


Data and text mining

i2b2-etl: Python application for importing electronic health data into the informatics for integrating biology and the bedside platform

Kavishwar B. Waghlikar ^{1,2,*}, Layne Ainsworth³, David Zelle⁴, Kira Chaney⁴, Michael Mendis³, Jeffery Klann^{1,2}, Alexander J. Blood^{1,4}, Angela Miller³, Rupendra Chulyadyo³, Michael Oates³, William J. Gordon^{1,3,4}, Samuel J. Aronson³, Benjamin M. Scirica^{1,4} and Shawn N. Murphy^{1,2}

¹Harvard Medical School, Boston, MA 02115, USA, ²Massachusetts General Hospital, Boston, MA 02114, USA, ³Mass General Brigham, Boston, MA 02199, USA and ⁴Brigham and Women's Hospital, Boston, MA 02115, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on February 8, 2022; revised on July 15, 2022; editorial decision on August 15, 2022; accepted on August 31, 2022

Abstract

Motivation: The i2b2 platform is used at major academic health institutions and research consortia for querying for electronic health data. However, a major obstacle for wider utilization of the platform is the complexity of data loading that entails a steep curve of learning the platform's complex data schemas. To address this problem, we have developed the i2b2-etl package that simplifies the data loading process, which will facilitate wider deployment and utilization of the platform.

Results: We have implemented i2b2-etl as a Python application that imports ontology and patient data using simplified input file schemas and provides inbuilt record number de-identification and data validation. We describe a real-world deployment of i2b2-etl for a population-management initiative at MassGeneral Brigham.

Availability and implementation: i2b2-etl is a free, open-source application implemented in Python available under the Mozilla 2 license. The application can be downloaded as compiled docker images. A live demo is available at <https://i2b2clinical.org/demo-i2b2etl/> (username: demo, password: Etl@2021).

Contact: kwaghlikar@mgh.harvard.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

The i2b2 platform is deployed at major academic health institutions for querying for electronic health records (EHR) (Abend *et al.*, 2009; Murphy *et al.*, 2010). The platform provides a user-friendly interface that allows researchers with no expertise in information technology to find patient cohorts using EHR data. I2b2 has been deployed as a critical component of research networks including National Patient-Centered Clinical Research Network (PCORNet) (Klann *et al.*, 2018), Accrual for Clinical Trials (Visweswaran *et al.*, 2018) and Consortium for Clinical Characterization of COVID-19 (4CE) (Brat *et al.*, 2020; Weber *et al.*, 2009).

The platform has been used for a wide spectrum of use cases including clinical-trial enrollment (Bucalo *et al.*, 2021), population management (Waghlikar *et al.*, 2019), biobanking (Castro *et al.*, 2021; Mate *et al.*, 2017; Segagni *et al.*, 2011), clinical decision support and epidemiological analysis (Klann and Murphy, 2013;

Murchison *et al.*, 2021; Piffner *et al.*, 2016; Segagni *et al.*, 2011; Waghlikar *et al.*, 2017a, b). However, despite its impact and open-source availability, the deployment of the platform is largely limited to large academic medical centers.

A major obstacle for wider utilization of the i2b2 platform is the difficulty in data loading, as it requires the IT staff to learn the complex data-schema internal to the platform that has a star topology (Murphy *et al.*, 2007). To address this issue, we have developed the i2b2-etl package that specifies a simple input format and abstracts away the complex processes to initialize the internal i2b2 schema. This will facilitate wider deployment and utilization of the platform.

2 Materials and methods

We have implemented a Python application, referred to as 'i2b2-etl' that imports data in the Comma Separated Value (CSV) format into

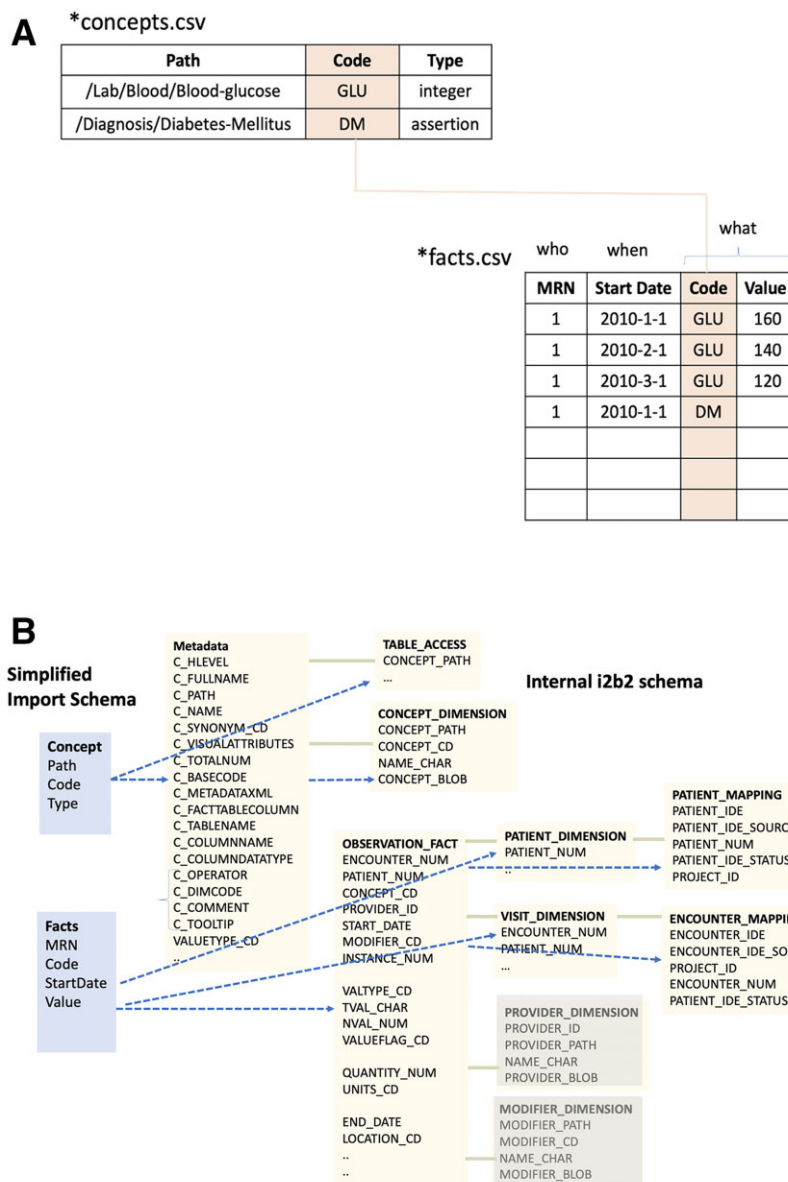


Fig. 1. (A) I2b2-etl accepts two types of CSV files—concept files and fact files. Concept files provide the meta-data or dictionary for the creation of an ontology hierarchy. The fact files contain patient data. For example, the first row represents that patient with medical-record number 1, had a ‘GLU’ of 160 on 1st January 2010. The concepts file clarifies that GLU is blood-glucose, which is a laboratory test performed on the blood, as indicated in the path, and it has a value of an integer. I2b2-etl uses the concept files to validate the facts while performing the import. (B) I2b2-etl parses in the input schemas shown on the left (blue background) and executes processes to populate the internal i2b2 tables shown on the right (yellow background). The metadata, table access and concept dimension tables are essential for the functionality to display ontologies and to query patient data using ontologies. These are automatically populated by the i2b2-tool. The observation-fact and dimension tables are internal i2b2 tables in a star topology that contain EHR data, and the mapping tables serve to de-identify the patient record numbers. These tables (except the provider and modifier dimensions) are automatically generated by i2b2-etl. Without i2b2-etl, all these internal tables need to be populated individually, which requires an in-depth understanding of the schema, a challenge that is now resolved by i2b2-etl (A color version of this figure appears in the online version of this article.)

the i2b2 platform. The source code is available in open source (<https://github.com/i2b2/i2b2-etl> and <https://github.com/i2b2/i2b2-etl-docker>), and as compiled containers in Dockerhub. The application can be downloaded as a docker images, that are compatible with all common Linux distributions, Windows and Mac-OSX systems.

For an online demo see, <https://i2b2clinical.org/demo-i2b2etl/> (username: demo, password: Etl@2021). After login navigates to ETL tab, press delete button and then choose upload files, selecting the CSV files from the [Supplementary Material](#). [Supplementary Appendix A](#) provides the steps to install i2b2-etl. [Supplementary Appendix B](#) describes command-line interaction. [Supplementary Appendix C](#) demonstrates the use of Gitlab to use SQL queries as inputs to i2b2-etl.

As shown in [Figure 1A](#), i2b2-etl accepts two types of CSV files—concept files and fact files, which contain the meta-data and data, respectively.

Concept files provide the meta-data or dictionary for the creation of the ontology hierarchy. Concept files consist of three columns: path, code and type, and their names end in ‘_concepts.csv’. Each row in the concept file corresponds to a node in the ontology hierarchy displayed in the i2b2-User Interface. The path specifies the unique location of the node in the hierarchy. The type can be integer, float, assertion, string or large-string, and the code is the abbreviated reference to the concept. For example, in the first row in the concept file shown in [Figure 1A](#), the path ‘/Lab/Blood/Blood-glucose’ leads to creation of the integer node called Blood-Glucose, as a child of the ‘Blood’ node. The ancestor nodes for ‘Lab’ and ‘Blood’ are automatically created.

The fact files contain patient data in four columns: medical record number (MRN), start-date, code and value. The name of fact files end in ‘_facts.csv’. Each row of the fact file provides the value of a specific observation (of a concept) for a patient referenced by the MRN starting at a particular point in time (start date). For example, the first row in the fact file shown in Figure 1A indicates that a value of 160 for blood glucose was observed on 1 January 2010, for the patient having a MRN 1.

I2b2-etl populates the tables in the internal i2b2 schema as elucidated in Figure 1B. The ontology hierarchy and SQL code snippets for ontology-based querying are auto-generated from the concept file and populated in the i2b2 metadata and concept dimension tables by the tool. The medical record numbers are converted into randomly generated integers and stored in the i2b2 patient-mapping table that is inaccessible to end-users. The latter can only access the randomized integers as patient numbers in the user interface. I2b2-etl performs validation of the input facts to ensure that each fact references a valid concept, has a value that conforms to its concept type, and that it has a valid time stamp.

We deployed i2b2-etl for a cardiovascular population health program at MassGeneral Brigham in Boston (Benson *et al.*, 2018; Blood *et al.*, 2020; Gordon *et al.*, 2021; Waghlikar *et al.*, 2019). The program involved daily interaction of navigators with patients. The resulting data were recorded in a relational database. To import this data into i2b2 for easy querying, we developed SQL queries to extract the project data into CSV files specified above. Next, we deployed i2b2-etl as a nightly job that executed the SQL queries and loaded the resulting CSV files into an i2b2 repository. The latter was setup specifically for the project using the i2b2 docker containers (Waghlikar *et al.*, 2018).

3 Results

Our i2b2-etl package allows importing of the ontology and patient-data as two simple CSV files. The cardiovascular program included data for 28 483 patients. Deployment of i2b2-etl resulted in 1395 concepts and over 4.7 million facts in the i2b2 repository, requiring 18 min for execution. The resultant i2b2-repository is used by the study staff to identify sub-cohorts in the population and to evaluate the program’s progress (Scirica *et al.*, 2021).

The novelty of i2b2-etl application is the simplified design of the input file schema, inbuilt de-identification and data validation. The input file schema abstracts away the complexity of the data schemas internal to the i2b2 platform (see Fig. 1B). This simplified mechanism will allow the IT staff at healthcare institutions to easily transform and load their institution’s EHR into the i2b2 platform.

Without i2b2-etl, the IT staff needs to thoroughly understand the complex schemas in i2b2 platform, in order to extract the data in conformance with the i2b2 schemas. As i2b2-etl can transform the simplified concept and file schemas into the platform’s schemas, IT staff is required to only focus on preparing SQL statements to yield the simplified input schemas. Moreover, with the integration of ETL module with Gitlab (Supplementary Appendix C), the entire ETL process can be automated, wherein Gitlab triggers the execution of SQL on the source database to extract the data as CSVs, which are then transformed into the i2b2 internal schemas and then loaded into the i2b2 platform’s database. Consequently, as the transform and loading steps are done by i2b2-etl, the IT staff only needs to focus on the extraction step, which can be performed by IT staff with minimal SQL expertise.

However, the simplified abstraction is at the expense of functionality in the i2b2 platform. I2b2-etl does not support querying using fields in the modifier, patient and visit dimensions of the i2b2 star topology. There are several alternative approaches that have been previously developed for importing EHR data into i2b2. Post *et al.* (2013a, b, 2016) have developed an application called Eureka that can load Excel files with custom schemas into the i2b2 database. Importing of EHR data into OMOP model has been used in several projects (Klann *et al.*, 2018, 2019; Majeed *et al.*, 2021; Rinner *et al.*, 2018). However, the major differentiator of i2b2-etl with the

alternative approaches is the simplified abstraction for the input, which allows users without advanced training in data modeling to rapidly import EHR into the i2b2 platform, thereby improving accessibility of the EHR for secondary use cases.

Funding

This work was supported by MassGeneral Brigham and National Institutes of Health [R00-LM011575, R01-HG009174 and R01HL151643].

Conflict of Interest: none declared.

References

- Abend, A. *et al.* (2009) Integrating clinical data into the i2b2 repository. *Summit Transl. Bioinform.*, 2009, 1–5.
- Benson, M.D. *et al.* (2018) A remote lipid management program improves appropriate statin use and cholesterol levels across a wide population of high cardiovascular risk patients. *J. Am. Coll. Cardiol.*, 71(11 Supplement), A1762.
- Blood, A.J. *et al.* (2020) Rationale and design of a navigator-driven remote optimization of guideline-directed medical therapy in patients with heart failure with reduced ejection fraction. *Clin. Cardiol.*, 43, 4–13.
- Brat, G.A. *et al.* (2020) International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *NPJ Digit. Med.*, 3, 109.
- Bucalo, M. *et al.* (2021) i2b2 to optimize patients enrollment. *Stud. Health Technol. Inform.*, 281, 506–507.
- Castro, V.M. *et al.* (2021) The mass general brigham biobank portal: an i2b2-based data repository linking disparate and high-dimensional patient data to support multimodal analytics. *J. Am. Med. Inform. Assoc.*, 29.
- Gordon, W.J. *et al.* (2021) Workflow automation for a virtual hypertension management program. *Appl. Clin. Inform.*, 12, 1041–1048.
- Klann, J.G. *et al.* (2019) Data model harmonization for the all of Us research program: transforming i2b2 data into the OMOP common data model. *PLoS One*, 14, e0212463.
- Klann, J.G. and Murphy, S.N. (2013) Supporting the health quality measures format in i2b2. *AMIA Jt. Summits Transl. Sci. Proc.*, 2013, 124.
- Klann, J.G. *et al.* (2018) Web services for data warehouses: OMOP and PCORnet on i2b2. *J. Am. Med. Inform. Assoc.*, 25, 1331–1338.
- Majeed, R.W. *et al.* (2021) Accessing OMOP common data model repositories with the i2b2 Webclient - algorithm for automatic query translation. *Stud. Health Technol. Inform.*, 278, 251–259.
- Mate, S. *et al.* (2017) On-the-fly query translation between i2b2 and samplify in the German Biobank Node (GBN) prototypes. *Stud. Health Technol. Inform.*, 243, 42–46.
- Murchison, C.F. *et al.* (2021) Racial differences in Alzheimer’s disease specialist encounters are associated with usage of molecular imaging and dementia medications: an Enterprise-Wide analysis using i2b2. *J. Alzheimers Dis.*, 79, 543–557.
- Murphy, S.N. *et al.* (2007) Architecture of the open-source clinical research chart from informatics for integrating biology and the bedside. *AMIA Annu. Symp. Proc.*, 2007, 548–552.
- Murphy, S.N. *et al.* (2010) Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Inform. Assoc.*, 17, 124–130.
- Pfiffner, P.B. *et al.* (2016) C3-PRO: connecting ResearchKit to the health system using i2b2 and FHIR. *PLoS One*, 11, e0152722.
- Post, A.R. *et al.* (2013a) Semantic ETL into i2b2 with eureka!. *AMIA Jt. Summits Transl. Sci. Proc.*, 2013, 203–207.
- Post, A.R. *et al.* (2013b) Temporal abstraction-based clinical phenotyping with eureka!. *AMIA Annu. Symp. Proc.*, 2013, 1160–1169.
- Post, A.R. *et al.* (2016) Metadata-driven clinical data loading into i2b2 for clinical and translational science institutes. *AMIA Jt. Summits Transl. Sci. Proc.*, 2016, 184–193.
- Rinner, C. *et al.* (2018) A clinical data warehouse based on OMOP and i2b2 for Austrian health claims data. *Stud. Health Technol. Inform.*, 248, 94–99.
- Scirica, B.M. *et al.* (2021) Digital care transformation: interim report from the first 5000 patients enrolled in a remote algorithm-based cardiovascular risk management program to improve lipid and hypertension control. *Circulation*, 143, 507–509.

- Segagni,D. *et al.* (2011) R engine cell: integrating R into the i2b2 software infrastructure. *J. Am. Med. Inform. Assoc.*, **18**, 314–317.
- Segagni,D. *et al.* (2011) The ONCO-I2b2 project: integrating biobank information and clinical data to support translational research in oncology. *Stud. Health Technol. Inform.*, **169**, 887–891.
- Visweswaran,S. *et al.* (2018) Accrual to clinical trials (ACT): a clinical and translational science award consortium network. *JAMIA Open*, **1**, 147–152.
- Wagholikar,K.B. *et al.* (2018) Implementation of informatics for integrating biology and the bedside (i2b2) platform as Docker containers. *BMC Med. Inform. Decis. Mak.*, **18**, 66.
- Wagholikar,K.B. *et al.* (2019) Phenotyping to facilitate accrual for a cardiovascular intervention. *J. Clin. Med. Res.*, **11**, 458–463.
- Wagholikar,K.B. *et al.* (2017a) Evolving research data sharing networks to clinical app sharing networks. *AMIA Jt. Summits Transl. Sci. Proc.*, **2017**, 302–307.
- Wagholikar,K.B. *et al.* (2017b) SMART-on-FHIR implemented over i2b2. *J. Am. Med. Inform. Assoc.*, **24**, 398–402.
- Weber,G.M. *et al.* (2009) The shared health research information network (SHRINE): a prototype federated query tool for clinical data repositories. *J. Am. Med. Inform. Assoc.*, **16**, 624–630.