1. Introduction

i2b2 and tranSMART were developed to provide clinical and translational investigators with the tools necessary to integrate medical record and clinical research data in the genomics age. The core of this is a highly flexible but simple i2b2 Common Data Model (CDM). The current version of the i2b2 CDM was released in July, 2020. This document not only describes the database tables and fields in the i2b2 CDM, but also provides a set of recommendations and best practices for using it.

The i2b2 CDM, which we initially developed in 2004, is based on a "star schema". Instead of separate tables for diagnoses, medications, and other data types, all patient observations are stored in a single "fact" table. A separate ontology describes the different codes that are placed in this fact table. As a result, institutions can use their own local codes, without having to map to common code sets. Furthermore, institutions can easily add new types of data to i2b2 and tranSMART just by extending the ontology. No changes to the database or software are needed. This enables software developers to build query, analysis, and visualization tools that are generalizable to different types of data and future-proof since the i2b2 CDM can remain stable over time.

Over 200 institutions worldwide use the i2b2 CDM to store and integrate coded electronic health record and medical claims data, notes, images, genomics, clinical trial data and more. It is highly scalable, with some instances containing billions of data facts for millions of patients and supporting queries from thousands of users. It is open source and has been implemented for Microsoft SQL Server, Oracle, and Postgres. The i2b2 CDM databases at different institutions can be linked and harmonized to form large federated data networks using the Shared Health Research Information Network (SHRINE) software. Because it is ontology based, the i2b2 CDM is well suited to address rapidly emerging public health crises, such as COVID-19, since codes for new tests and diagnoses only require updates to the ontology, not the database.

Other popular data models, such as OMOP CDM, have separate database tables for each data type. Although their schemas are more complex, they can be easier to learn for people who have not used ontology-based systems. (Nevertheless, the numerous benefits of the i2b2 CDM has led to its widespread adoption.) Aware of this, we have designed this documentation to start with the basics in a Quick Start guide to help new institutions gain familiarity with the i2b2 CDM (Chapter 2). Next, we provide a schema reference, introducing each of the tables and fields in the i2b2 CDM (Chapter 3). Then, a series of brief tutorials explain more complex ways of using the i2b2 CDM to model different data types (Chapter 4). Finally, we discuss advanced topics, such as security and performance optimization (Chapter 5).