \ days	All	365	180	90
Initial		778,015		
Dx, 18+, M/F		303,847		
CVD	45,710	44,228	43,950	43,667
People	39,498	35,457	32,767	29,414

Then we evaluated all the patients selected in each cohort (from Tables 4, 5 and 6) by comparing them to the reference Pauá study, using the private Keys table to map the CDM person_ids to the SI³ patient ids (confidential to the institution), whose results are presented in Table 8. It should be noted that Pauá had an update in October 2016, after the reference study was published, and so we re-executed the query for its patient cohort, resulting in P=28,074 patients selected in the cohort (1.4% increase) and overall total P+N=323,251 patients included (3.0% increase).

Table 8– Predictive scores for each cohort.

#	TPR	FPR	PPV	NPV	ACC	F1
1	.905	.041	.674	**.990**	.953	.772
2	.901	.040	.678	**.990**	.954	.774
3	.903	.040	.677	**.990**	.954	.774
4	.904	.041	.676	**.990**	.954	.773
5	.905	.041	.674	**.990**	.953	.772
6	**.907**	.052	.623	**.990**	.944	.738
7	**.907**	.052	.622	**.990**	.944	.738
8	**.907**	.052	.620	**.990**	.943	.736
9	.889	.040	.680	**.990**	.954	.775
10	.877	.031	.727	.988	**.960**	**.795**
11	.829	.027	.743	.983	**.960**	.784
12	.754	**.023**	**.752**	.976	.957	.753

Evaluation of the ROC curve

We complete the evaluation by plotting the ROC curve from the predictive scores in Table 8. Only the patients existing in the Pauá database were used, i.e., only 37,713 patients of those retrieved in the cohort #1 also existed in the Pauá database. The highest scores of each column are highlighted in bold, e.g., cohort #10 exhibited the highest F1-score.

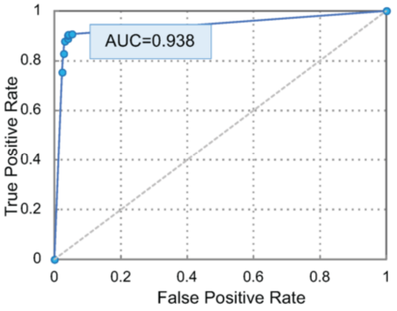

Figure 2– Empirical ROC curve for Table 8.

In this section we presented the results of the cohorts selected with Atlas from the InCor-CDM database. We also presented the effect of tweaking cohort parameters in the quality of the results (Table 8), which exhibited up to 80% F1-score, 75% Precision (PPV) and 91% Recall (TPR) at different settings. Figure 2 summarizes our results in an empirical ROC curve, exhibiting 0.938 of area under curve (AUC).

Discussion

The resulting InCor-CDM exhibits high agreement with the previous gold standard study [9]. This means that it is possible to estimate the same population-level effects (e.g. different medications) in both databases. InCor-CDM additionally benefits from the quality analyses implemented in the OHDSI Achilles tool, which warns about inconsistencies and errors found in the transformed data, and can be used for more advanced analysis in comparison with external CDM-based databases. For future work, we envision the comparison of patient cohorts between InCor-CDM and external CDM-based databases from other OHDSI work groups, further studying data quality, subpopulation characteristics between different institutes and evaluating risk scores for InCor patients.

Conclusions

We presented details of the migration process of a huge clinical database, the InCor's SI^3, to the OMOP CDM standard, an international format aimed at improving research in computer-aided medical systems. We presented the method used to extract, transform and load data between the databases commenting on the challenges regarding models, formats, terminology, and tools. We evaluated the quality of the resulting database, named InCor-CDM, by comparing cohorts obtained with the software OHDSI Atlas. We considered a previous cohort selection study used as ground truth; for a systematic comparison, we computed several information retrieval statistics and a ROC curve. The cohorts defined in Atlas exhibited from 62% to 75% precision, 75% to 91% of recall, 74% to 80% F1-score, and 0.938 of area under the ROC curve (AUC). The results indicate that the method that we employed was able to produce a new database that was both consistent with the original data and in accordance to the OMOP standard. The new database shall support a wide range of new research initiatives within the Heart Institute.

Acknowledgements

The authors thank FAPESP (grant #2018/11424-0), CNPq (grants #406550/2018-2 and #305580/2017-5) and InCor's Informatics Service: A. Longo, M. Rebelo, M. Yamaguti, R. Moreno, T. Vieira, V. Silva, V. Reis, W. Pereira and others.

References

[1] V.N. Gudivada, R. Baeza-Yates, and V.J. Raghavan, Big Data: Promises and Problems, *IEEE Computer* **48, 3** (2015), 20-23.

[2] A.K. Jha, D. Doolan, D. Grandt, T. Scott, and D.W. Bates, The use of health information technology in seven nations, *Int J Med Inform* **77, 12** (2008), 848-854.

[3] M.T. Abrahão, M.R.C. Nobre, and M.A. Gutierrez, Descriptive Statistics of 65000 Patients Treated for Myocardial Ischemia: data from routine electronic health records, *Evidence Live* **13** (2013).

[4] M.T. Abrahão, M.R.C. Nobre, and M.A. Gutierrez, The effectiveness of statins in the treatment of cardiovascular disease: Cross-sectional study with paired groups from electronic patient records, *Value Health* **16, 7** (2013), A518.

[5] M.T. Abrahão, T.J. Soares, F.A. Pires, M.A. Gutierrez, and M.R.C. Nobre, Data Warehouse com dados da Saúde Pública: Estudo de Caso sobre o Tratamento de Doença Cardiovascular Aterosclerótica no Estado de São Paulo, *CBIS* **12** (2010).

[6] G. Hripcsak, J.D. Duke N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *MedInfo* **2015** (2015), 574-578.

[7] D. Madigan, P.B. Ryan, M. Schuemie, P.E. Stang, J.M. Overhage, A.G. Hartzema, et al. Evaluating the Impact of Database Heterogeneity on Observational Study Results. *Am J Epidemiol* **178, 4** (2013), 645-651.

[8] S.S. Furuie, M.F.S. Rebelo, R.A. Moreno, M. Santos, N. Bertozzo, G.H.M.B. Motta, et al. Managing Medical Images and Clinical Information: InCor's Experience, *IEEE Trans Inf Technol Biomed* **11, 1** (2007), 17-24.

[9] M.T. Abrahão, M.R.C. Nobre, and M.A. Gutierrez, A method for cohort selection of cardiovascular disease records from an electronic health record system, *Int J Med Inform* **102** (2017), 138-149.

[10] D. Yoon, E.K. Ahn, M.Y. Park, S.Y. Cho, P. Ryan, M.J. Schuemie, et al. Conversion and Data Quality Assessment of Electronic Health Record Data at a Korean Tertiary Teaching Hospital to a Common Data Model for Distributed Network Research, *Healthc Inform Res* **22, 1** (2016), 54-58.

[11] E.A. Voss, R. Makadia, A. Matcho, Q. Ma, C. Knoll, M.J. Schuemie, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc* **22, 3** (2015), 553-564.

[12] R.H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F.M. Behlen, P.V. Biron, and A. Shabo, HL7 clinical document architecture, release 2. *J Am Med Inform Assoc* **13, 1** (2006), 30-39.

[13] D. Kalra, T. Beale, and S. Heard, The openEHR foundation, *Studies in Health Technol & Inform* **115** (2005), 153-173.

[14] K. Pommerening, and M. Reng, Secondary use of the EHR via pseudonymisation. *Studies in Health Technol & Inform* **103** (2004), 441-446.

[15] R.J. Bayardo, and R. Agrawal, Data privacy through optimal k-anonymization, *ICDE* **21** (2005), 217-228.

Address for correspondence

Daniel Mário de Lima <danielm@usp.br> (+55 11 2661 5545)