Designing Reliable Cohorts of Cardiac Patients across MIMIC and eICU

Catherine Chronaki[1,2], Abdullah Shahin[2], Roger Mark[2]

[1]HL7 Foundation, Brussels, Belgium
[2]Laboratory of Clinical Physiology, MIT, Cambridge, MA, USA

Abstract

The design of the patient cohort is an essential and fundamental part of any clinical patient study. Knowledge of the Electronic Health Records, underlying Database Management System, and the relevant clinical workflows are central to an effective cohort design. However, with technical, semantic, and organizational interoperability limitations, the database queries associated with a patient cohort may need to be reconfigured in every participating site. i2b2 and SHRINE advance the notion of patient cohorts as first class objects to be shared, aggregated, and recruited for research purposes across clinical sites.

This paper reports on initial efforts to assess the integration of Medical Information Mart for Intensive Care (MIMIC) and Philips eICU, two large-scale anonymized intensive care unit (ICU) databases, using standard terminologies, i.e. LOINC, ICD9-CM and SNOMED-CT. Focus of this work is lab and microbiology observations and key demographics for patients with a primary cardiovascular ICD9-CM diagnosis. Results and discussion reflecting on reference core terminology standards, offer insights on efforts to combine detailed intensive care data from multiple ICUs worldwide.

1. Introduction

Adoption of information technology to support clinical research has increased dramatically in the recent years. From 2005 to 2011, academic centers with data repositories for repurposing EHR data for research increased by 70% in the United States [1]. Notable was also the adoption of collaborative tools that enable teams to work together across organizations and time zones. Despite increasing use of information technology, differences in terminology and workflow combined with low or inconsistent adoption of standards limit the extent to which patient cohorts can be automatically assembled across clinical sites. The typical case is that the database queries need to be adjusted and reconfigured in each clinical site to address local coding systems and working practices that are reflected in the clinical data repository.

i2b2/SHRINE [2,3] proposed an architecture and a query language to advance the notion of patient cohorts as first class objects that can be shared, aggregated, and recruited for research purposes across clinical sites. The i2b2 workbench uses hierarchies to graphically compose patient cohort queries that are broadcasted to participating clinical sites, to receive the number of qualifying subjects and associated aggregated data. In this way, the time to identify a sufficient number of patients to analyse even complex clinical questions is significantly reduced. i2b2 data marts or data cells [4] along with ontology cells [5] carry a powerful notion of patient cohorts that extends across multiple sites through adaptors that facilitate mapping of concepts with support from terminology services and mapping tools. The i2b2 star schema paradigm centers on patient observations i.e. facts about the patient that are linked to specific data dimensions in an Entity-Value-Association. Observations are quantitative or factual data being queried, e.g. diagnosis, procedures, demographics, lab exams. Dimensions are groups of hierarchies and descriptors that define the facts e.g. concept, provider, visit, patient, or any other possible modifier. Actual coded concepts populate the ontology tables and facilitate mappings (see Figure 1).

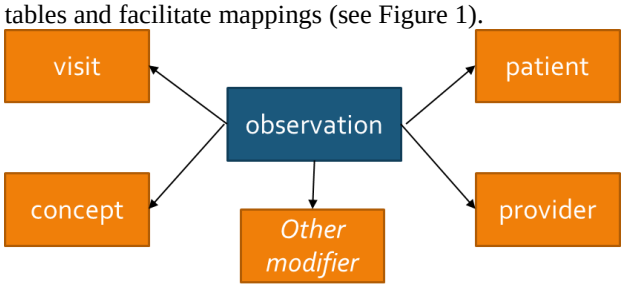

Figure 1: Main elements of the i2b2 star schema.

This paper reports on initial efforts to assess the coverage of concurrent queries to MIMIC and eICU, two large-scale anonymized ICU databases using the i2b2 design principles and standard terminologies.

Since 2003, MIMIC-II has served as a valuable resource to researchers worldwide offering detailed anonymized ICU data [6,7]. MIMIC-II v2.6 and MIMIC-III released in 2011 and 2015 respectively provide detailed data from ICU admissions in the Beth Israel Deaconess Intensive

Care Units from 2001 to 2011. MIMIC-III includes 57955 admissions of 48018 patients. All data, including clinical notes, have been de-identified and anonymized. In particular, all admission dates were shifted randomly, while time frames between clinical events remain intact.

The eICU Research Institute v3.0 data warehouse supports research initiatives on ICU patient outcomes, trends, and best practice protocols using data from the Philips eICU program currently operating in 35 states across USA. The eICU subset in this study (eICU_ADM version of March/April 2015) includes 731332 admissions in 500 ICU locations, mainly in 2011-12. De-identification preserves the year of admission, while the time of clinical events is presented as number of minutes/ seconds from admission. Clinical notes are not included.

The next generation of MIMIC aspires to be a *massive, detailed, high-resolution ICU data archive with complete medical records from patients admitted to intensive care worldwide.* Core facts such as lab observations, diagnosis, procedures, and medications may be coded with different level of detail in the data repositories of participating sites, thus presenting a formidable challenge to federated query processing and results aggregation.

The work reported in this paper looks into the terms or codes associated with demographics, diagnosis, and specific types of observations i.e. laboratory and microbiology in MIMIC and eICU and assesses the coverage of standard terminology systems and associated mappings. The evidence collected provides preliminary insights on the fitness of LOINC and SNOMED-CT (SCT) as reference core terminologies to support query and retrieval of patient cohorts with cardiovascular diagnosis.

2. Methods

The key issue shared with MIMIC-III and eICU is that they use alternative terms to describe the same concepts and moreover, the granularity or specificity of the terms is different. The i2b2 workbench uses hierarchies of coded concepts drawn from standard terminologies to compose patient cohorts using refined or general characteristics, and in this way, is able to address differences in the granularity of concepts used in specific clinical sites. Furthermore, i2b2 has developed transformation tools to benefit from the resources of the national center for biomedical ontology (http://www.bioontology.org) and has developed mapping tools to mitigate the co-existence of local and international coding and terminology standards in the participating sites. The i2b2 ontology and mapping tools ensure consistency and verification of the terminologies used in specific data repositories assisting with the assignment, verification, integration, export and import of the necessary mappings. To introduce MIMIC-III and eICU in an i2b2/SHRINE framework as presented in Figure 2, suitable terminology services and adapters need to be developed. They are needed to reformulate the query associated with a given patient cohort in MIMIC and eICU terms, and then transform any results received.

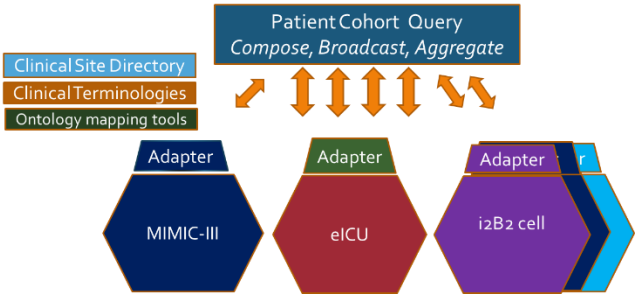

Figure 2: MIMIC-III and eICU integration components.

Table 1 presents the terminology standards used in i2b2, MIMIC and eICU and the reference terminology standards selected for evaluation in this paper. Since the scope of MIMIC is international, even though i2b2, MIMIC, and eICU use ICD9-CM, this work evaluates SCT as the reference terminology for diagnosis and microbiology observations. LOINC is adopted worldwide for lab observations and already most of MIMIC-III lab events are associated with a LOINC code.

Table 1. Reference terminologies for common elements.

Common element	i2b2/ SHRINE	MIMIC	eICU	Target
Gender	HL7 Admin Gender	Local	Local	SCT
Ethnicity	CDC	Local	Local	CDC
Labs	LOINC top 300	LOINC	Local	LOINC SCT
Microbio	LOINC (?)	-	Local	SCT
Diagnosis	ICD9-CM & hierarchy	ICD9-CM	ICD9-CM hierarchy	SCT

The coded values for the demographic traits *gender*, *age*, and *ethnicity* maintained in MIMIC and eICU differ. The value sets for gender adopted by HL7, LOINC, and SCT were the options considered with the objective to select a reference value set that would meaningfully express the *gender* of most eICU and MIMIC patients.

In MIMIC, the *age* of a patient at the time of admission can be computed by subtracting the date of birth from the date of admission. eICU stores the actual age of the patient in the *patients* table. HIPPA privacy regulations require to hide the exact age of patients 90 years or older at ICU admission. Therefore, when age is higher than 89 years, the string *">89"* is recorded in eICU. In MIMIC, if a patient's admission age is past the 89th year, the dates are shifted randomly so that age calculates at ~200 years. MIMIC-III has additional demographic data that are related to insurance, socioeconomic status and zip code. They were not considered, since most of them are not part of eICU.

In the case of laboratory observations, LOINC v2.22 of

December 22, 2014 was loaded in a database table and a mapping between lab types in MIMIC-III and eICU was first computed automatically and then reviewed manually adding appropriate LOINC codes where missing. Then, the mapping work was quality reviewed by two independent medical experts. Mapping followed the informal guidance of IHTSDO (SCT organization) step 2&3 (Figure 3).

Figure 3: IHTSDO guideline process for term mapping.

The current study selected patients with primary diagnosis in the cardiovascular/circulatory system (ICD9-CM 390.*-459.*). One-to-one maps were identified using the ICD9-CM to SCT map published by the National Library of Medicine, September 2014 edition. In MIMIC-III the *ICD9* table was used and the first ICD9-CM code with *sequence=1* is considered as the primary diagnosis. In eICU, ICU admissions are associated with coma-separated sequences of ICD9-CM codes that can be marked as "*active in discharge*". Each sequence is marked as "*primary*", "*major*" or "*other*" and is associated with a branch in the ICD9-CM disease hierarchy that classifies the patient's diagnosis. The % of ICU admissions in eICU and MIMIC that are covered by the *one-to-one* and *one-to-many* ICD9-CM to SCT maps give an indication of the coding style and type of mapping that is required. Microbiology observations are documented slightly differently in MIMIC-III and eICU, while *specimen (culturesite)*, *sensitivity (interpretation)*, *organism*, and *antibiotic* are part of a microbiology observation in both databases. The % of microbiology observations that can be expressed with pre-coordinated SCT terms, indicate the fitness of SCT in cross-ICU queries.

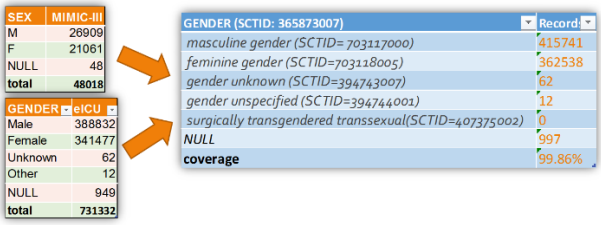

Figure 4: Gender expressed in SCT leads to 99% coverage.

3. Results

Demographics: HL7 administrative gender (https://www.hl7.org/fhir/valueset-administrative-gender.html) takes values in *('M', 'F', 'UN', null)* with *UN* standing for '*undifferentiated*'. LOINC adopts the WHO definition and accepts values for *sex* (http://r.details.loinc.org/LOINC/21840-4.html) in *('Male', 'Female', 'Other', 'Transsexual', 'Unknown'*). The SCT *findings* hierarchy includes the *gender* concept *(SCTID: 365873007)* with children: *'feminine gender', 'gender unknown', 'gender unspecified', 'masculine gender', 'surgically transgendered transsexual'*. Gender is expressed in MIMIC with the value set *('M', 'F', NULL)*, while eICU uses *('Male', 'Female', 'unknown', 'other', NULL).* Mapping both value sets to the SCT *Gender* concept results in coverage of 99.86% (see Figure 4). Specifying the age bracket of a patient cohort in MIMIC and eICU is trivial if the age computation in MIMIC for patients above 89 is adapted to yield ">89", the same as in eICU. Ethnicity in eICU takes values in *('African American', 'Asian', 'Other/unknown', 'Caucasian', 'Hispanic', 'Native American', NULL)*. Ethnicity in MIMIC is represented with 41 codes. Manually mapping MIMIC ethnicity codes to the eICU value set, covered 99.81% of the total ICU admissions. Organizing ethnicity codes in a hierarchy that expands to more detail as provided by MIMIC or the Center of Disease Control can support more refined ethnicity queries at the cost of smaller data sets.

Lab observations: MIMIC-III stores lab observations in the table *lab_events*. Lab observation types are identified by an internal code *itemid*. 718 lab observation types were identified of which 218 have an associated LOINC code. They account for 78.87% of the lab observation records in MIMIC. In eICU, lab observations are associated with an internal value set. There are 169 distinct lab observation types in 6 categories. Following a comprehensive mapping process, 103 lab types of eICU lab observations were associated to lab event types in MIMIC-III. When a LOINC code was not present in MIMIC, an appropriate one was identified. *As a result, 81.89% of eICU and 76.54% of MIMIC-II lab observations can be reached using LOINC codes.* Some eICU lab types are not considered lab events in MIMIC. For example, MIMIC considers *bedside glucose* as a *chart event* measured at bedside, while glucose is a *lab event* were a blood sample is taken to the lab for analysis. Such practice variations noted in the database structure, highlight the need to capture clinical context and workflow information.

Diagnosis: In eICU, admission diagnosis, diagnosis at discharge, as well as other diagnoses during the ICU stay are timestamped. MIMIC-III provides detailed de-identified admission and progress notes. Its *ICD9* records however, are not associated with the time and context when each of up to 33 ICD9-CM codes was documented.

In MIMIC-III, the ICU admissions that were associated with a first in sequence ICD9-CM code in [390.*-459.*] were selected. In eICU, admissions with a *primary* diagnosis sequence including a code in [390.*-459.*] and tagged "*cardiovascular/**" were selected. In eICU, 120 out of 1224 '*primary*' diagnostic sequences of ICD9-CM codes include a code in the range [390.*-459.*] and are associated with the "*cardio vascular/**" branch referring to 32643 admissions. In MIMIC-III 240 out of 2812 distinct ICD9-CM codes presented as first in sequence i.e. *primary* are in the range [390.*-459.*] and referred to 16112

admissions. MIMIC and eICU share 90 ICD9-CM codes within [390.*-459.*] showing significant variability in the selection of codes.

For 69/120 ICD9-CM codes in eICU, the NLM ICD9-CM to SCT *one to one* map offers a *single SCT* concept resulting *55.58% coverage of* admissions. In MIMIC-III, 108 of 240 ICD9-CM codes have *a one-to-one* map to an SCT concept capturing 54.7% of admissions. 48 of these 108 concepts are in the SCT *core set* and account for 83% of the covered ICU admissions of cardiovascular nature.

Using the NLM *one to many* map, 26/120 additional ICD9-CM codes from eICU identified with multiple SCT concepts including 6 codes that mapped each to a unique SCT concept and 7 codes mapped to *null*. Meanwhile, 73/240 additional disease codes from MIMIC-III were identified in the NLM map. 27 of these codes (2726 ICU admissions) mapped each to a unique SCT concept. For 16 of those, the concept was empty i.e. *null*. Many were the ones listed as '*other…*' the so called '*not otherwise specified'* which reflects the need for further analysis and synthesis of the related disease data. Overall, *the NLM SCT one-to-one and one-to-many maps covers 60.96% of the selected MIMIC-III admissions* of cardiac patients.

Microbiology observations: In eICU, microbiology observation records consist of: *sensitivity (interpretation)*, *organism*, *culturesite (specimen)*, and *antibiotic*. Sensitivity in MIMIC-III uses the value set ('*resistant'*, '*sensitive'*, *'intermediate', null)*. The corresponding value set in eICU for *Interpretation* is (*'R', 'S', 'I', 'P', null)*. Specimen of patients with a cardiovascular primary diagnosis were tested for one of 54 organisms in 1854741 tests, most frequently for *Escherichia coli*, *Staphylococcus aureus*, and *Klebsiella Pneumonia*. Organisms recorded in MIMIC-III are more granular: 309 organisms in 508696 tests of which 42 appear also in eICU, with the same three at the top in different order. The value sets of *specimen* (MIMIC-III) and *culturesite* (eICU) were mapped to SCT using the IHTSDO browser (browser.ihtsdotools.org). For 21 of 24 distinct *specimen* terms in eICU, suitable SCT concepts were found. This was not possible for "*Sputum, Tracheal Specimen*" and "*Sputum, Expectorated*" and "*Blood, Venipuncture*", cases where only post-coordinated expressions can be constructed. 93 specimen types in MIMIC-III were present in eICU, but in a more general form. For example, *blood culture*, a SCT concept used in eICU, can group several specimen types in MIMIC-III associated with SCT child-concepts of *blood culture*. Worth noting is also that some specimen types were in the SCT procedure hierarchy. 29 *antibiotics* were common among the 53 listed in eICU and the 30 listed in MIMIC for 99.86% and 79.66% of the recorded observations.

4. Discussion – Future Work

Assessing fitness of standard terminologies i.e. LOINC, ICD9-CM and SCT in queries across MIMIC-III and eICU is not easy. ICD9-CM has >13000 codes, SCT >350000 concepts, and LOINC >70000 terms. Value sets in use are typically 10% in size. MIMIC-III uses 218 LOINC terms, 103 shared with eICU. Only 90/240 ICD9-CM cardiac disease codes in MIMIC-III appear in eICU. Does this reflect different code practices or under-coding? Despite standardization efforts, mapping remains a very much needed complex tedious process for specialized expertise. Investing in ICU value sets and training could make the use of standard terminologies more effective. Hierarchies and ICU-specific value sets should help with varied query granularity, but that needs to be confirmed in clinical studies. Mapping common ICU diagnoses and value sets for microbiology events possibly with post-coordination, validated by scaling up prior studies to both ICU databases will no doubt elicit lessons and guidance that would advance the notion of patient cohorts as first class objects.

Acknowledgements

C. Chronaki would like to thank M. Feng, I. Silva, and L-W Lehmann for their help and support. This work is has been supported in part by EC Contract 64388 (AssessCT).

References

[1] Murphy S, Dubey A, Embi P, et al. Current State of Information Technologies for the Clinical Research Enterprise across Academic Medical Centers, CTS 2012;5:281-4

[2] Weber GM, Murphy SN, McMurry AJ, et al. The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. J Am Med Inform Assoc. 2009

[3] McMurry AJ, Murphy SN, Mac Fadden D, et al. SHRINE: enabling nationally scalable multisite disease studies. PLoS One. 2013;8(3):e55811.

[4] Philips, L. i2b2 Design Document, Ontology Management (ONT) Cell v.1.7.1, https://www.i2b2.org/software/files/PDF / current/Ontology_Architecture.pdf

[5] Donahoe J i2b2 clinical research chart design Document, v1.7 www.i2b2.org/software/files/PDF/current/CRC_Design.pdf

[6] Saeed M, Villarroel M, Reisner AT, et al. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database. Crit Care Med, 39(5):952–960, 2011.

[7] Scott DJ, Lee J, Silva I et al. Accessing the public MIMIC-II intensive care relational database for clinical research. BMC Medical Informatics and Decision Making 2013, 13:9

[8] Philips, eICU Research Institute v3.0 Detailed Design Document, V.000.DD.71, 2013-09-13, version 6.1.

Address for correspondence.

Catherine Chronaki
38-40 Square de Meeus, Brussels, 1000, Belgium
Email: euoffice@HL7.org