# A Proposed Star Schema and Extraction Process to Enhance the Collection of Contextual & Semantic Information for Clinical Research Data Warehouses

Michael Blechner[T]
Department of Pathology, UCHC, Farmington, USA
mblechner@uchc.edu

Rishi Kanth Saripalle[T] and Steven A Demurjian
Dept. of CSE, UCONN, Storrs, USA
rishikanth@engr.uconn.edu, steve@engr.uconn.edu

**Abstract** - *In the past decade, clinical patient data has played a pivotal role in clinical and translational research in support of new treatment options, medical interventions, drug development, etc. In support of this process, researchers require massive integrated data sets generated via a health information exchange (HIE) to centralize and automate the development and maintenance of a clinical research data warehouse (CRDW). The data harvested from the CRDW is obtained by cleansing transactional clinical databases (TCD) used for daily clinical activities. Traditionally, TCD schema and CRDW data models only capture conceptual patient data, often neglecting to address the contextual and semantic information attached to such data that is crucial for clinical analysis. In this paper, we propose a star schema and associated extraction process to enhance the collection of contextual and semantic information in support of CRDW that leverages HL7 Clinical Document Architecture in conjunction with the Reference Information Model.*

***Keywords- health information exchange, clinical translational research, clinical data warehouses, heatlh care standards.***

## I. INTRODUCTION

For the past decade, the federal government and various medical organizations have been pushing for the adoption of information technology (IT) to digitalize patient information in the form of an Electronic Health Record (EHR). This directional change was due to the introduction of the HITECH Act of 2009[1] and the associated meaningful use guidelines. The HITECH Act provides financial incentives for clinical organizations to implement EHRs and share electronic patient data with other organizations using the health information exchange (HIE). The primary goals are to: improve patient quality care by reducing medical errors due to rapid access to digitalized patient's data and current best practices; efficiently share and transfer patient medical data via HIE to reduce costs; and promote evidence-based medicine [2] that utilizes clinical data for clinical and translational research (CTR) to accelerate the transition of effective medical practices from the bench to the bedside and from the bedside into the community.

In support of CTR, many research institutions have built clinical research data warehouses (CRDW) which are analytical relational database systems built on a dimensional database schema, as illustrated in Figure 1. These systems capture information from multiple transactional clinical databases (TCDs) on a daily basis and are built via a scheduled process of extract-transform-load (ETL) operations. The transformed clinical data is placed into a CRDW using fact and dimension tables that are structured into star or snowflake schemas [3]. One popular approach to CRDW is Informatics for Integrating Biology and the Bedside (i2b2) [4, 19], a framework that harnesses medical data aggregated from multiple TCDs and varied organizations for research purposes which is of high value to CTR. While i2b2 has proven successful for large organizations, a significant percentage of clinical data is gathered by small outpatient clinics and community hospitals that do not have the financial resources to build or contribute to a CRDW. As a result of an incomplete representation of a population's clinical data which can limit the CRDW 's utility may lead to biased conclusions for hypotheses.

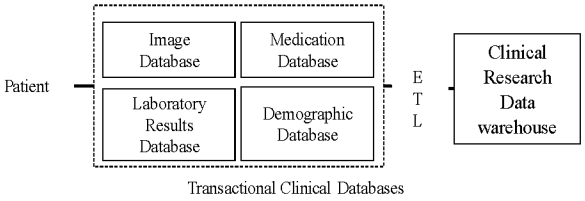


**Figure 1:** Example illustrating CRDW and various TCD's.

i2b2 also suffers from an inability to capture vital semantic and contextual information related to patient data which are very crucial for interpreting that data during scientific discoveries. Additionally, HIE is almost exclusively document-based, i.e., single discrete data elements are not exchanged but rather entire documents are shared that contains crucial data. HIE relies heavily on the HL7 Clinical Document Architecture (CDA) [5, 6, 7] to represent documents that can contain structured and semantically encoded data such as physical exam findings, clinical lab

[T] The author's research is supported by a grant from the CT Institute for Clinical and Translational Science at University of Connecticut Health Center.

results, patient diagnoses, therapeutic interventions, etc.

In this paper, our objective is to leverage EHR generated HL7 CDA documents [6, 7] in conjunction with regional HIE systems to inexpensively and automatically populate a regional CRDW to produce a data warehouse containing a more complete data set that is more representative of a population's clinical care and maintains the contextual relationships between the data. Towards this objective, this paper proposes a star schema and associated extraction process that takes as input EHR CDA documents to automatically populate a CRDW. Our work also employs existing technologies and standards (e.g., SNOMED [8], LOINC [9], MeSH [10]).

The rest of the paper has four sections. Section 2 provides background on HIE, an HL7 CDA example, and a discussion on limits of CRDWs. Section 3 introduces our proposed star schema and extraction process to enhance the collection of contextual and semantic information for CRDWs. Section 4 details related work on CRDWs. Finally, Section 5 concludes the paper and discusses ongoing work.

## II. Background and Motivation

***Health Information Exchange:*** HIE provides the capability to electronically move clinical information between disparate health care information systems while maintaining the semantic meaning of the information. This is accomplished in part by using CDA documents which provide a mechanism for semantically encoding clinical information. The electronic exchange of clinical information is still not widely available in the United States and the required technology is still evolving. However, the financial incentives and the meaningful use guidelines of the HITECH Act have generated a great deal of progress. The Integrating the Healthcare Enterprise (IHE) [11] specifications have become widely accepted standards for implementing healthcare data exchange. An HIE system compliant with these specifications enables clinical documents generated by an EHR or other clinical system at one institution to be registered in a centralized database and be later retrieved by a clinician at another institution. An IHE compliant HIE as three main components: the Patient Identity Cross-Reference (PIX) server to manage patient identification across different institutions; the document registry server that stores metadata about existing documents; and document repositories which store the actual clinical documents to be exchanged. Frequently, repositories belong to individual institutions and are located behind institutional firewalls enabling each organization to maintain full control of their clinical data. The IHE specifications provide the mechanisms for the HIE to retrieve documents from these private repositories. Alternatively, an HIE can have a central repository to hold documents for institutions that do not have the internal infrastructure or support for their own repository. The IHE specifications provide the ability to query the repositories for a patient's documents based on his assigned identifier in the PIX server. While HIE systems facilitate exchange of clinical data for patient care including the reporting of required data to state public health agencies, at present, they have not been developed or implemented to support CTR.

***CDA Standard and Example:*** The electronic exchange of clinical information between EHR systems will continue to increase due to the financial incentives of the HITECH Act. Clinical data to be exchanged between systems is first wrapped within a CDA document, an XML-based standard for the electronic exchange of clinical documents. A CDA document is composed of a header, which identifies and classifies the document, and a body, which is the core of the document holding the patient's clinical data.

In the CDA Release 1, the header section contains well-structured content, but the body of the document is typically unstructured text as shown in Figure 2. The header section has various components such as: *patientRole* to capture the patient information and demographics; *author* to capture the person or device that represents the entity which created the document; *legalAuthenticator* to specify the clinician who has legally signed the document and attested to its content; etc. The header can also contain details of the specific clinical encounter, related consents given by the patient, related documents, and the confidentiality assigned to the document as a whole.

However, in CDA Release 1, the body is wrapped around a *nonXMLBody* tag indicating unstructured information. The unstructured representation and lack of semantics in the body of CDA R1 means that these documents are of limited or no value for CTR. This limitation is addressed in the CDA R2 standard [3]. The CDA R2 model is highly expressive and allows clinical data to be represented in the form of statements. The header implementation in CDA R2 is identical to that of R1, but the body is structured into sections, whose content can be encoded using standard medical vocabularies. As shown in Figure 3, the CDA body is wrapped by using a *structuredBody* tag and is built recursively into nested sections. Each section can contain a single narrative block and any number of CDA entries and external references.

***Limitations of CRDWs:*** Imagine a scenario where a state or regional HIE supports a clinical document

exchange for the region's outpatient practices using a number of relatively small community hospitals and an academic university institution.

```
<ClinicalDocument>
<!– CDA Header-->
    <template/>
      <recordTarget>
        <patientRole>
         <addr></addr>
         <patient> </patient>
           <patientRole>
             <author>
               <addr></addr>
               <assignedPerson> </ assignedPerson >
             </author>
            <patientRole>
       <assignedPerson> </ assignedPerson
      </recordTarget>
<!– CDA Body-->
<component>
      <nonXMLBody>
        <text mediaType="application/rtf" ></text>
      </nonXMLBody>
</component>
</ClinicalDocument>
```

**Figure 2:** CDA Header and Unstructured Body.

Conceptually, HIE provides the technical infrastructure to aggregate larger clinical data across the region. But, this data is not readily searchable by a clinical researcher at the academic hospital. The university hospital may have a CRDW stocked with data from their TCDs. However, many of the patients seen at the university hospital are also seen by other physicians outside the university. The clinical data from these encounters will not be represented in the universities TCDs or CRDW. In addition, if a researcher at the university wants to gather data regarding the effects of a specific therapy on the treatment of colon cancer s/he would be limited to the data available from patients cared for at the university hospital; clinical encounters outside of the university hospital will not be represented and this fact may bias any analysis performed on the data in an effort to represent the entire region's population. Feeding data from smaller institutions into a common CRDW or having each institution build their own CRDW is not realistic financially.

Additionally, clinical data is often embedded in a context that may be crucial to interpreting the data. For example, the administration of a drug can elicit an adverse response in a patient which may also be manifest one or more clinical findings or conditions. These findings can have additional qualifiers such as severity. One researcher may initially be interested in querying for adverse drug reactions but later may need to know the manifestations of these reactions. Another researcher may be interested in patients who experience anaphylaxis and may need to know that in some (but not all) of these patients the anaphylaxis was a manifestation of a drug reaction. These types of contextual information can easily be represented using CDA R2 standards by nesting clinical statements, but a typical CRDW design may have a difficult time retaining this information. Further, parsing this data from a CDA document and transforming it into a dimensional relational database schema (e.g., star schema or snowflake schema) requires parsing logic to be more complicated and specific as to the document contents and thus fragile and susceptible to breakage with subtle changes in document structure.

To illustrate this limitation, consider Figure 3, a sample CDA R2 and the challenges of representing the encoded knowledge in a typical CRDW. Using this as a basis, Figure 4 provides a structured summary of the data in this XML fragment. The medical facts from Figure 3 observed in Figure 4 are:

1. The patient has a penicillin drug allergy. An observation (Observation-1) was made on the patient with concept code 416098002 (Code-1) from SNOMED CT with a human readable value of "Drug Allergy". This observation has value (Value-1) which has been assigned a code from RxNorm of 7016 with a human readable value of "Penicillin".

2. The Patient had Hives (Observation-2) has a concept code of "Assertion" (Code-2) and a value (Value-2) which has been assigned a SNOMED CT code of 247472004 with a human readable value of "Hives".

3. The patient's hives was a manifestation of his/her penicillin allergy. Observation-1 is related to Observation-2 using the entryRelationship tag (Relationship-1) with a typecode of "MFST". The facts Observation-1, Relationship-1 and Observation-2 form a clinical statement (Statement-1), where Observation-1 is the subject, Relationship-1 is the predicate and Observation-2 is the object.

4. This documented drug allergy is still active. Observation-3 has a concept code of 33999-4 (Code-3) which has a display name of "Status". Observation-3 has a value (Value-3) which has been assigned a SNOMED CT code of 55561003 which has a human readable value of "Active". This observation is contained within an entryRelationship tag (Relationship-2) with a typeCode of "REFR" or refers to, indicating that the active status refers to the enclosing act tag which represents the drug allergy as a whole. The facts in Observation-1, Relationship-

2 and Observation-3 form a clinical statement (Statement-2), where Observation-1 is the subject, Relationship-2 is the predicate and Observation-3 is the object.

```
<entry typeCode="DRIV">
  <act classCode="ACT" moodCode="EVN">
Observation 1 <entryRelationship typeCode="SUBJ" inversionInd="false">
      <observation classCode="OBS" moodCode="EVN">
Code 1          <code code="416098002" codeSystem="…."
                displayName="drug allergy" codeSystemName="../>
Value 1         <value xsi:type="CD" code="70618" codeSystem="…"
                  displayName="Penicillin" codeSystemName="…">
                </value>
Relationship 1  <entryRelationship typeCode="MFST" inversionInd="true">
Observation 2   <observation classCode="OBS" moodCode="EVN">
Code 2            <code code="ASSERTION" codeSystem="…"/>
                  <statusCode code="completed"/>
Value 2         <value xsi:type="CD" code="247472004" codeSystem="…"
                  displayName="Hives"></value>
                </observation>
              </entryRelationship>
Relationship 2  <entryRelationship typeCode="REFR">
Observation 3   <observation classCode="OBS" moodCode="EVN">
Code 3          <code code="33999-4" codeSystem="…"
                  displayName="Status" statusCode code="completed"/>
Value 3         <value xsi:type="CE" code="55561003" codeSystem = "…."
                    displayName="Active"> </value>
                </observation>
              </entryRelationship>
            </observation>
          </entryRelationship>
        </act>
</entry>
```

**Figure 3:** Sample CDA instance.

The sample XML in the Figure 3 demonstrates the rich semantic and contextual information that can be encoded using CDA R2. Representing these clinical statements using a typical dimensional CRDW like i2b2 requires three separate entries in the fact table. For example, Statement-1 has three entities Observation-1, Relationship-1 and Observation-2 which requires three rows: one for storing Observation-1 with its concept code (Code-1) and its value (Value-1), a second for storing Observation-2 with its concept code (Code-2) and its value (Value-2)

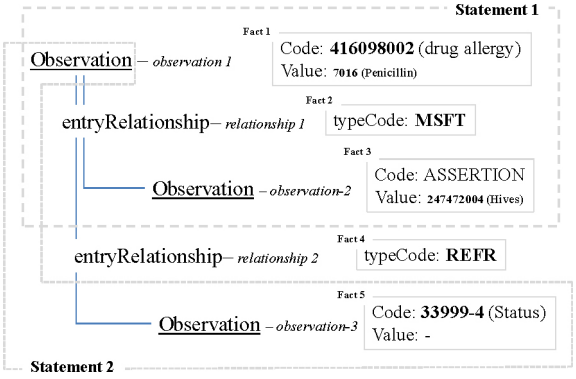


**Figure 4**: Statements and Facts from Figure 3.

and a third for Relationship-1 and its typeCode (MFST). However, if we attempt to add Observation-3 and Relationship-2, the parsing logic to accomplish this within i2b2 is extremely complex and as a result it becomes impossible to maintain all of the relationships between the different observations. This complexity also affects the query process; user queries for a single concept like penicillin or hives remains fairly straightforward but queries that relate hives with penicillin are very difficult to write in i2b2.

## III. Star Schema & Extraction Process

In this section, we present our proposed star schema and extraction process that takes EHR CDA documents as input and automatically populates a CRDW with a data set that is more representative of a population's clinical care. The extracted data is rich in *contextual information* – in the sense if the recorded patient's present medical condition is standalone or caused due to external factors or consequence of previous medical actions etc. and *semantic information* – which captures the nature of the medical event such as clinical observations, laboratory tests, doctors action, medical vocabulary etc. on the patients data that can be vital in support of meaningful use, evidence-based medicine, and accountable care. In Section III.A, we present the star schema that is a multi-dimension fact table capable of representing the observations (see Figure 4) from the CDA document (see Figure 3). In Section III.B, we detail the algorithms for the extraction process to take the CDA document and create an instance of the star schema. In Section III.C, to bring the concepts together more clearly, we provide an example of a star schema using our approach from a sample CDA document.

### A. *Proposed Star Schema*

In this section, we present our proposed star schema for capturing a complete data set of contextual and semantic information represented in an EHR structured using CDA R2 standards. Our initial aim was to leverage i2b2, an open-source platform that provides tools for managing and mining clinical data for research. The i2b2 framework consists of a collection of cells, each offering a distinct service made available through web services. Specifically, we had planned to use the i2b2 Data Repository Cell which is the dimensional CRDW that stores the clinical data used by other i2b2 cells. Due in part to the limitations described in Section 2 and our desire to capture contextual and semantic information, our research work has evolved to propose a star schema that is not compatible with the current i2b2 Data Repository Cell, but still incorporates other aspects

from the i2b2 design. The intent is to have our star schema be integrated with i2b2, as well as to work in a standalone mode.

The proposed star schema preserves the contextual and semantic information from the CDA R2 documents by designing a **CDA Statement Fact Table** and supporting it with new/reused dimension tables to form a star schema as shown in Figure 5; note that the dashed boxes in Figure 5 for Concept, Visit, Patient, and Observer dimensions are borrowed from the i2b2 schema. The CDA Statement Fact Table is structured to capture the details of each clinical statement encoded using CDA clinical statements. As described in Section 2 and shown in Figure 4, clinical statements in CDA R2 contain one or more instances of the Act class or Act subclasses. Multiple Acts within a clinical statement are connected by instances of the entryRelationship class. The Fact Table and dimension tables in Figure 5 are:

- **CDA Statement Fact Table**: Table I is main fact table is the core of the star schema to reference all of the other dimension tables.
- **Act Dimension:** Table II stores contextual information on each act occurring in the clinical statements. An instance of the Act class *represents any clinical action that can be documented.* Act subclasses include Observation, Procedure, Encounter, etc. The semantics of the act are described by the attributes such as *classCode, moodCode, negationInd, priorityCode*, etc.
- Substance Dimension: Table III stores information on medications administered to the patient and various attributes such as intake frequency, route prescribed for intake, and dosage value. In CDA, this data is contained within an instance of the *substanceAdminitration* class, is a subclass of *Act.*
- Predicate Dimension: Table IV stores relationships using the entryRelationship between two acts. .

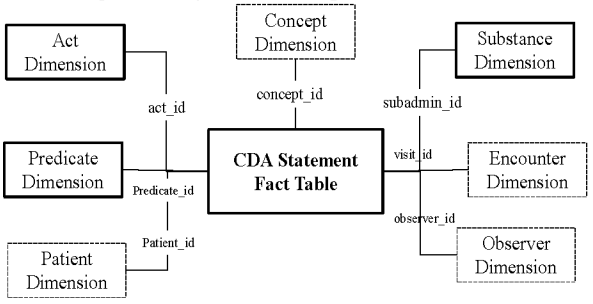


**Figure 5:** Proposed Star Schema for Contextual and Semantic Information.

**Table I: CDA Statement** fact table columns.

| Column Name | Column Description | Referring Dimension Table |
|---|---|---|
| Fact_id | Fact id | - |
| Patient_num | Refers Patient Dimension Table | Patient |
| Provider_num | Refers to Practitioner or Provider Dimension Table | Observer |
| Encounter_num | Refers to Encounter Dimension Table | Visit |
| Act_id | Refers to Act Dimension Table | Act |
| Subject_act_id | Refers to Act Table | Act |
| Subject_concept_code | Refers to the Concept Dimension Table | Concept |
| Subject_concept_value | Value of the Subject code which can also be a another concept | Concept |
| Substance_id | Refers to Substance Dimension Table | Substance |
| Predicate | Relationship in the between the concepts | - |
| Predicate_id | Refers to the Predicate Dimension Table | Predicate |
| Object_act_id | Refers to Act Dimension Table | Act |
| Object_concept_code | Refers to the Concept Dimension Table | Concept |
| Object_concept_value | Value of the Object code which can also be a another concept | Concept |
| Status | Status of the statement. | - |
| Parent_fact_id | Refers to Fact_id for linking multiple statements | CDA Statement Fact |

**Table II: Act Dimension** table columns.

| Act Column | Act Column Name |
|---|---|
| Act_Id | Act Primary Key. Referred in the CDA Statement fact table |
| Act_Class_Code | Type of Act |
| Act_Mood_Code | Mood of the Act |
| Act_Neg_Code | Captures the negation about the Act |
| Act_Priority_Code | Priority assigned by the Role to the Act |
| Act_Start_Date | Act Start date |
| Act_End_Date | Act End date |
| Act_Status | State of the Act |

**Table III: Substance Dimension** table columns.

| Medication Column Name | Medication Column Description |
|---|---|
| Sbad_Id | Primary Key. Referred in the CDA Statement fact table |
| Sbad_Freq | Medication frequency |
| Sbad_Freq_Unit | Medication frequency unit such as hour, day, month |
| Sbad_Route_Code | How was the medication taken |
| Sbad_Dose_Quantity | Dose of the medication |
| Sbad_Dose_Unit | Unit of dose such as puff, mg |
| Sbad_Admin_Code | |
| Sbad_Value_Type | Type of the value such as string, integer, double etc. |

**Table IV: Predicate Dimension** table columns.

| Column Name | Column Description |
|---|---|
| Predicate_id | Primary Key. Referred in the CDA Statement fact table |

| | |
|---|---|
| Predicate_inv_Ind | Roles of the source and target Acts were Reversed |
| Predicate_neg_ind | The Relationship is negated |
| Predicate_sep_Ind | Act is intended to be interpreted independently of the target Act |

The other dimension tables Concept, Visit, Patient and Observer are reused from current i2b2 schema. Using the proposed CDA Statement Fact Table, a single clinical statement (Statement-1 from Figure 4) will be inserted as a single row and multiple clinical statements can be connected using *parent_fact_id*. For example, for storing Statement-1 (Figure 4), the Observation-1 with code (Code-1) will be stored under the *subject_concept_code* column and its value (Value-1) is stored under the *subject_concept_value* column; the Relationship-1 typecode is stored under predicate column; and Observation-2 code (Code-2) is stored under the *object_concept_code* column and its value (Value-2) is stored under the *object_concept_value* column. The entities act, observation, and entityRelationship and attributes are shown in Figure 5.

To place the proposed schema into an appropriate perspective, suppose that Figure 1 is expanded to populate both i2b2 CRDW and our proposed star schema of Figure 5. In this case, two different ETL processes are performed: one which retains the *contextual and semantic* information in the form of clinical statements (Section III.B) and the other which strips of *contextual* data to meet the specifications of i2b2 schema. With this architecture and data model, users can query for data with a particular concept similar to the current i2b2 such as "select all patients with allergy to penicillin" and also go beyond i2b2 limitations to perform contextual and semantic queries such as "select all the patient with drug allergy to penicillin which causes severe manifestation of hives". The queries can also check for active status of the observation made. Thus, the proposed star schema accompanied and extraction algorithms (see Section III.B) will create instances (see Section III.C) that will allow a query engine to be guild that is able to provide richer semantic and contextual data set for CTR.

### *B. Extraction Algorithms*

The CDA extraction algorithm is built using Model Driven Health Tools (MDHT) library [12], a runtime Java based API built using UML, EMF architectures, and the CDA specifications itself. The MDHT is developed under the umbrella of the Open Health Tools project [13] dedicated to narrow the gap between health care and information technology. The algorithm is divided into multiple modules, where each module performs a specific function such as decoding observations or acts or procedures, medications, and database transaction manager, etc. The main entry point of the algorithm is shown in the Figure 6, where: the input CDA document is read as a file and parsed using MDHT API (Line 1); the document sections are iterated (Lines 2 and 3) for identifying list of Acts and Medications (Lines 5 Line 6) for that section (Figure 3); and individual modules are called for decoding the encoded knowledge (Lines 7 and 8). The output is a list of clinical statements (Figure 4) which can be stored in the star schema (Figure 5).

The *decodeObservations* module as shown in the Figure 7 is called from main program (Line 7) and is responsible for extracting clinical statements code-value pairs from various *Observations* encoded in a

**INPUT**: CDA Document; **OUTPUT** : List of Clinical Statements

```
1.  doc = CDAUtil.load("CDAExample.xml")
2.  sections= doc.getAllSections()
3.  while (sections.next())
4.  {
5.      List acts = section.getActs();
6.      List med = sections.getSubstanceAdministrations();
7.      decodeObservations();
8.      deocdeMedications();
9.  }
```

**Figure 6**: Main Module of Extraction Algorithm.

section. The input observations are iterated to list all the relationships associated with that observation (Line 14). If the end of the relationship is again another observation, the *decodeObservation* module is recursively called on the end observation, or the observation is decoded to obtain the code-value pairs. This module recursively traverses from one observation to another using the relationship between them and the clinical statements are built backwards. Consider the sample facts in Figure 4, first Observation-1 is encountered which is related to Observation-2 using Relationship-1. The module is called on Obervation-2 which is a leaf node of the tree and hence decoded first. The function is returned to Observation-1 which is decoded along with Relationship-1.

**Function** decodeObservations

**INPUT**: List of Acts ; **OUTPUT**: Observation Code – Value pairs

```
12. List facts;
13. while(acts.hasNext()){
14.     Relationships r = act.getEntryRelationships();
15.     while(r.hasNext()){
16.         Act a = r.getObservation();
17.         if(a !=null)
18.             DecodeActs(r.getObservation);
19.         else
20.             facts.add(a.getCode(), a.getValue);
21.     }
22. }
```

**Figure 7**: *deocodeObservations* Module.

The *decodeMedication* module called from the main program (Line 8) is shown in Figure 8, is responsible for deciphering medication and other metadata associated with it. The input medications are iterated to list all of the relationships associated with that medication (Lines 24 to 26) and traverse the path recursively (Lines 27 to 29) to decode the medications code value (Line 31) and other metadata if represented. Similar algorithms for decoding other entities such as *procedure, organizer, visit, insurance,* etc., have been developed, but are omitted due to space limits.

```
Function decodeMedications
INPUT: List of Medications ; OUTPUT: Medication Code - Value pairs
23.  List medications;
24.  while(meds.hasNext()){
25.       Relationships r = med.getEntryRelationships();
26.       while(r.hasNext()){
27.            Med m = r.getObservation();
28.            if(m !=null)
29.                 DecodeMedications(m);
30.            else
31.                 medications.add(m.getCode(), m.getValue);
32.       }
33.  }
```

**Figure 8**: *deocodeMedications* Module.

## C. *Instance of Proposed Schema*

In this section, we illustrate an instance of the star schema using the extraction algorithm discussed in Section III.B on small set of test data obtained from the NIST website [14]. Figure 9 illustrates a limited number of instances of the CDA Statement Fact Table and Act, Predicate, and Medication Dimensions. This represents the decoding of *observations* and *medication* values from the CDA documents using the extraction algorithm. The sample XML fragments in Figure 3 have two facts (Statememt-1 and Statement-2) as shown in Figure 4 and are captured as two rows (Fact id (1) and Fact_id(2)) in the CDA Statement Fact Table at top of Figure 9. As the two statements have a common observation (Observation-1), the two rows have the same value for the Subject_Act_id column. The two distinct values for Object_Act_id column capture the entities Observation-2 and Observation-3. The predicate row (Predicate_id(1)) captures the type of Relationship-1; (predicate_id(2)) captures the type of Relationship-2; and any other metadata about these relations in the Predicate dimension table. Since, the two rows (statements) belong to the same Act or section, they are connected using Parent_Fact_id.

The algorithm's merger module (not shown in Section III.B) is responsible for identifying the common act (which can be a subject or object) between the two rows to build the final clinical statement when displaying the query results to the user. In this

| Fact id | Act id | Subject Act Id | Subject Concept Code | Subject Concept Value | Substance id | Predicate | Predicate id | Object Act id | Object Concept Code | Object Concept Value | Parent fact Id |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 416098002 | 70618 | NULL | MFST | 1 | 3 | ASSERTION | 247472004 | NULL |
| 2 | 1 | 2 | 416098002 | 70618 | NULL | REFR | 2 | 4 | 33999-4 | 55561003 | 1 |
| 3 | 5 | 6 | 419511003 | 1191 | NULL | REFR | 3 | 7 | 33999-4 | 55561003 | NULL |
| … | … | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. |
| 10 | 13 | 14 | NULL | NULL | 1 | PRCN | 10 | 15 | ASSERTION | 56018004 | 9 |

| Act Id | Act_Class Code | Act_Mood Code | Act_Neg_Code | Act_Priority_Code |
|---|---|---|---|---|
| 1 | ACT | EVN | NULL | NULL |
| 2 | OBS | EVN | NULL | NULL |
| 3 | OBS | EVN | NULL | NULL |
| … | ….. | …… | ….. | ….. |
| 7 | OBS | EVN | NULL | NULL |

| Predicateid | Predicateinv Ind | Predicate_neg_ind | Predicate_sep_Ind |
|---|---|---|---|
| 1 | False | False | False |
| 2 | False | False | False |
| 3 | False | False | False |
| ….. | ….. | …… | …… |
| 10 | False | False | False |

| Sbad_id | Sbad | Sbad_freq | Sbad_freq_unit | Sbad_route_code | Sbad_dose_quantity | Sbad_dose_unit | Sbad_admin_code |
|---|---|---|---|---|---|---|---|
| 1 | 307782 | NULL | NULL | IPINHL | 2 | NULL | 415215001 |

**Figure 9**: Sample Instance of Proposed Schema.

example, the observation (Observation-1) is the common feature (Subject_Act_id(2)) between the two rows. The code and value columns (both subject and object) are further supported with other columns to identify the type of standard vocabulary the rows are referring to in the concept dimension table. When the user queries for a concept, the input is run on both subject and object code-value columns to provide the complete data set.

## IV. Related Work

In this section, we provide brief overview of selected research efforts on CRDW that are relevant to our work. Wisniewski, et al. [15] have developed a CRDW to recognize and report trends in antimicrobial use and resistance that causes hospital-acquired infections which result in increased health care costs and patient morbidity and mortality. CRDW data is gathered from TCDs in microbiology, pharmacy, radiology, medical records, etc., but doesn't capture semantic and contextual information about the patient data or their visit. Einbinder, et al. [16] have proposed a CRDW for serving clinical research, academic projects, business intelligence, etc., using TCDs such as medical records, billing systems, laboratory systems, cardiac surgery, etc. The system has a GUI for queries but does not consider semantic and contextual data.

Lyman, et al. [17] have developed a technique for sharing information across multiple systems for integrating disparate biomedical data to facilitate research, health promotion, and quality assessment with TCDs mapped to HL7 RIM model and data is transferred to the centralized CRDW. The Ohio state University Medical Center has developed Information Warehouse (IW) [18], an decision support, which has evolved into a comprehensive informatics platform supporting basic, clinical, and translational research, with TCDs for clinical data, research data, a development platform for building business/research applications, and, business intelligence environment assisting in reporting in all function areas.

## V. Conclusions and Ongoing Research

In this paper, we proposed a star schema and associated extraction process based on CDA documents to more easily create and populate a regional CRDW that contains a more complete data set that is more representative of a population's clinical care and captures the contextual and semantic relationships between the data. To support this work, we presented background material on HIE, CDA, and i2b2 limitations in Section 2. This lead to the proposal of the star schema, its associated extraction algorithms, and a sample instance in Section 3. To place our work in perspective, we reviewed related CRDW efforts in Section 4. We believe our approach is an important first step in making the use of CRDW more widespread, particularly to smaller organizations and hospitals. Our ongoing work is focused on finalizing extraction algorithms, and establishing a working CRDW test bed based on the star schema that demonstrates the ease of repository creation.

## References


[1] HITECH Act, http://healthit.hhs.gov/programs

[2] S. Timmermans and A. Mauck, "The Promises And Pitfalls Of Evidence-Based Medicine", *Health Affairs*, Vol. 24, no. 1, pp. 18-28, 2005.

[3] Data Warehousing Guide, Oracle 9i, 2002.

[4] i2b2, https://www.i2b2.org/

[5] HL7 CDA R1, http://www.hl7.org/index.cfm?ref=nav.

[6] HL7 CDA R2, http://www.hl7.org/index.cfm?ref=nav.

[7] K. W. Boone, "The CDA™ book", Springer, 2011.

[8] SNOMDED http://www.ihtsdo.org/snomed-ct/

[9] LOINC http://loinc.org/

[10] MeSH http://www.ncbi.nlm.nih.gov/mesh

[11] Integrating the Healthcare Enterprise http://www.ihe.net/

[12] Model Driven Health Tools , https://www.projects.openhealthtools.org/sf/projects/mdht/

[13] Open Health Tools, http://www.openhealthtools.org/index.htm

[14] HL 7 CDA R2 Sample Documents, http://xreg2.nist.gov/cda-validation/downloads.html

[15] M. F. Wisniewski, et al,, "Development of a Clinical Data Warehouse for Hospital Infection Control", *Journal of American Medical Informatics Association*, Vol. 10, pp. 454-462, 2003.

[16] J.S. Einbinder, et al., "Case study: a data warehouse for an academic medical center", *Journal Healthcare Information Management* Vol. 15, no. 2, pp. 165-75, 2001.

[17] J.A. Lyman, et al., "Mapping From a Clinical Data Warehouse to the HL7 Reference Information Model", *Proc. of 2002 AMIA Fall Symposium*

[18] J. Kamal, et al., "Information warehouse - a comprehensive informatics platform for business, clinical, and research applications", *Proc. of 2010 AMIA Fall Symposium.*

[19] S. Murphy, et al. 'Optimizing Healthcare Research Data Warehouse Design through Past COSTAR Query Analysis". *Proc. of 1999 AMIA Fall Symposium.*