Enhancing the Representational Power of i2b2 through Referent Tracking

Jonathan C. Blaisure, MSc[1,2], Werner M. Ceusters, MD[1],[2]
[1]Institute for Healthcare Informatics, University at Buffalo, Buffalo, New York, USA;
[2]Department of Biomedical Informatics, University at Buffalo, Buffalo, New York, USA

Abstract

The Informatics for Integrating Biology and the Bedside (i2b2) software platform has proven successful in leveraging clinical enterprise data for the identification of cohorts of patients satisfying certain demographic, phenotypic and genetic criteria in support of further studies. An unanswered question thus far is whether i2b2 search criteria could include characteristics of assertions themselves, e.g. diagnoses, rather than what the assertions (observations) are about, e.g. diseases. This would allow, for instance, to find cohorts of patients for which different providers have been in disagreement about what condition the patient is suffering from. Previous research has shown that this requires more explicit detail about, and unique identification of, two sorts of entities: those that directly or indirectly contribute to the coming into existence of such observations and those that are either explicitly mentioned or merely implied in the assertions. Our research here demonstrates that i2b2's modifier system can be used to represent the relationships between observations and their explicit or implied referents on the one hand, and between relevant referents themselves on the other hand, both in combination with the storage of explicit unique instance identifiers for these observations and referents in i2b2's fact table. While this approach adheres to i2b2's base functionality and implementation specifications, it makes explicit ambiguities and confusions that would otherwise remain undetected.

Introduction

The Informatics for Integrating Biology and the Bedside (i2b2) software platform is designed to identify on the basis of enterprise data, in particular electronic healthcare records (EHR), cohorts of patients with a specific phenotype, and this in support of further studies[1]. Since there exist many biases and idiosyncrasies in EHR data that hamper their secondary use[2], we have been exploring the extent to which applying the principles of Ontological Realism[3] and Referent Tracking (RT)[4] to Extract-Transfer-Load (ETL) procedures can improve data quality in secondary use repositories built on common data models (CDM). The strategy involves: (1) to identify, and where possible remove, ambiguities and (2) to represent explicitly what is *implied* in certain assertions in the source EHRs and the CDMs[5-7].

Diagnoses are specifically prone to biases and inconsistencies[2], and this was also exemplified in our data warehouse[8]. From an ontological perspective, a correct diagnosis is minimally about the configuration formed by four components: (1) the patient, (2) the condition in the patient referenced by the diagnosis, (3) the type that the condition instantiates and (4) the relationships in which they stand[9]. When a diagnosis is inaccurate with respect to one or more of these components – most common is an erroneous assignment of the type – it fails at the level of compound expression but is still about the other components at the level of individual reference[10]. In mainstream EHR systems and common data model approaches, however, it is unfortunately common practice to represent diagnoses explicitly as being about the patient and the *type* of condition, but not explicitly about the condition which is diagnosed *itself*. Thus when a diagnosis of 'benign colon polyp' is followed by a diagnosis of 'malignant colon polyp', it is hard to infer whether it was the 1st, originally benign, polyp that turned malignant or whether the 2nd diagnosis is about another polyp than the 1st diagnosis was about[4]. Hogan has demonstrated that adequate representation of all entities relevant to a diagnosis is possible in a RT system by exploiting its ability for making explicit reference[11] not only to the four components as first-order entities, but also to the assertions about them as second-order entities, either in isolation or in combination[12].

Does the current i2b2 data model has similar representational capabilities? If so, that would allow querying the system not only for combinations of genetic, phenotypic or demographic characteristics of patients, but also for characteristics of assertions themselves. Examples of such queries with respect to diagnosis would be: which patients have seen some diagnosis be revised, or had different providers disagreeing about what condition they were suffering from. A more complex question would be: how many patients have at least one disease about which different providers have been in disagreement for more than N years, but later reconciled their opinions. Disagreements in diagnosis range from 5% to 75% depending on what the diagnosis is about (anatomopathology, psychiatric disease, radiographic imaging, …) while up to 30% of such disagreements involve some sort of diagnostic error[13]. Being able to differentiate such patient cohorts is thus important, either to eliminate such cases from secondary use data repositories, to identify them as such so that caution can be used when making inferences from such repositories, or to make them the topic of further inquiry towards quality improvement or policy making.

Background

The i2b2 platform is composed of a backend comprising several ‘cells’ that implement a number of web services and use an Extensible Markup Language (XML) based messaging syntax to communicate amongst each other as well as with client applications[1]. Each cell encapsulates its own business logic as well as access to data objects behind standard Web interfaces. Examples of such cells are the Clinical Research Chart (CRC) and the ontology cell (Ont Cell). The CRC functions as the integrated data repository for i2b2 and is organized as a star schema warehouse. The star schema consists of a ‘fact table’ surrounded by ‘dimensional tables’. Dimensional tables hold the sort of data about entities referenced in the fact table independent of the facts themselves. In i2b2’s fact table (**Table 1**), each row – or certain collection of rows – represents (roughly) an observation made by a provider about a patient in the context of some encounter. This table links to dimensional tables holding more information about first-order entities such as providers, patients, and encounters, as well as to a Concept_Dimension and Modifier_Table. Metadata for the latter tables is further described in the Ontology Cell which defines hierarchies as well as synonyms and modifiers. The notion of modifiers is used to provide more details about observations, examples being ‘route’ and ‘dose’ in relation to the concept of ‘medication’, and ‘initial’ and ‘discharge’ in relation to the concept of ‘diagnosis’.

Table 1. Fields in the structure of the Observation_Fact table relevant to our project (adapted)[14].

Key	Column Name	Column Definition
PK	Encounter_Num	Encoded i2b2 patient visit number
PK	Patient_Num	Encoded i2b2 patient number
PK	Concept_Cd	Code for the observation of interest (i.e. diagnoses, procedures, medications, lab tests)
PK	Provider_Id	Practitioner or provider id
PK	Start_Date	Starting date-time of the observation
PK	Modifier_Cd	Code for modifier of interest.
PK	Instance_Num	Encoded instance number that allows more than one modifier to be provided for each Concept_Cd. Each row will have a different Modifier_Cd but a similar Instance_Num.
	Valtype_Cd	Format of the concept. N = Numeric; T = Text.
	Tval_Char	Used in conjunction with Valtype_Cd = T or N. When Valtype_Cd = T: stores the text value. When Valtype_Cd = N contains one of E (Equals), NE (Not equal), L (Less than), LE (Less than and Equal to), G (Greater than), GE (Greater than and Equal to).
	Nval_Num	Used in conjunction with Valtype_Cd = N to store a numerical value

With the standard i2b2 web client, clinical researchers can retrieve patient counts that satisfy certain criteria expressed as Boolean queries on phenotypic observations or genetic traits. The architecture of i2b2 allows software developers to expand the range of questions that can be asked without changing the i2b2 database model. This can be achieved, for example, by using the i2b2 web client (1) on an i2b2 instance created with more complex ontologies and Extract-Transfer-Load (ETL) procedures[15] or (2) in combination with other query tools[16], or (3) by implementing other user interfaces[17]. However, a limitation of i2b2, so it is argued, is that for some use cases, e.g. finding undiagnosed patients with rare complex disorders by comparing their phenotype with that of diagnosed patients, a formal and precise list of phenotypic and/or demographic criteria might not be available and a search on similarity metrics more convenient[18]. The question is whether queries about disagreements in diagnoses as we have in mind belong also to this category.

Ontological Realism aims the development of high-quality ontologies that faithfully represent what is general in reality with the further goal to use these ontologies to make heterogeneous data collections comparable[3]. Central here is the idea that the world should be conceived as including entities of two sorts: ‘particulars’ (or ‘instances’) and ‘types’. Particulars are the sorts of entities described through observations performed for example in the lab or clinic. Types are to be understood as the counterparts in reality of the general terms used in the formulation of scientific theories. For each given type, there are many particulars that are its instances and the existence of certain similarities amongst certain particulars within some domain, for example healthcare, allows types to be depicted as nodes in a graph, each node standing in certain relations to other nodes. Types do not only exist for first-order entities such as patients, bacteria and lab tests, but also for second-order entities such as ‘assertions’, ‘facts’ or ‘observations’ about patients, i.e. the sort of entities EHRs and data repositories based on common data models are composed of.

However, whenever an assertion is made about some patient, it is almost never *exclusively* about that patient, except under some very strict interpretation of aboutness[10]: some might indeed hold that when they assert that 'this patient has three children' they do not assert that these children have that patient as parent, although they will not deny that an inference to that effect can be made by anybody who has adequate knowledge about the 'laws of nature' that govern parenthood and childhood. Under a less strict interpretation of aboutness, the statement 'this patient is 6.3 feet tall' can be argued to be not only about the patient but also about another entity which is his length, this length being 6.3 feet. Similarly, the presence of the ICD-code 'E11.21 - Type 2 diabetes mellitus with diabetic nephropathy' in a patient's chart is not just about the patient, but also about a second entity which is the disease that inheres in this patient and which has been diagnosed as being diabetes of the stated type, as well as about a third and a fourth entity: each of the two kidneys of that patient, at least if he still has two kidneys! Assertions of the sort mentioned above presuppose the existence of even more entities than those of which the assertion is directly and indirectly about. An ICD-code in a diagnosis field of an EHR presupposes the existence of an interpretative process which resulted in the diagnosis, this process in turn presupposes the existence of somebody who performed the interpretation on the basis of further entities such as lab tests, clinical examinations, and so forth[19]. The presence of '6.3' in a patient's record's field labeled 'length in feet' presupposes that the patient's length was measured, this act of measuring the length of that specific patient being an entity in its own right which in turn presupposes the existence of a measurement instrument.

In summary, each 'observation' found in an EHR is the final stage of a registration process – i.e., the process which resulted in the data elements being part of the record – which followed an observation process during which bodily features on the side of the patient became associated with one or other form of representation. Each of these processes have participants – healthcare workers carrying out or assisting in the process, measurement instruments, devices, and so forth – and each of these participants contributes in one or other form to the accuracy and reliability of the data.

Referent Tracking[4] (RT) has been developed as a framework to be maximally explicit about particulars[11] and this in accordance with the principles of Ontological Realism as implemented in the Basic Formal Ontology (BFO)[20] which has been shown to be the ontology whose terms are most often-reused in the BioPortal collection of biomedical ontologies[21]. Central in RT is the requirement for explicit unique identification by means of Instance Unique Identifiers (IUI) for all entities on the side of the patient and his environment that are involved in the observation and registration processes. In EHRs, only particulars of a few types are uniquely identified in an explicit way such as patients, certain healthcare workers such as providers and usually also patient-provider contacts. What is further uniquely identified, be it implicitly, are combinations of data elements as rows in the EHR tables defined by a unique primary key column or a combination of columns composing the primary key. These system-internal identity assignments are useful to keep track of data provenance, their authorized use, and also system-internal quality control. But only rarely are they used to register 'observations' about these data elements themselves, an exception being diagnoses which a provider in some systems can annotate with a confidence level or an indication that the observation was entered in error.

The lack of unique identification for all salient entities related to assertions leads to an important level of data reduction. For example, configurations in which at time t_n the relationship R obtains between entity e_1 of type C_x and entity e_2 of type C_y, become reduced to assertions to the effect that the relationship R obtains between *some* entity of type C_x and *some* entity of type C_y. A further reduction occurs when EHR data are pooled into repositories such as i2b2 in which explicit unique identification is restricted to patients, providers, and visits, and implicit unique identification to 'observations'.

In this paper, we explore ways in which some of the principles of Referent Tracking and Ontological Realism can be used to determine which entities that directly or indirectly contribute to the coming into existence of 'observations' in i2b2's Observation_Fact *should* be represented to provide more *explicit* detail about the precise background and context of these 'observations' so that it would become possible to query an i2b2 server for cohorts of patients not only on the basis of combinations or temporal sequences of what is believed to be facts about the patients, but also including belief revisions that have occurred over time, specifically about diagnoses. We also explore the extent to which representations of this sort *can* be accommodated for in i2b2's current database schema thereby applying the principle of 'faithful representation of reality' not only to what is happening in the patient's body (1st order reality), but also to that what is stated about the patient (2nd order reality).

Methodology

Hogan provided a complete analysis of how the simple case of '*a single patient with a single disease, as diagnosed on a single occasion by a single physician*' is to be represented in a Referent Tracking system[12]. This analysis was carried out using the abstract tuple syntax of Referent Tracking[22] with the goal to have an exhaustive representation of all first- and second-order entities that must exist – including the relationships that obtain between them – for the assertions in the EHR to be an accurate representation of the intended portion of reality. Our effort took the form of a real-life case study extracted from our EHR system and with a more complicated scenario consisting of six encounters with the same patient[8], but this time also including the fact that over time different diagnoses concerning the patient's conditions have been recorded by two different providers, including different diagnoses by the same provider. Our goal here is not to be exhaustive, but rather to assess the extent to which information of a sort expressed by means of Referent Tracking tuples can also be represented in the current i2b2 schema to be able to answer the sort of questions mentioned earlier. The series of encounters is represented in **Table 2**.

Table 2. Scenario selected for our case study derived from a case history reported on earlier[8].

	Encounter ID	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**
	DateTime	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**
Diagnosis Type	**Provider ID**	**P1**	**P2**	**P1**	**P2**	**P2**	**P1**
DT1: Type 1 Diabetes Mellitus - Uncontrolled		Ins	Act	Res		Res	
DT2: Type II diabetes mellitus with ketoacidosis				Ins	Act		Act
DT3: Type 2 Diabetes Mellitus - Uncomplicated, Uncontrolled				Ins	Err		
DT4: Acanthosis nigricans				Ins	Act	Act	Act

Legend. An entry in the cells of the columns E1 … E6 indicates the creation of a new data element in the EHR of the patient in this scenario at the time indicated by 'DateTime'. What sort of data element was created, is indicated by the cell entry: 'Ins' – a diagnosis of the type specified in the 'Diagnosis Type' column was entered; 'Act' – it was confirmed that the diagnosis of the specified diagnosis type is still 'active' during the encounter; 'Res' – the disease that lead to the diagnosis of the specified diagnosis type was specified to be resolved; 'Err' – the diagnosis made earlier was erroneous.

Our first step was to identify in the database of the EHR system from which our i2b2 server is populated the transactions that are generated for the patient history as described in **Table 2**, and the data elements that resulted from these transactions. Next, each data element – or combinations thereof such as the combination of a 1st data element containing a drug name, a 2nd one a numeral and a 3rd one a measurement unit – that qualifies as an assertion was subjected to an analysis with the goal to identify all particulars that are explicitly or implicitly referenced in the assertion. We used to that end an earlier developed expansion algorithm that identifies all particulars, as well as the relationships amongst them and the types they instantiate, that exist or must have existed for the assertion (1) to have come into existence itself, and (2) to be a faithful representation of reality[23]. From this list we then eliminated the particulars that are not required to be represented into i2b2 to be able to answer the sorts of questions we mentioned earlier. To the relevant particulars, we assigned an IUI. We studied the i2b2-documentation as well as notes and discussions in i2b2 community and development platforms in an attempt (1) to fully grasp what the i2b2 authors intended the different tables and fields to be filled with and to identify (2) the best fields to store the IUIs, and (3) the types that need to be represented in the Ontology cell so that the questions can be answered through the standard i2b2 web client and without changing the i2b2 data model insofar possible. We also studied EHR to i2b2 conversion experiences reported on in the scientific literature specifically to compare our interpretations of the documentation with those of other groups.

Results

Table 3 lists the entities from our scenario of which the IUIs should be represented in i2b2's CRC, restricted to those relevant for our intended queries.

Storing the IUIs for patients was found to be straightforward and several alternatives exist. One solution would be, since i2b2 allows optional columns to be added to the Patient_Dimension table, to create such a column, for example with the column name 'IUI' and data type varchar(n) with n large enough to hold the sorts of IUIs used in the local Referent Tracking system and to add a row in the Code_Lookup table with the required entries for the fields Table_CD, Column_CD, Code_Cd and Name_Char set to 'Patient_Dimension', 'IUI', Crc_Column_Descriptor' and 'Instance Unique Identifier' respectively. An alternative might be to store IUIs in the encrypted Patient_Ide field of the Patient_Mapping table and setting the Patient_Ide_Source field of that table to the IUI of the Referent Tracking system from which the patient IUI is drawn.

Table 3. Relevant entities from the analyzed scenario to be tracked in a RT-compatible i2b2 repository.

IUI	Description	Type	Valid time
IUI-1	The patient.	BFO:Material Entity	t1
IUI-2	The disease in IUI-1 which provider IUI-P1 diagnosed as being of type DT2 during encounter IUI-E3.	OGMS:Disease	t2
IUI-3	The disease course comprised of all pathological processes that are realizations of disease IUI-2.	OGMS:Disease Course	t3
IUI-4	The acanthosis nigricans in the skin of patient IUI-1.	OGMS:Disorder	t4
IUI-5	Part of the disease course of patient IUI-1 which provider IUI-P1 diagnosed during encounter IUI-E1 as being of type DT1 and which he declared to be resolved during encounter IUI-E3.	OGMS:Process	t5
IUI-P1	The provider identified by means of Provider ID 'P1' in the EHR used in our scenario.	BFO:Material Entity	t-p1
IUI-P2	The provider identified by means of Provider ID 'P2' in the EHR used in our scenario.	BFO:Material Entity	t-p2
IUI-En *	The encounters identified resp. by means of Encounter ID 'En' – 'n' ranging from 1 to 6 – in the EHR used in our scenario.	BFO:Process	t-en
IUI-DT|E|P|C *	The data elements in the EHR as represented in **Table 2**. For each of the 13 data elements, the respective IUI is formed by replacing 'DT|', 'E|' and 'P|' with the numeral of resp. the corresponding diagnosis type, the encounter ID and the provider ID, and 'C' with the cell entry. For example, the IUI for the data element represented in the right bottom cell of **Table 2** is 'IUI-462Act'.	IAO:Representation	t- DT|E|P|C
IUI-6	The composite data element formed by the data element IUI-231Ins and IUI-431Ins.	IAO:Representation	t6

Legend. Types (3rd column) are preceded by the abbreviation of the ontology they are drawn from (BFO=Basic Formal Ontology[20]; OGMS=Ontology for General Medical Science[19], IAO=Information Artifact Ontology[10]). 'Valid time' (4rd column) indicates (a) for continuants: the BFO:Temporal Region during which the BFO:instance relation holds between the entity denoted by the IUI and the indicated Type, and (b) for processes: the BFO:Temporal Region in which they are located. '*' in the IUI column indicates that an IUI-template is specified rather than an individual IUI.

Similar alternatives can be used for encounter IUIs: either by adding an optional column to the Visit_Dimension table or by storing the IUIs in the encrypted Encounter_Ide field of the Encounter_Mapping table and setting the Encounter_Ide_Source field of that table to the IUI of the RT system from which the encounter IUI is drawn. The two alternative solutions for patient IUIs and encounter IUIs do not work for provider IUIs: the Provider_Dimension table is not documented as allowing optional columns, and there is no table for mapping providers[14]. A solution here might be to include the provider IUI at the end of the path specified in the Provider_Path field of the Provider_Dimension, or, if the IUI is shorter than 50 characters, in the Provider_ID field.

More complex is storing the relevant information (IUIs, types of which the referents of these IUIs are instances of, relationships between these referents, and so forth) for the observations and for the disease-related entities (i.e. disease, disease course, and disorder)[19] on the side of the patient in the Observation_Fact table (**Table 1**). Our solution makes extensive use of modifier codes (**Table 4**) that could be drawn from relevant ontologies, in particular updated versions of the Information Artifact Ontology[10] and the Ontology for General Medical Science[19]. For readability, we divided the resulting part of the Observation_Fact table in four separate tables. **Table 5** lists the entries that could be created for the observations made with respect to diagnosis DT2 during the encounters E3, E4 and E6 (**Table 2**). It assumes that the patient does indeed have a disease condition, and that it is that condition that both providers are making, resp. confirming, a diagnosis about. **Table 6** shows a similar solution created for the observations made with respect to diagnosis DT1 during the encounters E1, E2 and E3. **Table 7** displays a solution for dealing with erroneous observations while, finally, **Table 8** shows a solution to represent diagnostic disagreement.

Table 4. Proposed modifier codes for relating observations to what they are about.

Modifier_Cd	Description
Compl:cCause	indicates that the IUI in the Tval_char field denotes the observation referencing the cause of the complication.
Compl:cEffect	indicates that the IUI in the Tval_char field denotes the observation referencing the effect of the complication.
Compl:Reference	indicates that the IUI in the Tval_char field denotes the data element that references the complication.
Compl:tCause	indicates that the IUI in the Tval_char field denotes the entity which is the cause of the complication.
Compl:tEffect	indicates that the IUI in the Tval_char field denotes the entity which is the effect of the complication.
Dx:cDisagreement	indicates that the IUI in the Tval_char field denotes an observation for which there is disagreement between providers.
Dx:cError	indicates that the IUI in the Tval_char field denotes the observation which states a previous observation to be erroneous.
Dx:cInitial	indicates that the diagnosis is an initial diagnosis.
Dx:Reference	indicates that the IUI in the Tval_char field denotes the observation (thus a reference) in the source system.
Dx:Referent	indicates that the IUI in the Tval_char field denotes the patient's condition, i.e. the referent which the observation is about.
Dx:tActive	modifier code indicating that the observation states the referent condition still to be active.
Dx:tComplOf	indicates that the IUI in the Tval_char field denotes the entity of which the referent of the diagnosis is a complication.
Dx:tEnd	indicates that the IUI in the Nval_char field denotes a reference to the end date of the patient's condition under the assumption that the referenced entity exists.
Dx:tOnset	indicates that the IUI in the Nval_char field denotes a reference to the start date of the referent of the diagnosis, thus to the patient's condition under the assumption it exists.
Dx:tRealOf	Short for the referent's 'realization of'[19], indicating that the IUI in the Tval_char field denotes the disease of which the referent of the diagnosis is the disease course.
Dx:tResolved	indicates that the referent of the diagnosis was resolved, i.e. ceased to exist.
Dx:tType	modifier code indicating that the value in the Tval_char field denotes a class label from an ontology.

Discussion

It is clear from these results that it is possible to set up an i2b2 instance that is compatible with the principles of Referent Tracking and Ontological Realism. The proposed method guarantees that the standard i2b2 web client can be used to identify cohorts of patients based on criteria not only including references to first-order entities (demographics, phenotypes, genetics) but also including data elements through which these first-order entities are referenced, as well as first-order entities implied to exist for such data elements to make sense. Although it would have been possible – most likely – to use other tables or specific fields that are allowed in the current data model, we preferred to work primarily through the modifier system in order to avoid ambiguities, confusions and conflations that we believe to exist in this data model and the available documentation.

One issue, for example, that needs more clarification is what an '*observation'* in i2b2's terminology precisely means. It is stated that an 'observation' is '*simply a recording or a notation of something*', thereby providing the following clarification: '*For example, the observation of 'diabetes' recorded in the database as a 'fact' at a particular time does not mean that the condition of diabetes began exactly at that time, only that a diagnosis was recorded at that time (there may be many diagnoses of diabetes for this patient over time)*'[14]. When taken at face value, I2b2 takes here the position that in their terminology an 'observation' is not something in first-order reality on the side of the patient – thus neither that what is observed, nor the process of observing that something – but rather a second-order entity, more concretely a data element – or a collection of data elements – resulting from observing something followed by an act of registering.

Table 5. Storing IUIs of observations – with respect to diagnosis DT2 during the encounters E3, E4 and E6 – and what they are about in relevant fields of the Observation_Fact table.

	Encounter_num	**Patient_num**	**Concept_cd**	**Provider_ID**	**Start_date**	**Modifier_cd**	**Instance_num**	**ValType_Cd**	**Tval_char**	**Nval_num**
R1	E3	1	DT2	P1	D3	@	1	<null>	<null>	<null>
R2	E3	1	DT2	P1	D3	Dx:Reference	1	T	IUI-231Ins	<null>
R3	E3	1	DT2	P1	D3	Dx:cInitial	1	<null>	<null>	<null>
R4	E3	1	DT2	P1	D3	Dx:Referent	1	T	IUI-3	<null>
R5	E3	1	DT2	P1	D3	Dx:tOnset	1	N	L	n(D3)
R6	E3	1	DT2	P1	D3	Dx:tEnd	1	N	G	n(D3)
R7	E3	1	DT2	P1	D3	Dx:tRealOf	1	T	IUI-2	<null>
R8	E3	1	DT2	P1	D3	Dx:tType	1	T	OGMS:Disease Course	<null>
R9	E4	1	DT2	P2	D4	@	1	<null>	<null>	<null>
R10	E4	1	DT2	P2	D4	Dx:Reference	1	T	IUI-242Act	<null>
R11	E4	1	DT2	P2	D4	Dx:tActive	1	<null>	<null>	<null>
R12	E4	1	DT2	P2	D4	Dx:Referent	1	T	IUI-3	<null>
R13	E4	1	DT2	P2	D4	Dx:tEnd	1	N	G	n(D4)
R14	E6	1	DT2	P1	D6	@	1	<null>	<null>	<null>
R15	E6	1	DT2	P1	D6	Dx:Reference	1	T	IUI-261Act	<null>
R16	E6	1	DT2	P1	D6	Dx:tActive	1	<null>	<null>	<null>
R17	E6	1	DT2	P1	D6	Dx:Referent	1	T	IUI-3	<null>
R18	E6	1	DT2	P1	D6	Dx:tEnd	1	N	G	n(D6)

Legend. 'n(Dm)' in Nval_num column: indicates a conversion of date 'Dm' to a numerical value. '@' is the value used by i2b2 for fields in which a <null> value is not allowed, but otherwise no meaningful value can be given.

Table 6. Storing IUIs with respect to diagnosis DT1 during the encounters E1, E2 and E3.

	Encounter_num	**Patient_num**	**Concept_Cd**	**Provider_ID**	**Start_date**	**Modifier_cd**	**Instance_num**	**ValType_Cd**	**Tval_char**	**Nval_num**
R19	E1	1	DT1	P1	D1	@	1	<null>	<null>	<null>
R20	E1	1	DT1	P1	D1	Dx:Reference	1	T	IUI-111Ins	<null>
R21	E1	1	DT1	P1	D1	Dx:Initial	1	<null>	<null>	<null>
R22	E1	1	DT1	P1	D1	Dx:Referent	1	T	IUI-5	<null>
R23	E1	1	DT1	P1	D1	Dx:tOnset	1	N	L	n(D1)
R24	E1	1	DT1	P1	D1	Dx:tEnd	1	N	G	n(D1)
R25	E1	1	DT1	P1	D1	Dx:tType	1	T	OGMS:DisCourse	<null>
R26	E2	1	DT1	P2	D2	@	1	<null>	<null>	<null>
R27	E2	1	DT1	P2	D2	Dx:Reference	1	T	IUI-122Act	<null>
R28	E2	1	DT1	P2	D2	Dx:Active	1	<null>	<null>	<null>
R29	E2	1	DT1	P2	D2	Dx:Referent	1	T	IUI-5	<null>
R30	E2	1	DT1	P2	D2	Dx:tEnd	1	N	G	n(D2)
R31	E3	1	DT1	P1	D3	@	1	<null>	<null>	<null>
R32	E3	1	DT1	P1	D3	Dx:Reference	1	T	IUI-231Res	<null>
R33	E3	1	DT1	P1	D3	Dx:tResolved	1	<null>	<null>	<null>
R34	E3	1	DT1	P1	D3	Dx:Referent	1	T	IUI-5	<null>
R35	E3	1	DT1	P1	D3	Dx:tEnd	1	N	LE	n(D3)

Table 7. Storing IUIs of observations with respect to diagnosis DT3 during the encounters E3 and E4.

	Encounter_num	Patient_num	Concept_cd	Provider_ID	Start_date	Modifier_cd	Instance_num	ValType_Cd	Tval_char	Nval_num
R36	E3	1	DT3	P1	D3	@	1	<null>	<null>	<null>
R37	E3	1	DT3	P1	D3	Dx:Reference	1	T	IUI-331Ins	<null>
R38	E3	1	DT3	P1	D3	Dx:Initial	1	<null>	<null>	<null>
R39	E3	1	DT3	P1	D3	Dx:Referent	1	T	IUI-3	<null>
R40	E4	1	DT3	P2	D4	@	1	<null>	<null>	<null>
R41	E4	1	DT3	P2	D4	Dx: Reference	1	T	IUI-342Err	<null>
R42	E4	1	DT3	P2	D4	Dx:cError	1	T	IUI-331Ins	<null>
R43	E4	1	DT3	P2	D4	Dx:cDisagreement	1	T	IUI-331Ins	<null>

Table 8. Relating diagnosis DT2 to DT4 during encounter E3 (see **Table 5** for initial entries to DT2).

	Encounter_num	Patient_num	Concept_cd	Provider_ID	Start_date	Modifier_cd	Instance_num	ValType_Cd	Tval_char
R44	E3	1	Complication	P1	D3	@	1	<null>	<null>
R45	E3	1	Complication	P1	D3	Compl:Reference	1	T	IUI-6
R46	E3	1	Complication	P1	D3	Compl:tCause	1	T	IUI-2
R47	E3	1	Complication	P1	D3	Compl:tEffect	1	T	IUI-4
R48	E3	1	Complication	P1	D3	Compl:cCause	1	T	IUI-231Ins
R49	E3	1	Complication	P1	D3	Compl:cEffect	1	T	IUI-431Ins
R50	E3	1	DT4	P1	D3	@	1	<null>	<null>
R51	E3	1	DT4	P1	D3	Dx:Reference	1	T	IUI-431Ins
R52	E3	1	DT4	P1	D3	Dx:cInitial	1	<null>	<null>
R53	E3	1	DT4	P1	D3	Dx:Referent	1	T	IUI-4
R54	E3	1	DT4	P1	D3	Dx:tType	1	T	OGMS:Disorder
R55	E3	1	DT4	P1	D3	Dx:tComplOf	1	T	IUI-2

From this description alone it is unclear whether ‘*database’* in the clarification cited denotes only the Observation_Fact table in the i2b2 system or, perhaps in addition, the database of the source system in which this diagnosis was recorded.

The presence of five different date fields in the observation_fact table provides arguments for the thesis that only the data element originally registered in the source database is referred to as the ‘observation’ These date fields are start_date (starting date-time of the observation), end_date (end date-time for the observation), update_date (date obtained from the source system at which the row was updated by the source system), download_date (date the data was downloaded from the source system) and import_date (date the data was imported into the observation_fact table). In this sense, the observation_fact table does not contain observations, but rather data elements that are created in the observation_fact table through a process of downloading (at the download_date) copies of the observations from the source system, followed by an upload process (at the import_date) into the i2b2 system. For clarity, we will henceforward use the term ‘source observation’ to denote original data elements in the source system and ‘observation representation’ to denote those elements in the observation_fact table that correspond with the source observation.

If we are correct in this assumption, then another issue is the precise meaning of ‘start_date’ and ‘end_date’ in the Observation_Fact table. Would the start_date of a source observation be the date that the data element is created in the EHR? If so, to what does then the end_date correspond to? The date it is deleted? That would be odd since most EHR systems are deletionless. Might start- and end_date roughly identify the temporal period during which the provider referenced in the Observation_fact table considers the source observation to be faithful to reality? That might then lead to another confusion at the level of the source observation when these dates do not originate from direct entry by the provider in the EHR at the time the source observation, e.g. a diagnostic assertion, is created, but are obtained through subsequent EHR entries, thus *distinct* source observations over time. This confusion can be avoided by committing to an ontology that accepts source observations to change and grow over time, similar to how database tables may grow over time. The solution we adopted here is to perceive each new data element in the EHR as a new source observation, and accepting that source observations are not only about first-order entities, but can also be about other source observations. We therefore use only the start_date field in which we enter the datetime the source observation was created, while we use the modifiers Dx:tEnd and Dx:tOnset to reference the condition on the side of the patient under the assumption that it exists (e.g. R5 and R6 in **Table 5**).

Assuming that the referenced entity exist, does not force us to assume that the source observation is veridical. Thus it might very well be that the patient has some condition, but that the diagnosis about that condition is erroneous. By conceiving source observations as instances of IAO:Representation this can perfectly be represented since the requirement for a representation is that it is *intended* to be about something, but not that it is veridical[10]. When an instance of IAO:Representation is truly about what it is intended to be about, then it is also an instance of IAO:Information Quality Entity (IQE) and therefor a *concretization* of an instance of IAO:Information Content Entity (ICE). As an example, both R4 in **Table 5** and R39 in **Table 7** have IUI-3 as referent of the respective diagnoses, yet, both cannot be true at the same time. It is in R42 that the erroneous diagnosis is identified and in R43 that it is declared to be a disagreement.

From this it follows that a determination of what a diagnosis in an EHR is precisely about in combination with whether or not the diagnosis is accurate, is crucial for determining what entities exist, and what they are precisely instances of. That is the reason for our choice to let diagnoses DT2 and DT3 to be about the patient's disease course (IUI-3), rather than about the disease (IUI-2) which is realized in the disease course (R7 in **Table 5**). Indeed, that a condition is 'under control' or 'uncontrolled' is not a characteristic of the disease itself, but rather how the disease evolves without or despite appropriate treatment. It is for this reason that we do not use the Concept_Cd field as a reference to the type the referent of the source observation is an instance of, but rather the modifier Dx:tType in combination with an entry in Tval_char as for example in R25 in **Table 6**. This approach also allowed us to be neutral as to whether provider P1 was correct in declaring the patient first to have had diabetes type I which then was resolved (R33 in **Table 6**). Indeed, since till to date DM type 1 is assumed not to be curable, there most likely never was in this patient another disease entity of type DM than IUI-2, but it was this entity that was misdiagnosed as DM1, and then, 2nd mistake, declared as resolved, rather than having been entered in error. By taking the referent of the initial diagnosis, i.e. IUI-5, to be a part of a disease course, it leaves both possibilities open: if P1 was right, then IUI-5 is the disease course related to the first disease (another one than the one referenced by IUI-2). If he was wrong, then IUI-5 would be a part of IUI-3. This ambiguity could be expressed by a marker to that end in the field Confidence_Num (not shown here) in the Observation_Fact table to express our assessment of accuracy of R33.

One limitation of our approach is that applying the method correctly requires good insight in the principles of ontological realism. Unfortunately they are not commonly taught in medical informatics and data science although the need thereto was already recognized in 1978[24]. A limitation of this paper, though not of the proposed approach, is that only one specific case is discussed and therefor that it does not demonstrate that the approach is generalizable. Future work includes a thorough analysis and description of all scenario types encountered, the implementation thereof in our institution's i2b2-server and an analysis of whether the approach is generalizable to other platforms.

Conclusion

By making extensive use of i2b2's modifier system we have been able to extend the range of queries that can be issued through the standard i2b2 web client so as to include not only criteria based on first-order entities such as demographics and phenotypic configurations, but also criteria about the assertions in which such first-order entities are referenced. At the heart of the solution is (1) making explicit distinction between data elements and what they are about, and (2) unique identification of all entities referenced directly, or implied to exist. While this approach adheres to i2b2's base functionality and implementation specifications, it avoids ambiguities and confusions that would otherwise remain undetected.

Acknowledgement: This work was supported in part by Clinical and Translational Science Award NIH 1 UL1 TR001412-01 from the National Institutes of Health.

References

1. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124-30.
2. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30-7.
3. Smith B, Ceusters W. Ontological realism: A methodology for coordinated evolution of scientific ontologies. Appl Ontol. 2010;5(3-4):139-88.
4. Ceusters W, Smith B. Strategies for referent tracking in electronic health records. J Biomed Inform. 2006;39(3):362-78.
5. Blaisure J, Ceusters W. Business Rules to Improve Secondary Data Use of Electronic Healthcare Systems Informatics for Health 2017; 24-26 April 2017.; Manchester, United Kingdom.2017. p. 5.
6. Ceusters W, Blaisure J. A Realism-Based View on Counts in OMOP's Common Data Model. 14th International Conference on Wearable, Micro- and Nanotechnologies for Personalized Health; 14-16 May; Eindhoven, The Netherlands 2017. p. 7.
7. Blaisure J, Ceusters W. Improving the 'Fitness for Purpose' of Common Data Models through Realism Based Ontology. AMIA 2017 Annual Symposium November 7, Washington, DC: AMIA; 2017.
8. Bona JP, Ceusters W. Replacing EHR structured data with explicit representations. International Conference on Biomedical Ontology; July 27-30; Lisbon, Portugal. 2015. p. 85-6.
9. Hogan WR, Ceusters W. Diagnosis, misdiagnosis, lucky guess, hearsay, and more: an ontological analysis. J Biomed Semantics. 2016;7(1):54.
10. Smith B, Ceusters W. Aboutness: Towards Foundations for the Information Artifact Ontology. International Conference on Biomedical Ontology; July 27-30; Lisbon, Portugal2015. p. 47-51.
11. Ceusters W, Chiun Yu Hsu, Smith B. Clinical Data Wrangling using Ontological Realism and Referent Tracking. CEUR Workshop Proceedings. 2014;1237:27-32.
12. Hogan W. Representing the aboutness of a diagnosis. Medical Informatics Europe; April 24-26, 2018; Goteborg, Sweden.2018.
13. Institute of Medicine. Improving Diagnosis in Health Care. Balogh EP, Miller BT, Ball JR, editors. Washington, DC: The National Academies Press; 2015. 472 p.
14. Donahue J. Data Repository (CRC) Cell. 2016 09/06/2016. Report No.: 1.7.08-004.
15. Post AR, Krc T, Rathod H, Agravat S, Mansour M, Torian W, et al. Semantic ETL into i2b2 with Eureka! AMIA Jt Summits Transl Sci Proc. 2013;2013:203-7.
16. Murphy SN, Avillach P, Bellazzi R, Phillips L, Gabetta M, Eran A, et al. Combining clinical and genomics queries using i2b2 - Three methods. PLoS One. 2017;12(4):e0172187.
17. Wang TD, Plaisant C, Quinn AJ, Stanchak R, Murphy S, Shneiderman B. Aligning temporal data by sentinel events: discovering patterns in electronic health records. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Florence, Italy. 1357129: ACM; 2008. p. 457-66.
18. Garcelon N, Neuraz A, Benoit V, Salomon R, Kracker S, Suarez F, et al. Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse. J of Biomedical Informatics. 2017;73(C):51-61.
19. Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. Summit on translational bioinformatics. 2009;2009:116-20.
20. Arp R, Smith B, Spear AD. Building ontologies with Basic Formal Ontology. Cambridge, Massachusetts: Massachusetts Institute of Technology; 2015. xxiv, 220 pages.
21. Ochs C, Perl Y, Geller J, Arabandi S, Tudorache T, Musen MA. An empirical analysis of ontology reuse in BioPortal. J Biomed Inform. 2017;71:165-77.
22. Ceusters W, Manzoor S. How to track Absolutely Everything? In: Obrst L, Janssen T, Ceusters W, editors. Ontologies and Semantic Technologies for the Intelligence Community Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press; 2010. p. 13-36.
23. Rudnicki R, Ceusters W, Manzoor S, Smith B. What particulars are referred to in Electronic Health Record data? A case study in integrating Referent Tracking into an EHR application. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2007:630-4.
24. Kent W. Data and reality : basic assumptions in data processing reconsidered. Amsterdam: North-Holland; 1978.