Published online: 14.06.2019

Developing HL7 CDA-Based Data Warehouse for the Use of Electronic Health Record Data for Secondary Purposes

Fabrizio Pecoraro[1] Daniela Luzi[1] Fabrizio L. Ricci[1]

[1] National Research Council, Institute for Research on Population and Social Policies, Rome, Italy

Address for correspondence Fabrizio Pecoraro, PhD, National Research Council, Institute for Research on Population and Social Policies, Via Palestro 32, Rome 00185, Italy (e-mail: f.pecoraro@irpps.cnr.it).

ACI Open 2019;3:e44–e62.

Abstract

Background The growing availability of clinical and administrative data collected in electronic health records (EHRs) have led researchers and policy makers to implement data warehouses to improve the reuse of EHR data for secondary purposes. This approach can take advantages from a unique source of information that collects data from providers across multiple organizations. Moreover, the development of a data warehouse benefits from the standards adopted to exchange data provided by heterogeneous systems.

Objective This article aims to design and implement a conceptual framework that semiautomatically extracts information collected in Health Level 7 Clinical Document Architecture (CDA) documents stored in an EHR and transforms them to be loaded in a target data warehouse.

Results The solution adopted in this article supports the integration of the EHR as an operational data store in a data warehouse infrastructure. Moreover, data structure of EHR clinical documents and the data warehouse modeling schemas are analyzed to define a semiautomatic framework that maps the primitives of the CDA with the concepts of the dimensional model. The case study successfully tests this approach.

Conclusion The proposed solution guarantees data quality using structured documents already integrated in a large-scale infrastructure, with a timely updated information flow. It ensures data integrity and consistency and has the advantage to be based on a sample size that covers a broad target population. Moreover, the use of CDAs simplifies the definition of extract, transform, and load tools through the adoption of a conceptual framework that load the information stored in the CDA in the data warehouse.

Keywords

- data warehouse
- dimensional model
- Health Level 7
- Clinical Document Architecture
- electronic health record
- secondary uses

Background and Significance

There is a growing number of electronic health record (EHR) systems developed to gather clinical and administrative data during the different encounters between the patient and the health care professionals. The majority of the EHR systems are developed for a single institution (e.g., hospital) or a single provider (e.g., general practitioners [GP]) and/or gather data for a specific target population (e.g., diabetics, investigational patients) limiting the information collected to a restricted part of the individual's care pathway. Even if they are primarily developed to support service delivery as well as improve the communication between clinicians, there is a tendency to extend their functionalities to use their data to feed clinical data warehouses for secondary purposes,[1,2] such as biomedical research,[3] epidemiological studies,[4] ambulatory clinical care,[5] pharmacovigilance,[6] comorbidity detection,[7] or to alert providers of potential clinical risks.[8] Moreover, this solution is

received
August 1, 2018
accepted after revision
April 11, 2019

DOI https://doi.org/10.1055/s-0039-1688936.
ISSN 2566-9346.

adopted by different parties (e.g., hospital, GP, specialists)[9,10] and at different organizational level (i.e., local, regional, and national authorities).[11,12] Several benefits of data warehousing in health care have been already demonstrated[13,14] under different perspectives, such as supporting clinical research,[15–18] decision making,[16,19–21] and the accomplishment of strategic business objectives.[22]

In our approach, we consider the cross-institutional EHR that provides a more comprehensive description of the patient's health status with a complete and consolidated lifetime medical history described by the different types of clinical documents generated by different providers and available across multiple health care organizations.[23] The EHR integrates heterogeneous information systems in a distributed environment, including systems developed for primary and secondary care (e.g., GP's and specialist's record), ambulatory and hospital settings (e.g., laboratory and radiology information system), etc. The development of a data warehouse based on EHR takes advantage on the already standardized data model adopted to harmonize and integrate the different EHR source systems developed by standard bodies such as Health Level 7 (HL7) and openEHR. Moreover, these standards are based on common vocabularies that specify the exact meaning of clinical data despite cultural and language differences, using widely adopted standard terminologies such as Systematized Nomenclature of Medicine, Logical Observation Identifiers Names and Codes (LOINC) or International Classification of Diseases.

In this procedure, a crucial issue is the design and implementation of the extract, transform, and load (ETL) tools that aim to structure data to be easily extracted and analyzed under a statistical point of view. This task generally requires the integration of data provided by different source systems taking into account, for instance, operating systems, communication protocols, and database management systems. In our approach, the adoption of HL7 Clinical Document Architecture (CDA) documents allow us to access data already harmonized within each type of CDA specification moving the integration issue from a source system to a document template point of view.

Materials and Methods

In this article, we propose a methodological approach to facilitate the design and development of an ETL tool that extracts information from the source EHR system, transforms data according to the snowflake schema representation, and then loads them in a specialized data warehouse that processes them for secondary purposes. Thus, a semi-automatic conceptual framework is defined and implemented to facilitate these transformation procedures mapping the primitives of the data warehouse dimensional model with the HL7 CDA classes. In particular, this paper describes: 1. the data warehouse architecture; 2. the ETL design process providing a formal definition of the conceptual framework based on the first-order logic; and 3. the application of this formal definition to implement the ETL tool using the eXtensible Stylesheet Language (XSL) to produce, as a result, the XSL Transformations (XSLT) document. An example of the transformation is provided to demonstrate the feasibility of our approach. The application of this methodology is tested considering on the one hand different CDA specifications to describe the same business process, and on the other hand on a set of Continuity of Care Documents (CCDs) to define a clinical dashboard on patients with diabetes.

Data Warehouse Architecture

►Fig. 1 shows a possible data warehouse solution considering two perspectives: on the left side a three-layer architecture is presented taking into account the system point of view, whereas on the right side the different formats adopted to represent data in each layer are reported to highlight the different transformations needed to extract data from heterogeneous data sources and integrate them to be used for secondary purposes.

The data source layer is represented by a set of legacy systems and repositories (e.g., GP's electronic health care record, laboratory information systems, radiology information systems) that manage health care and administrative information related to citizens. The identification of information systems to be included in this layer represents a critical process to be performed for the success of a data warehouse project, as it needs the specification of the role played by selected data sources in the design and development of the data warehouse and data marts. From the model point of view, as highlighted in the right side of ►Fig. 1, the source information systems can represent data using different formats, such as relational databases, XML, flat/binary files, and spreadsheets. Moreover, as each format can map data on the basis of different models, there is a need to implement procedures that match the different schemas adopted, considering that even schemas for identical concepts may have structural and naming differences, schemas may model similar but nonidentical contents, may be expressed in different data models, may use similar words with different meanings, etc. Moreover, to effectively extract data from source systems it is necessary to manage the distinct set of characteristics of each data source, such as differences in database management systems, operating systems, and communications protocols. This makes it necessary to implement an ETL tool (ETL_1) that consolidates these different representations in a common data model. To achieve this aim some authors have proposed to use a module called wrapper[24] that is responsible for data gathering from different sources, data cleansing, format conversions, as well as data integration. Data managed by each wrapper can therefore be loaded in the data warehouse and then in the relevant data mart.

However, a change in a data warehouse schema makes the revision of each wrapper a not straightforward task. Therefore, it is necessary to include an intermediate stage between the data sources and the data warehouse tiers in the architecture. This middleware system, called operational data store (ODS), contains detailed and integrated data with

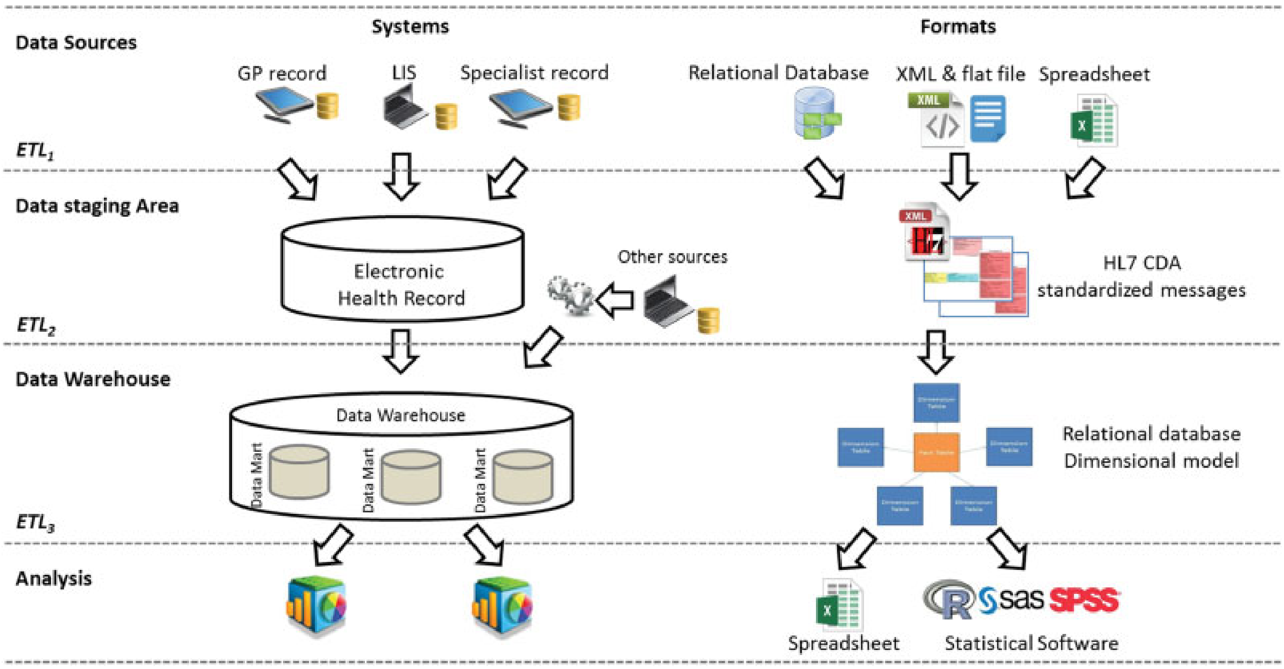


Fig. 1 Data warehouse description. On the left side, the data warehouse architecture is presented from the source systems point of view, and on the right side, the different data formats used to represent the information at each layer of the architecture are reported.

specific constraints including referential integrity that ensures data accessibility by relevant units. For the purpose of this study, in the proposed architecture the ODS can be represented by the EHR and modeled as a central repository that contains individual's structured clinical documents modeled using the HL7 CDA. These documents are subsequently used to feed the data warehouse based on an On-Line Analytical Processing (OLAP) approach that facilitates the integrated analysis to develop specific dashboards based on business processes and clinical indicators. To perform this task, a second ETL tool (named ETL_2 in ►Fig. 1) has to be designed and implemented. This stage of the data warehouse design process is the central part of this article and is described in the following paragraphs focusing the attention on the conceptual framework adopted to implement a tool that semiautomatically extracts and transforms the EHR data in a data warehouse dimensional model representation.

Finally, the analysis layer concerns tools and techniques for data analysis, such as data mining, reporting, and OLAP tools. For instance, they can be used to define a set of clinical indicators, as highlighted in the following. From the model point of view, it means to translate information in different data formats to be analyzed by the relevant tools, from an Excel spreadsheet to specific statistical software.

ETL Process Design

ETL process design is one of the main phases of the data warehouse lifecycle. Its general framework entails three main steps: (1) data are extracted from different data sources, and then (2) transformed and cleansed before being (3) loaded to the data warehouse. As highlighted in the previous paragraph, the proposed three-layer architecture comprises two ETL processes. In the first one, data are extracted from the legacy sources and transformed to be stored in the ODS that is represented by the EHR, where information are structured using standardized HL7 CDA documents. The second ETL process extracts data collected in these documents and transforms them to be loaded in the data warehouse. In this article, the attention is focused on the design and development of a tool to model this latter ETL process proposing a conceptual framework that maps the HL7 CDA components with the primitives of the data warehouse logical model. The feasibility of the proposed approach is demonstrated providing a case study based on the laboratory results collected in the CCD[25,26] to define clinical outcome indicators for quality assessment, such as percentage of patients with a vital sign parameter within a specific threshold (e.g., glycated hemoglobin under 7%). In the next paragraphs both the dimensional model and the HL7 CDA schema are described using formalism based on the first-order logic and subsequently mapped to define the conceptual framework.

Formal Definition of the Dimensional Model

A widely accepted formalization of the data warehouse conceptual modeling is the dimensional model that is represented as a fact table surrounded by independent dimensions.[27] The former specifies the measurements of a business process performance in a qualitative and/or quantitative way (e.g., episodes of care, clinical outcome), whereas each dimension describes a collection of reference information about a measurable event collected in the fact table (e.g., time, patient, location, providers). There are two main types of dimensional models depending on the level of denormalization of the dimensions: (1) star schema where dimensions are modeled using independent denormalized tables which are not related

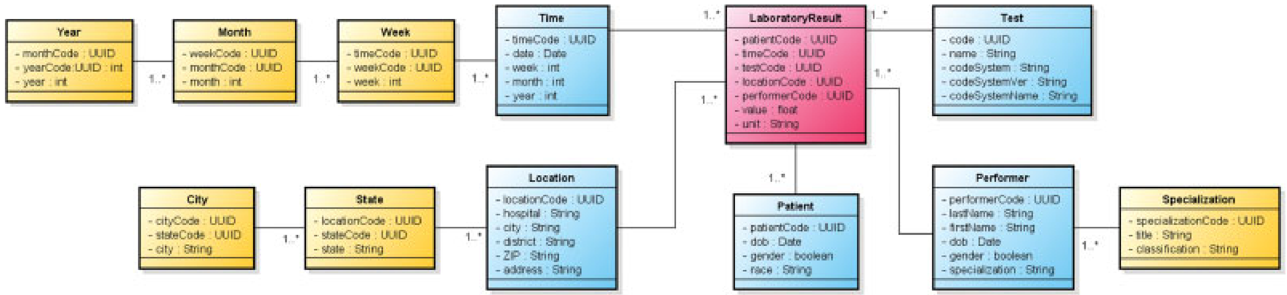


Fig. 2 Snowflake schema dimensional model. Example of a dimensional model representing a snowflake schema composed by a fact LaboratoryResult related to five dimensions: two denormalized tables (Test and Patient) and three hierarchies (Time, Performer, and Location).

with each other; and (2) snowflake schema where each dimension is represented by a normalized hierarchy. Moreover, it is possible to combine multiple star schemas to define a constellation model that contains multiple fact tables that share the same dimension tables. The formalization of the snowflake schema is proposed in Definition 3 taking into account the one-to-many relationship between two classes (Definition 1) and the hierarchy structure that is a central concept to model dimensions (Definition 2). A high-level representation of this schema is shown in ►Fig. 2 considering the case study of laboratory results modeled using the CCD document. It is composed by a LaboratoryResult Fact table surrounded by five dimensions: two denormalized tables (Test and Patient) and three normalized hierarchies (Time, Performer, and Location).

Definition 1. One-to-Many Association

This type of association relates two classes C_1 and C_2 in which an element of C_1 may be linked to many elements of C_2, and an element of C_2 may be linked to only one element of C_1.

Let $c(a_j, 1 \leq j \leq n)$ be a tuple describing the n properties of an object, a class C is defined as $C\{c_i\}$ with $1 \leq i \leq k$.

Let C_1 and C_2 be two classes as defined above and the one-to-many association between them is defined as $R(C_1,C_2)$ which satisfies the following conditions:

- $\forall c_1 \in C_1\ \exists c_2 \in C_2 \mid c_2.FK = c_1.PK$
- $\forall c_2 \in C_2\ \exists!\ c_1 \in C_1 \mid c_1.PK = c_2.FK$

where PK and FK are specific properties (attributes) of the relevant class. PK uniquely identifies each instance of the class c_1 and FK specifies the c_1 instance related to the instance of the class c_2.

An example of a one-to-many association is reported in ►Fig. 2 where the patientCode primary key of the Patient dimension (C_1) is associated with one or more entries of the patientCode foreign key of the LaboratoryResult fact (C_2).

Definition 2. Hierarchy

A hierarchy is a data model where data are organized into a tree-like structure with a cascade series of classes related to a many-to-one relationships.

Let $\{C_1,\dots,C_k\}$ be a set of classes, the hierarchy is defined as $H(C_1,\{C_i, 2 \leq i \leq k\})$ which satisfies the following condition:

- $\forall i \in \{1,\dots,k-1\},\ \exists!\ R_i \mid R_i = R(C_i, C_{i+1})$

The set of ordered classes of the hierarchy is returned by the function $f_s(H) = \{C_1,\dots,C_k\}$.

An example of a hierarchy is highlighted in ►Fig. 2 where the classes Time, Week, Month, and Year represent the $\{C_1,C_2, C_3,C_4\}$ classes, respectively.

Definition 3. Snowflake schema

This type of dimensional model is composed by a fact table related to a set of dimensions represented by normalized hierarchy.

Let F be a class, defined as $F\{c_i\}$ with $1 \leq i \leq k$ where $c(a_j, 1 \leq j \leq n)$ is a tuple describing the n properties of an object. Let $R_{sf} = \{R_{sf_i}, 1 \leq i \leq m\}$ be a set of one-to-many relationships and $H = \{H_i, 1 \leq i \leq m\}$ a set of hierarchies, the snowflake schema is defined as $S_{sf}(F,R_{sf})$ which satisfies the following condition:

- $\forall i \in \{1,\dots,m\},\ \exists!\ R_{sf_i} \in R_{sf} \mid R_{sf_i} = R(F, H_i.C_1)$

►Fig. 2 highlights an example of a snowflake schema composed by a fact *F* LaboratoryResult related to five dimensions described by three denormalized tables (Test, Patient, and Location) and two hierarchies (Time and Performer).

HL7 CDA

CDA Release 2 Level 3 records clinical observations and services in a mark-up structured standard document based on the six backbone classes of the HL7 Reference Information Model (RIM): Act, ActRelationship, Participation, Entity, Role, and RoleLink. Each business process can be modeled by decomposing it into an elementary description based on a speech act[28] describing the action performed or scheduled (i.e., represented by the Act class). Moreover, the triple <Participation, Role, Entity> modeled as a hierarchy is adopted to describe subjects/objects involved in the process as well as the role played by them within the action.[29]

For instance, considering the portion of the CDA reported in ►Fig. 3 that models the laboratory results collected in the CCD, the triples <recordTarget, patientRole, Patient> and <recordTarget, patientRole, Organization> are used to represent the medical record belonging to the relevant ClinicalDocument and the triple <performer, associatedEntity, Person> specifies the practitioner that performed a specific event, such as an Observation. The RIM triple and its relationship with the relevant Act are formalized respectively in Definitions 4 and 5.

Definition 4. HL7 Hierarchy

Similarly to Definition 2, a HL7 hierarchy is a data model where data are organized into a tree-like structure with a

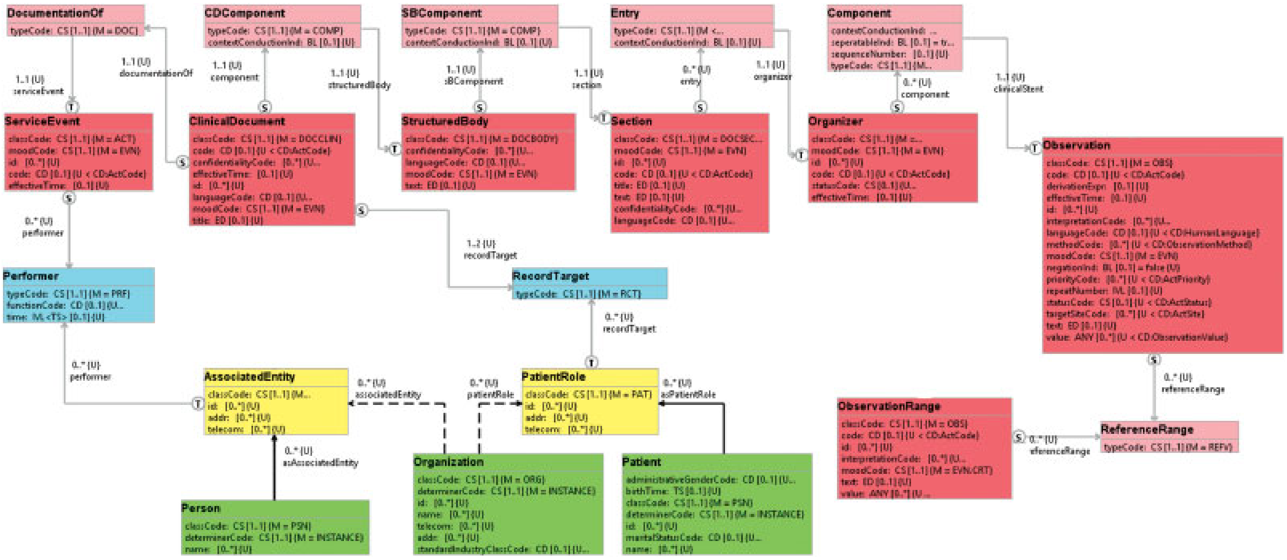

DocumentationOf
typeCode: CS [1..1] {M = DOC}
CDComponent
typeCode: CS [1..1] {M = COMP}
contextConductionInd: BL [0..1] {U}
SBComponent
typeCode: CS [1..1] {M = COMP}
contextConductionInd: BL [0..1] {U}
Entry
typeCode: CS [1..1] {M <...
contextConductionInd: BL [0..1] {U}
Component
contextConductionInd: ...
seperatableInd: BL [0..1] = tr...
sequenceNumber: [0..1] {U}
typeCode: CS [1..1] {M...
1..1 {U}
serviceEvent
1..1 {U}
documentationOf
1..1 {U}
component
1..1 {U}
structuredBody
1..1 {U}
sBComponent
1..1 {U}
section
0..* {U}
entry
1..1 {U}
organizer
0..* {U}
component
1..1 {U}
clinicalStent
T
S
ServiceEvent
classCode: CS [1..1] {M = ACT}
moodCode: CS [1..1] {M = EVN}
id: [0..*] {U}
code: CD [0..1] {U < CD:ActCode}
effectiveTime: [0..1] {U}
ClinicalDocument
classCode: CS [1..1] {M = DOCCLIN}
code: CD [0..1] {U < CD:ActCode}
confidentialityCode: [0..*] {U...
effectiveTime: [0..1] {U}
id: [0..*] {U}
languageCode: CD [0..1] {U...
moodCode: CS [1..1] {M = EVN}
title: ED [0..1] {U}
StructuredBody
classCode: CS [1..1] {M = DOCBODY}
confidentialityCode: [0..*] {U...
languageCode: CD [0..1] {U...
moodCode: CS [1..1] {M = EVN}
text: ED [0..1] {U}
Section
classCode: CS [1..1] {M = DOCSEC...
moodCode: CS [1..1] {M = EVN}
id: [0..*] {U}
code: CD [0..1] {U < CD:ActCode}
title: ED [0..1] {U}
text: ED [0..1] {U}
confidentialityCode: [0..*] {U...
languageCode: CD [0..1] {U...
Organizer
classCode: CS [1..1] {M =...
moodCode: CS [1..1] {M = EVN}
id: [0..*] {U}
code: CD [0..1] {U < CD:ActCode}
statusCode: CS [0..1] {U...
effectiveTime: [0..1] {U}
Observation
classCode: CS [1..1] {M = OBS}
code: CD [0..1] {U < CD:ActCode}
derivationExpr: [0..1] {U}
effectiveTime: [0..1] {U}
id: [0..*] {U}
interpretationCode: [0..*] {U...
languageCode: CD [0..1] {U < CD:HumanLanguage}
methodCode: [0..*] {U < CD:ObservationMethod}
moodCode: CS [1..1] {M = EVN}
negationInd: BL [0..1] = false {U}
priorityCode: [0..*] {U < CD:ActPriority}
repeatNumber: IVL [0..1] {U}
statusCode: CS [0..1] {U < CD:ActStatus}
targetSiteCode: [0..*] {U < CD:ActSite}
text: ED [0..1] {U}
value: ANY [0..*] {U < CD:ObservationValue}
0..* {U}
performer
1..2 {U}
recordTarget
Performer
typeCode: CS [1..1] {M = PRF}
functionCode: CD [0..1] {U...
time: IVL <TS> [0..1] {U}
RecordTarget
typeCode: CS [1..1] {M = RCT}
0..* {U}
recordTarget
0..* {U}
referenceRange
0..* {U}
performer
AssociatedEntity
classCode: CS [1..1] {M...
id: [0..*] {U}
addr: [0..*] {U}
telecom: [0..*] {U}
0..* {U}
associatedEntity
0..* {U}
patientRole
PatientRole
classCode: CS [1..1] {M = PAT}
id: [0..*] {U}
addr: [0..*] {U}
telecom: [0..*] {U}
0..* {U}
asPatientRole
ObservationRange
classCode: CS [1..1] {M = OBS}
code: CD [0..1] {U < CD:ActCode}
id: [0..*] {U}
interpretationCode: [0..*] {U...
moodCode: CS [1..1] {M = EVN,CRT}
text: ED [0..1] {U}
value: ANY [0..*] {U...
0..* {U}
referenceRange
ReferenceRange
typeCode: CS [1..1] {M = REFV}
0..* {U}
asAssociatedEntity
Person
classCode: CS [1..1] {M = PSN}
determinerCode: CS [1..1] {M = INSTANCE}
name: [0..*] {U}
Organization
classCode: CS [1..1] {M = ORG}
determinerCode: CS [1..1] {M = INSTANCE}
id: [0..*] {U}
name: [0..*] {U}
telecom: [0..*] {U}
addr: [0..*] {U}
standardIndustryClassCode: CD [0..1] {U...
Patient
administrativeGenderCode: CD [0..1] {U...
birthTime: TS [0..1] {U}
classCode: CS [1..1] {M = PSN}
determinerCode: CS [1..1] {M = INSTANCE}
id: [0..*] {U}
maritalStatusCode: CD [0..1] {U...
name: [0..*] {U}

Fig. 3 Health Level 7 (HL7) Clinical Document Architecture (CDA) schema based on the Reference Information Model (RIM). Portion of the CDA document to represent the data collected in the laboratory result business process modeled using the Continuity of Care Document (CCD) implementation.

cascade series of classes related with a many-to-one relationship. In the HL7 RIM this tree-like structure is composed, respectively, by the classes Participation, Role, and Entity.

Let us consider C_P, C_R and C_E be partitions of the class C representing, respectively, a Participation, a Role, and an Entity of the HL7 RIM. The HL7 Hierarchy is defined as $H_{HL7} = H(C_P C_R, C_E)$.

An example of HL7 Hierarchy (H_{HL7}) is reported in ►Fig. 3 where the three classes Performer, AssociatedEntity, and Person represent respectively the Participation (C_P), the Role (C_R), and the Entity (C_E) classes.

Definition 5. HL7 Act-Hierarchy Association

Similarly to Definition 1, this association is a one-to-many relationship between a Participation class of a HL7 Hierarchy and an Act class.

Let H_{HL7} be a HL7 Hierarchy, the HL7 Act-Hierarchy association is defined as $R_{A\text{-}H} = R(C_A, H_{HL7}, C_P)$, where C_P is the Participation class of the hierarchy H_{HL7} and C_A is the Act class related with it.

An example of this association is shown in ►Fig. 3 where the Participation Performer (C_P) is related with the Act class ServiceEvent (C_A).

Often the main Act modeled by the clinical document is associated with other Acts to indicate, for instance, temporal, logical, or structural order of related events or to group similar events. The RIM represents this association using the ActRelationship class that links a source Act with a target Act involving two one-to-many associations. As a result, the association between two Acts is modeled on the basis of the one-to-many relationship, as formalized in Definition 6.

Definition 6. HL7 Act-Relationship

Similarly to Definition 1, this association is a one-to-many relationship between two Act classes of the HL7 RIM. In particular, an element of a source Act may be linked with many elements of a target Act, and an element of the target Act may be linked to only one element of the source Act.

Let C_A^S, C_A^T be two partitions of the Act class C_A. The HL7 Act-Relationship is defined as $R_{A\text{-}A} = R(C_A^S, C_A^T)$, , where C_A^S and C_A^T are respectively the source and the target class of the association.

An example of this relationship is presented in ►Fig. 3 where the Act classes Observation and ObservationRange describe respectively the source (C_A^S) and the target (C_A^T) classes of the relationship ReferenceRange (R_{A-A}).

The Act specializations and their relationships are the basis to define the backbone of the clinical document. In particular, the root class of all CDA documents is the Act ClinicalDocument that is composed by Sections, each one collecting a set of events modeled using the Act class of the ClinicalStatement pattern. A ClinicalStatement is a choice structure containing a set of Act specializations depending on the event to be documented, such as Observation and Procedure. The CDA Backbone can be considered as a hierarchy and formalized in Definition 7.

Definition 7. CDA Backbone Hierarchy

Similarly to Definition 2, a CDA backbone hierarchy is a data model where data are organized into a tree-like structure with a cascade series of classes related to a many-to-one relationship. In the HL7 CDA this tree-like structure is composed by a set of Acts.

Let $\{C_{A_1},\ldots,C_{A_k}\}$ be an ordered set of classes, the CDA Backbone Hierarchy is defined as $H_B = H(C_{A_i}, 1 \le i \le k)$ where $\forall C_{A_i} \in f_s(H_B)$ is an Act class of the RIM.

In the portion of the CDA document reported in ►Fig. 3, the following Act classes are identified to define the CDA Backbone Hierarchy (H_B): ClinicalDocument (C_{A_1}), StructuredBody (C_{A_2}), Section (C_{A_3}), Observation (C_{A_4}), and ObservationRange (C_{A_5}).

Both the HL7 Hierarchy and the CDA Backbone represent the main components used to define the CDA model as formalized in Definition 8.

Definition 8. CDA Model

This type of model comprises a CDA backbone hierarchy composed by a set of Act classes each one related to a set of HL7 hierarchies.

Let H_B be the CDA Backbone Hierarchy, $\{R_{A-Hi}, 1 \le i \le m\}$ be a set of HL7 Act-Hierarchy Associations defined on a set of HL7 Hierarchies $\{H_{HL7_i}, 1 \le i \le m\}$,, and $\{R_{A-Ai}, 1 \le i \le n\}$ be a set of HL7 Act-Relationships, the CDA model is defined as $S_{CDA}(H_B, \{R_{A-Hi}, 1 \le i \le m\}, \{R_{A-Ai}, 1 \le i \le n\})$.

►Fig. 3 shows a portion of a CDA model composed by a set of Act classes (ClinicalDocument, StructuredBody, Section, Organizer, Observation, ObservationRange) of the Backbone Hierarchy (H_B) as well as by the following HL7 Hierarchies: <specimen, specimenRole, PlayingEntity> , <recordTarget, patientRole, Patient> , <participant, associatedEntity, Person> , and <performer, associatedEntity, Person> .

Mapping the Dimensional Model with the CDA Schema

The conceptual framework proposed in this section is described considering the 4-step life-cycle described by Kimball and Ross[27]: starting from the business process to be modeled and the level of detail to be captured, we first identify the fact and subsequently the dimensions of the model. The steps to perform this mapping are summarized in ►Table 1.

When these steps are completed, the dimensional model can be refined by denormalizing the schema hierarchies as well as by pruning and grafting unnecessary attributes and tables. Moreover, given that in the CDA schema attributes are represented using complex data types (derived from the RIM), the process requires the adoption of resolving techniques that analyze a complex CDA attribute and store each property in a single column of the relevant class of the dimensional model schema. For instance, the attribute *code* of the Observation class of the CDA is described by different elements: a *code* value and a *codeSystem* identifier to specify an externally defined coding scheme. Moreover, the *codeSystem* may have a *codeSystemName* helping human interpretation. It may also have an optional *displayName* element containing the text that was originally written. These four elements are mapped in four attributes of the relevant class of the dimensional model schema. These steps of the mapping framework are described in details in the following sections as well as in the implementation procedures described in ETL process implementation paragraph.

Choose the Business Process and Declare the Grain

In this step, the activities performed by the health care organization (e.g., laboratory results, patient encounter workflow, physician orders) are determined and prioritized to identify the business process to be modeled, based on different criteria such as significance, feasibility, complexity, and data quality of source systems. The determination of a business process can be based on the HL7 RIM Act class that represents "an intentional action that can be either executed, ordered, planned, and must be documented." For this reason, it is important to identify which Act is a feasible representation of the event taking into also account the level of granularity to be represented. In this perspective, the CDA is mainly described by three Acts: (1) ClinicalDocument that provides an high level description of the document; (2) Section that aggregates in a single "narrative block" core patient-specific data based on common clinical conventions

Table 1 Conceptual mapping of HL7 CDA model and snowflake schema

Let $S_{CDA}(H_B, \{R_{A-Hi}, 1 \le i \le m\}, \{R_{A-Ai}, 1 \le i \le n\}$ be the CDA model (Def. 8) and $S_{sf}(F,R_{sf})$ be the snowflake schema (Def. 3) to map the S_{CDA} elements in the S_{sf} the following steps should be performed	
Step 1. Identify the Fact. This class of the dimensional model can be chosen among the Act classes of the CDA depending on the data to be collected and the analysis to be done	
Let C_A^F be an Act class of the CDA Backbone Hierarchy $H_B((C_A, 1 \le i \le r-1), C_A^F, (C_A, r+1 \le i \le l))$ (Def. 7) so that $C_A^F \in f_s(H_B)$ used to represent the Fact class *F* of the dimensional model $S_{sf}(F,R_{sf})$ (Def. 3)	
Step 2. Identify the Dimensions. Starting from the Act chosen and a Fact of the dimensional model, the CDA document is navigated in both parent–child and child–parent direction to include the relevant Act classes and HL7 Hierarchies to the data warehouse dimensional model	
a	Each class C_A^T related to the class C_A^F as a target of an HL7 Act-Relationship $(R_{A-A}(C_A^F, C_A^T))$ (Def. 6) is included in the dimensional model along with its related HL7 Hierarchies • $\forall R_{A-A} \mid R_{A-A} = R(C_A^F, C_A^T) \rightarrow R_{A-A} \in R_{sf}$
b	Each Act class C_A^S included in the CDA Backbone Hierarchy $H_B^{'} = H_B(C_{A_i}, 1 \le i \le r-1)$ (Def. 8) and related to the class C_A^F as a source of an HL7 Act-Relationship (Def. 6), is included in the dimensional model • $\forall C_{A_i} \in f_s(H_B^{'}) \mid \exists R_{A-A}(C_A, C_A^F) \rightarrow R_{A-A} \in R_{sf}$ Each HL7 Hierarchy (Def. 4) and each Act related to the class C_{A_i} (HL7 Act-relationship, Def. 6) are included in the dimensional model • $\forall C_{A_i} \in f_s(H_B^{'}), \forall H_{HL7} \mid \exists R_{A-H}(C_A, H_{HL7}) \rightarrow R_{A-H} \in R_{sf}$ • $\forall C_{A_i} \in f_s(H_B^{'}), \forall C_A^T \mid \exists R_{A-A}(C_A, C_A^T) \rightarrow R_{A-A} \in R_{sf}$
c	Each H_{HL7} HL7 Hierarchy (Def. 4) related to the class C_A^F by an HL7 Act-Hierarchy association $R_{A-H}(C_A^F, H_{HL7}, C_P)$ (Def. 5) is included in the dimensional model • $\forall H_{HL7} \mid \exists R_{A-H}(C_A^F, H_{HL7}, C_P) \rightarrow R_{A-H} \in R_{sf}$

Abbreviations: CDA, Clinical Document Architecture; HL7, Health Level 7.

Note: The definitions reported in the ETL process design paragraph are used to map each primitive of the CDA model with a concept of the dimensional model.

(e.g., laboratory test results, medications, problems, procures); and (3) clinicalStatement, that is, a choice structure that represents the content of a specific action (e.g., observation, procedure, substance administration).

In particular, in the CCD implementation the class Section through the attribute code is a standardized identifier to classify clinical business processes coded using the LOINC nomenclature. Some examples of the business processes that can be described on the basis of the CCD documents are reported in the following:

- Alert: allergies, adverse reactions, and alerts.
- Medication: patient's current medications and pertinent medication history.
- Procedure: interventional, surgical, diagnostic, and therapeutic procedures, and treatments pertinent to the patient historically.
- Result: results of observations generated by laboratories, imaging procedures, and other procedures.

As highlighted previously, in this article the attention is posed on the results achieved in the business process case study that collects individual's clinical findings, such as blood pressure, heart rate, body mass index, and glycated hemoglobin. The portion of the CCD document describing this business process is shown in ►Fig. 3.

Once the business process is detected, the next step is to declare the grain of the model depending on the level of detail of the information to be collected in the data warehouse.[30] A high level means that each transaction (e.g., blood pressure) is stored in the data warehouse, whereas a low level indicates that the information is stored after a summarization (e.g., average value of the blood pressure over a specific period). The grain declaration is based not only on the objective of the data warehouse design, but also on the granularity of data contained in the clinical document.

Identify the Fact

A Fact describes the relevant event to be analyzed trough qualitative and quantitative measures that represent the performance of the business process and that could be analyzed using statistical methods. Looking at the CDA model, the fact can be chosen between the Act classes comprised in the CDA Backbone Hierarchy given that they represent "measurement of health care business processes." The choice depends on the purpose of the analysis to be performed, on the indicators to be developed, and on the event to be investigated. For this reason, in our approach the Acts that define the CDA Backbone can be considered as suitable candidates to identify the Fact of the dimensional model depending on the purpose of the analysis to be performed and on the indicators to be developed.

This phase is described in the step 1 of the conceptual mapping lifecycle reported in ►Table 1 and also proposed again in the following to simplify the readiness of the framework. The Fact is selected among the Acts of the CDA Backbone Hierarchy $f_s(H_B)$.

Examples of Act that can describe related actions and events that constitute health care services are reported in ►Table 2 providing some examples of the business processes and measures.

Step 1. Identify the Fact. This class of the dimensional model can be chosen among the Act classes of the CDA depending on the data to be collected and the analysis to be done

Let C_A^F be an Act class of the CDA Backbone Hierarchy $H_B\left(\left(C_A, 1 \le i \le r-1\right), C_A^F, \left(C_A, r+1 \le i \le l\right)\right)$ (Def. 7) so that $C_A^F \in f_s(H_B)$ used to represent the Fact class *F* of the dimensional model $S_{sf}(F, R_{sf})$ (Def. 3).

Once the Fact has been determined, its attributes are analyzed to define measures that represent a qualitative or quantitative evaluation of the business process and that could be analyzed using statistical methods. In the RIM numerical information is collected in the Act class attributes modeled with quantity (i.e., QTY) or physical quantity data type (i.e., PQ), whereas qualitative data are specified using coded data types (e.g., CV, CE, CD). For instance, the Observation class contains two measures described by the attribute value and interpretationCode that represents respectively a quantitative and qualitative measure of the event observed. Considering the SubstanceAdministration event, three quantitative measures can be detected: doseQuantity, rateQuantity, and maxDoseQuantity that model the medication quantity given per dose, the rate at which the dose is to be administered, and the maximum medication dose given over time.

Identify the Dimensions

In this article, dimensions are determined based on the Zachman framework[31] that provides a systematic of information related to the investigated event: who (persons), what (the fact), when (the time), where (the location), why (the reason), and how (the manner). To identify suitable candidates to derive dimensions, we start analyzing the two main structural components of the CDA document related to the Fact class: (1) Acts that captures the meaning and purpose of each association with the main event as well as additional actions to determine, for instance, why the event has been performed or the criteria used to evaluate the event outcome; (2) HL7 Hierarchy that describes the functions of subjects and objects involved in a specific process, identifying, for instance, who performed it (i.e., performer), for whom it was done (i.e., subject), and where it was done (i.e., location). These data are captured through the attribute typeCode of the Participation class that specifies its meaning and purpose using a controlled vocabulary defined by HL7.

To map each HL7 Act-Relationship with the dimensional model, the CDA schema is navigated in two directions, starting from the Act class chosen as a fact table in the previous step (C_A^F). These steps are described in steps 2a and b of the process reported in ►Table 1 and also proposed again in the following to simplify the readiness of the framework.

In particular, in step 2a each Act class related to the C_A^F is included in the dimensional model along with its related HL7 Hierarchies, whereas in step 2b, starting from the C_A^F, the CDA Backbone Hierarchy is recursively navigated in a target-source (i.e., child–parent) direction and each Act is included in the

	Step 2. Identify the Dimensions. Starting from the Act chosen and a Fact of the dimensional model, the CDA document is navigated in both parent–child and child–parent direction to include the relevant Act classes and HL7 Hierarchies to the data warehouse dimensional model
a	Each class C_A^T related to the class C_A^F as a target of an HL7 Act-Relationship (($R_{A-A}\left(C_A^F, C_A^T\right)$)) (Def. 6) is included in the dimensional model along with its related HL7 Hierarchies. • $\forall R_{A-A} \mid R_{A-A} = R\left(C_A^F, C_A^T\right) \rightarrow R_{A-A} \in R_{sf}$
b	Each Act class C_A^S included in the CDA Backbone Hierarchy $H_B^{'} = H_B(C_{A_i}, 1 \leq i \leq r-1)$ (Def. 8) and related to the class C_A^F as a source of an HL7 Act-Relationship (Def. 6), is included in the dimensional model. • $\forall C_{A_i} \in f_s\left(H_B^{'}\right) \mid \exists R_{A-A}\left(C_{A_i}, C_A^F\right) \rightarrow R_{A-A} \in R_{sf}$ Each HL7 Hierarchy (Def. 4) and each Act related to the class C_{A_i} (HL7 Act-Relationship, Def. 6) are included in the dimensional model. • $\forall C_{A_i} \in f_s(H_B^{'}), \forall H_{HL7} \mid \exists R_{A-H}\left(C_{A_i}, H_{HL7}\right) \rightarrow R_{A-H} \in R_{sf}$ • $\forall C_{A_i} \in f_s(H_B^{'}), \forall C_A^T \mid \exists R_{A-A}\left(C_{A_i}, C_A^T\right) \rightarrow R_{A-A} \in R_{sf}$

dimensional model. These classes provide additional information at low-level of detail and can be included in the model by a direct link with the Fact. Moreover, HL7 Hierarchies and target Acts related to the source class are included in the model to capture additional relevant information.

The second component to be studied is the HL7 Hierarchy. This is a particularly suitable element to represent a dimension given that it captures the functions of subjects and objects involved in a specific process, identifying, for instance, who performed it (i.e., performer), for whom it was done (i.e., subject), and where it was done (i.e., location). These data are captured through the attribute typeCode of the Participation class that specifies its meaning and purpose using a controlled vocabulary defined by HL7. *H_{HL7}* hierarchies related to C_A^F in an HL7 Act-Hierarchy Association ($R_{A-H}\left(C_A^F, H_{HL7}.C_P\right)$) are included in the model, as described in step 2c of the process shown in ►Table 1 and also proposed again in the following to simplify the readiness of the framework.

	Step 2. Identify the Dimensions
c	Each *H_{HL7}* HL7 Hierarchy related to the class C_A^F by an HL7 Act-Hierarchy association $R_{A-H}\left(C_A^F, H_{HL7}.C_P\right)$ is included in the dimensional model • $\forall H_{HL7} \mid \exists R_{A-H}\left(C_A^F, H_{HL7}.C_P\right) \rightarrow R_{A-H} \in R_{sf}$

►Table 3 summarizes examples of the different components of the CDA that can be used to identify a dimension of the schema, reporting the type and the name of the component as well as its description and the related Act class of the backbone. For instance, the recordTarget HL7 Hierarchy related to the Act ClinicalDocument describes the patient involved in the event. Steps 2a to c of the mapping process can be applied to the components reported in the portion of the CCD model to define a first draft of the snowflake schema as shown in ►Fig. 4. It is composed by a Fact denoted by the Act Observation surrounded by three dimensions: (1) ObservationRange to define specific thresholds of the laboratory test; (2) ServiceEvent related with the hierarchy <performer, assignedRole, assignedPerson> to define the main health care service described in the document as well as the health care provider who carries out the laboratory test; and (3) the hierarchy <recordTarget, patientRole, Patient> that specifies the patient involved in the event.

Moreover, there are attributes of the Fact that can be specifically used to define dimension keys in the fact class that is not related to a dimension table. These are called degenerate dimensions and are useful to group-related fact

Table 2 Example of Act classes that can be used to represent a fact table of the dimensional model

CDA class	Description	Example of processes	Measures
Act	General event that is being done, has been done, can be done, or is intended or requested to be done	Every type of health-related activity	Not applicable
Encounter	An interaction between a patient and health care participant(s) to provide service(s) or assessing the health status of a patient	Specialist and GP visits	lengthOfStayQuantity (quantity of time when the subject is expected to be or was resident at a facility as part of an encounter)
Observation	Action performed to determine an answer or a result value	Vital signs, clinical results in general, and also diagnoses, findings, and symptoms	Value (data determined by the observation) as well as interpretationCode (a qualitative interpretation of the observation)
Procedure	An event whose immediate and primary outcome (postcondition) is the alteration of the subject physical condition	Conservative procedures such as reduction of a luxated join, including physiotherapy such as chiropractic treatment	Not applicable
Substance administration	The act of introducing or otherwise applying a substance to the subject	Chemotherapy protocol; Drug prescription; Vaccination record	doseQuantity (amount of the therapeutic agent), as well as rateQuantity (the speed with which the substance is dispensed)

Abbreviation: CDA, Clinical Document Architecture.

Table 3 Example of suitable dimensions derived from the CDA components

HL7 component	Name	Example
Clinical document		
Hierarchy (Def. 4)	recordTarget	Patient involved in the event
	performer	Physician/Practitioner that performed the event
	responsibleParty	Participant with legal responsibility
	location	Health care facility where the event occurred
	participant	Other involved participants not mentioned by other classes
Act-Relationship (Def. 6)	ServiceEvent	The main Act such as a colonoscopy being documented
	EncompassingEncounter	Primary encounter in which the documented acts took place.
Section		
Hierarchy (Def. 4)	subject	Target of the entries recorded in the document
Clinical statement		
Hierarchy (Def. 4)	**performer**	Physician/Practitioner that performed the clinical event
	specimen	Part of entity typically the subject target of the observation
	participant	Other involved participants not mentioned by other classes
Observation		
Act-Relationship (Def. 6)	**ObservationRange**	Specifies a range of values for a particular observation
Substance administration		
Hierarchy (Def. 4)	consumable	Substance taken up or consumed as part of the administration

Abbreviations: CDA, Clinical Document Architecture; HL7, Health Level 7.
Note: Names in bold represent dimensions with one-to-many relationships.

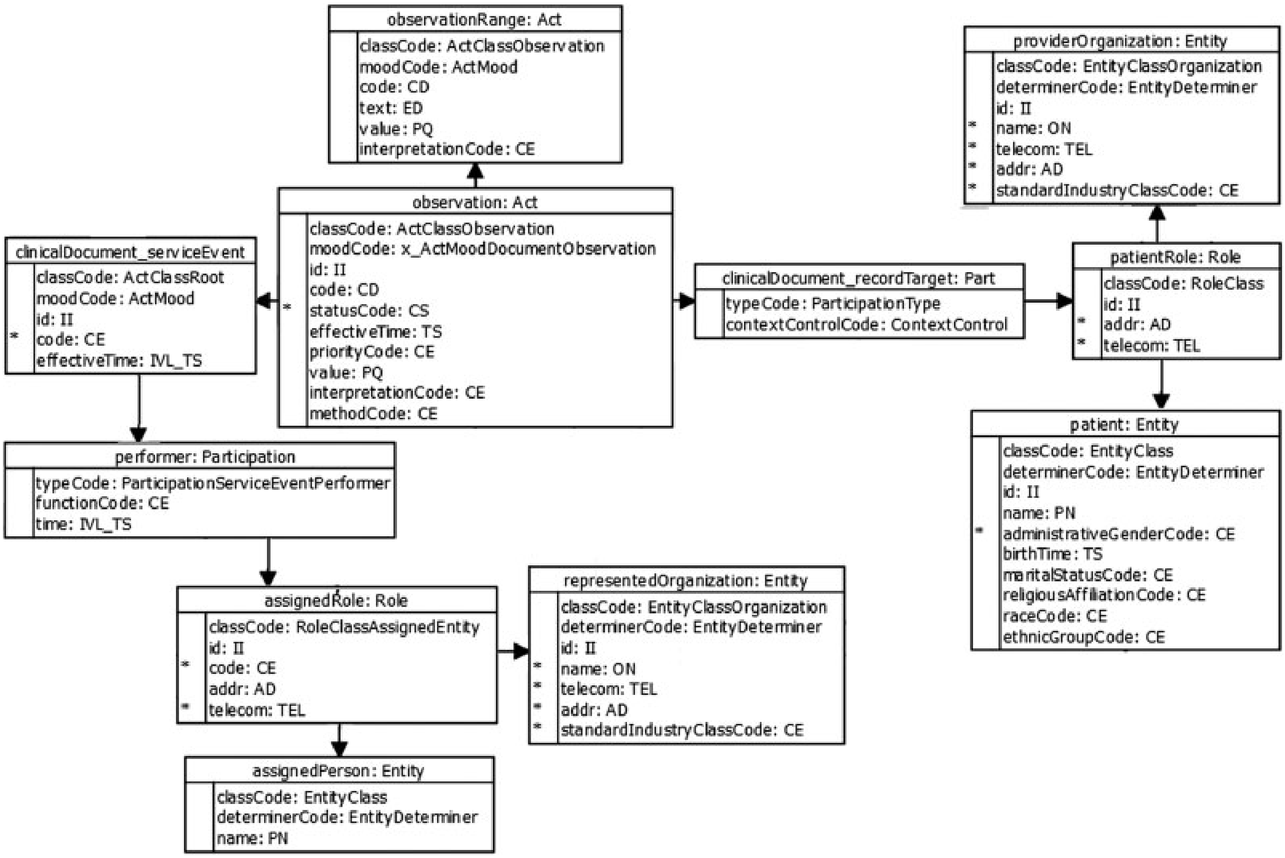


Fig. 4 First draft of the dimensional model designed on the basis of the Clinical Document Architecture (CDA). The snowflake schema is represented by the fact Observation surrounded by three dimensions: ObservationRange to specify the thresholds of the event observed; recordTarget and ServiceEvent to describe, respectively, the patient involved in the event and the main service provided. The class ServiceEvent is related with the hierarchy performer that specifies the provider who performed the test.

rows. A generic Act of the RIM contains several attributes that can be mapped in a degenerate dimension, such as the code that classifies the particular kind of Act and the statusCode that specifies the state of the Act (e.g., active, cancelled). Another important attribute of the Act class that can be modeled as a degenerate dimension is the effectiveTime that describes time/date when the event took place. However, this is often modeled into different classes of a hierarchy to analyze business process data over different dates or date ranges, such as weeks, week days, months, and individual days.

Refinement of the Dimensional Model

The initial conceptual schema designed on the basis of the CDA components, results in a high-level normalized data model making the operation of insert, update, and delete highly efficient also minimizing the size of the data stored. This representation is typically adopted in the design of databases intended for Online Transaction Processing characterized by a high volume of small transactions, such as updating an EHR. Conversely, a database designed for analytical purposes is characterized by a low volume of transactions often limited to insert information and accessing them using complex queries. For such models, the use of a denormalized schema facilitates business intelligence applications also improving the performance in data retrieval and aggregation.[27]

Denormalization of a hierarchy $H(\{C_1,...,C_n\})$ is performed by collapsing the attributes of the classes $\{C_2,...,C_n\}$ in the class C_1. For instance, the HL7 Hierarchy <subject, RelatedSubject, SubjectPerson> used to define the patient involved in a specific Observation is generally mapped in a single dimension class composed by all attributes of the HL7 Hierarchy classes. This approach can also be applied when the Role class is associated with two Entities called player and scoper as depicted in ►Fig. 4 where a PatientRole is linked to the patient (player) and to the providerOrganization (scoper) from which the patient will receive services. The results of this denormalization are shown in ►Fig. 5 where all the attributes of PatientRole, Patient, and ProviderOrganization are collapsed in a single class clinicalDocument_recordTarget.

However, health care business processes often require the adoption of many-to-many relationships to represent multiple records of a specific dimension associated with the fact table. For instance, when different practitioners deliver care to an individual over different distinct time intervals or when a specialist visit is performed due to multiple diagnosis. In these cases, the hierarchy cannot be fully denormalized and a bridge class should be used to model the many-to-many relationship between the fact and the hierarchy.[32] The result of this denormalization is depicted in ►Fig. 5 where all the attributes of AssignedPerson and RepresentedOrganization classes are collapsed in the assignedEntity class and the performer_bridge class is used to model the bridge class.

Another important step to be performed to refine the dimensional model is to resolve complex data types. In fact, several attributes of the CDA are coded using a complex data type that consists in a set of fields used to describe the value along with its properties. For instance, the attribute code of the class Observation of the CDA is described by different elements: a code value and a codeSystem identifier to specify an externally defined coding scheme. Moreover, the codeSystem may have a codeSystemName helping human interpretation. It may also have an optional displayName element containing the text that was originally written. A possible solution to represent a complex data type is to store each property in a single column of the relevant table excluding properties that are not needed for the business process analysis. For instance, a CD (Concept Descriptor) can be mapped using only two attributes: code and codeSystem to store the code of the event occurred and the system used to represent it. Moreover, different attributes of the RIM assume multiple values, such as the interpretationCode that specifies a set of rough qualitative interpretation of an Observation based on a HL7 nomenclature (e.g., "is decreased," "is below alert threshold," "is moderately susceptible"). These attributes can be modeled either creating a separate table related to the fact to store all the instances reported in the document or capturing only a single value in a specific attribute of the fact table, such as the first reported in the document.

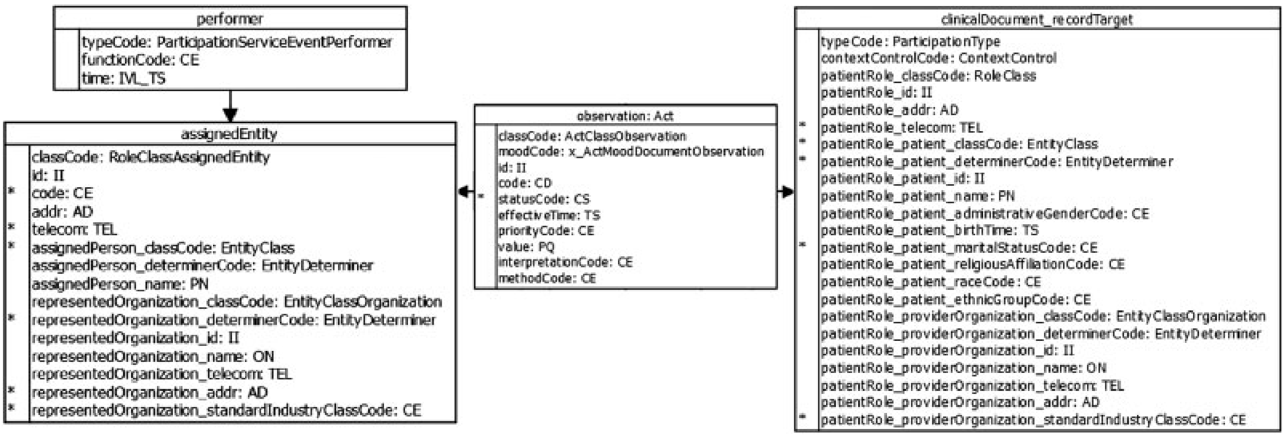

performer
typeCode: ParticipationServiceEventPerformer
functionCode: CE
time: IVL_TS
assignedEntity
classCode: RoleClassAssignedEntity
id: II
* code: CE
addr: AD
* telecom: TEL
* assignedPerson_classCode: EntityClass
assignedPerson_determinerCode: EntityDeterminer
assignedPerson_name: PN
representedOrganization_classCode: EntityClassOrganization
* representedOrganization_determinerCode: EntityDeterminer
representedOrganization_id: II
representedOrganization_name: ON
representedOrganization_telecom: TEL
* representedOrganization_addr: AD
* representedOrganization_standardIndustryClassCode: CE
observation: Act
classCode: ActClassObservation
moodCode: x_ActMoodDocumentObservation
id: II
code: CD
* statusCode: CS
effectiveTime: TS
priorityCode: CE
value: PQ
interpretationCode: CE
methodCode: CE
clinicalDocument_recordTarget
typeCode: ParticipationType
contextControlCode: ContextControl
patientRole_classCode: RoleClass
patientRole_id: II
patientRole_addr: AD
* patientRole_telecom: TEL
* patientRole_patient_classCode: EntityClass
* patientRole_patient_determinerCode: EntityDeterminer
patientRole_patient_id: II
patientRole_patient_name: PN
patientRole_patient_administrativeGenderCode: CE
patientRole_patient_birthTime: TS
* patientRole_patient_maritalStatusCode: CE
patientRole_patient_religiousAffiliationCode: CE
patientRole_patient_raceCode: CE
patientRole_patient_ethnicGroupCode: CE
patientRole_providerOrganization_classCode: EntityClassOrganization
patientRole_providerOrganization_determinerCode: EntityDeterminer
patientRole_providerOrganization_id: II
patientRole_providerOrganization_name: ON
patientRole_providerOrganization_telecom: TEL
patientRole_providerOrganization_addr: AD
* patientRole_providerOrganization_standardIndustryClassCode: CE

Fig. 5 Denormalization of a Health Level 7 (HL7) Hierarchy. The clinicalDocument_recordTarget hierarchy is denormalized by collapsing all attributes in the recordTarget participation class (right side). The performer hierarchy is denormalized by collapsing the attributes of the Entities in the Role assigneEntity (left side). The Participation class (performer) models a bridge to represent a many-to-many relationship with the Fact.

The last step to refine the dimensional model is to remove the unnecessary information by pruning and grafting it. Pruning a class implies to remove it together with its related classes creating a target in a relationship. All classes and attributes are excluded from the schema and therefore cannot be used to aggregate or to perform a query. For instance, pruning the class patientRole shown in ►Fig. 4 will also drop the classes patient and providerOrganization. Conversely, grafting a class means to delete it and to move the relationships with its targets to its sources. For instance, grafting the class patientRole moves the relationships with the classes patient and providerOrganization to the class clinicalDocument_recordTarget. The model can be further refined by removing attributes that are not of interest for the analysis to eliminate unnecessary level of detail.

ETL Process Implementation

Starting from the first-order logic description, the semiautomatic framework proposed in this article is shown in ►Fig. 6 highlighting two main subprocesses.

The first part of the conceptual framework concerns the generation of the XSLT document using a definition engine based on the node specified by the user to represent the fact of the dimensional model as well as the rules defined by the Kimball lifecycle. Moreover, the relevant CDA schema is considered to identify RIM stereotype of each element as well as the cardinality of each relationship, while the data type schema specifies the cardinality and the type of data of each attribute of a specific node. The XSLT definition engine is described in the next paragraph.

In the second part of the workflow, the XSLT document processes a CDA represented using the XML format to produce an output XML document (Transformation of the CDA) that is further managed and transformed to be mapped into a relational, object-relational or XML-native database (Store of the XML). In this perspective, different XML data warehouse architectures have been proposed in the literature to represent complex data as XML documents, such as XCube,[33] X-Warehousing,[34] and XML-OLAP,[35] to be physically integrated into an ODS and further analyzed using statistical and business intelligence methodologies. These representations converge toward a unified model that differ in the number of XML documents used to store facts and dimensions.[36] In this article, transformed XML documents are organized on the basis of X-Warehousing architecture, where each XML embeds the facts stored in the original CDA document as well as their related dimensions. This transformation is performed by a XSLT processor, such as the Open Source SAXON XSLT engine developed by Saxonica Limited (saxon.sourceforge.net).

Note that to comply with the privacy regulations, the original CDA document must be anonymized. However, this activity has not been discussed in the article given that it has to be applied to the CDA before applying the proposed conceptual framework.

Generation of the XSLT Document

►Fig. 7 reports the four main components (i.e., templates) of the XSLT document, each one identified by a specific pool using the Business Process Model and Notation.

In particular, it highlights the different activities to be executed to transform a CDA structured document in a XML document that is described in the following:

(1) Main: It finds all the nodes that match with the class chosen by the data warehouse designer to represent the fact table of the dimensional model (e.g., Observation). Starting from each node, it navigates the XML document in both directions: each ancestor is explored by the Examine Ancestor Node template, while each child is analyzed by the Examine Node template. The portion of the XSLT implementing this template is reported in the following:

```
<xsl:template match=”/”>
<model>
<xsl:for-each select=”//observation”>
<xsl:element name=”{name(.)}”>
<!–examines each child of the Fact node–>
<xsl:for-each select=”*”>
<xsl:call-template name=”examineNode”>
```

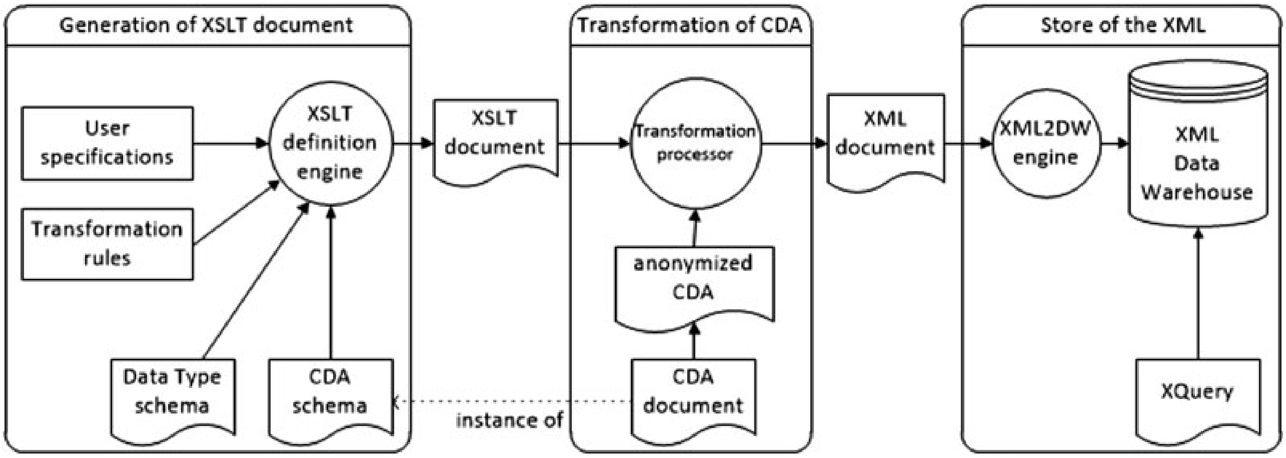


Fig. 6 Transformation process to load a Clinical Document Architecture (CDA) document in a Data Warehouse. The process concerns three main subprocesses: in the first one the eXtensible Stylesheet Language Transformations (XSLT) document is generated and used by the second subprocess to transform the CDA document to be loaded in the target data warehouse (Store of the XML).

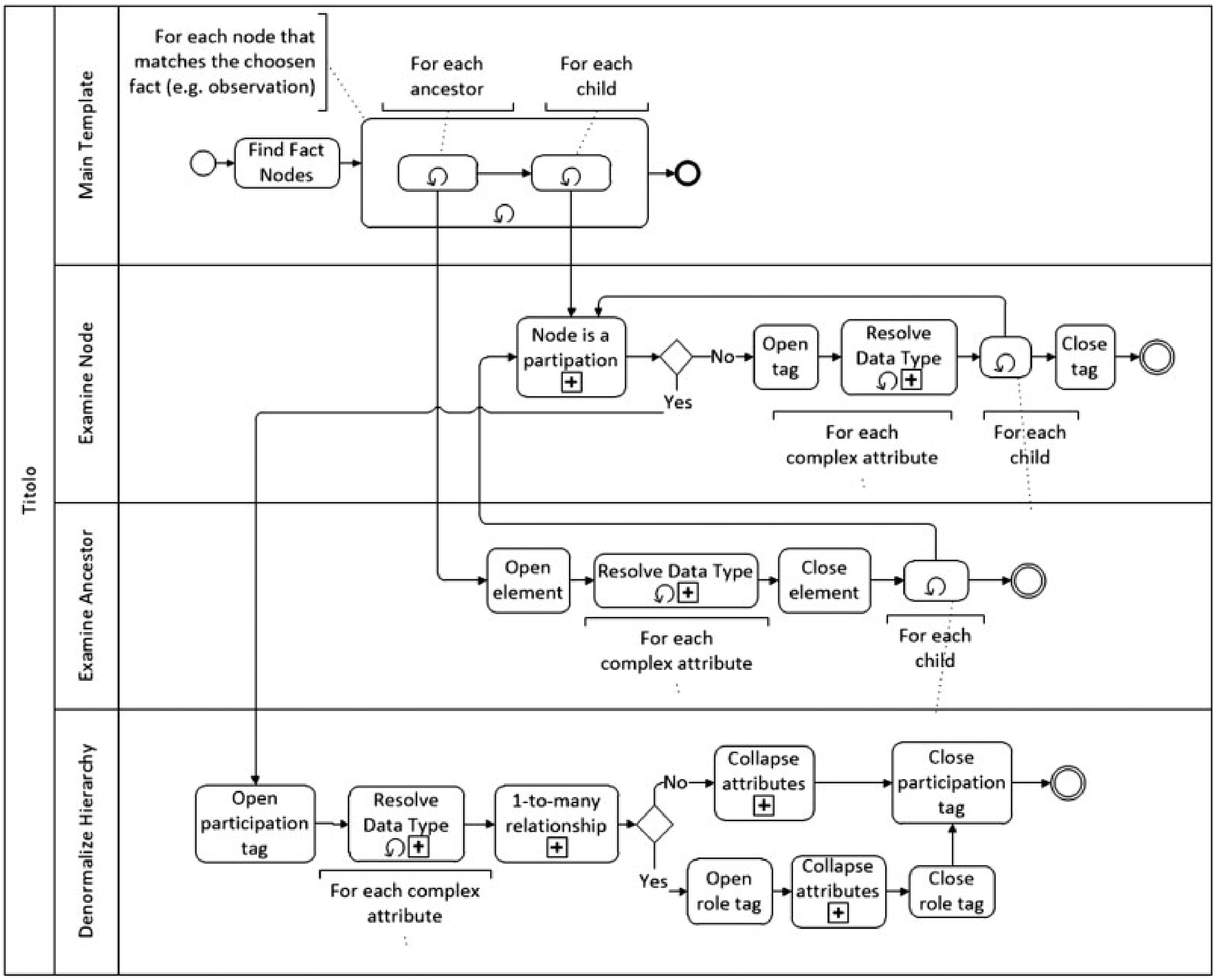


Fig. 7 Business process to generate the eXtensible Stylesheet Language Transformations (XSLT) document. It represents the main activities to be performed to define the XSLT document. In particular, each swim lane identifies a specific template of the transformation schema. Moreover, complex functions are shown using a rectangle with the plus sign against the bottom line.

```
<xsl:with-param name=”node” select=”.” />
</xsl:call-template>
</xsl:for-each>
<!–examines each ancestor of the Fact node–>
<xsl:variable name=”ancList” select=”ancestor-or-self::*” />
<xsl:for-each select=”$ancList”>
<xsl:if test=”parent::*”>
<xsl:call-template name=”examineAncestorNode”>
<xsl:with-param name=”node” select=”./.” />
<xsl:with-param name=”toRemove” select=”name(.)” />
</xsl:call-template>
</xsl:if>
</xsl:for-each>
</xsl:element>
</xsl:for-each>
</model>
</xsl:template>
```

(2) Examine Ancestor Node: It includes the node passed as input in the transformed document considering its resolved attributes. Moreover, each child is analyzed by the Examine Node template.

(3) Examine Node: It checks if the stereotype of the node received as input is a Participation. In this case, the node is passed to the Denormalize Hierarchy, otherwise it is included in the output document along with its resolved attributes. Moreover, each child is recursively analyzed by this template to be included in the output document. Once all children have been analyzed, the tag of the relevant node is closed.

(4) Denormalize Hierarchy: Starting from a participation node, the 4-ple <Participation, Role, Entity Player, Entity Scoper> is analyzed and a denormalized node is reported taking into account the multiplicity of the relationship between the participation and the act class. If the multiplicity is 1-to-1, the complex attributes of role and entity nodes are resolved by the Resolve Data Type function and collapsed in the output schema as children of the participation node using the function Collapse attributes. Otherwise if the relationship is 1-to-many, the hierarchy cannot be fully denormalized and a bridge

class is needed. To accomplish this task, the attributes of entity nodes are resolved and included in the schema as children of the role node. The portion of the XSLT implementing the template devoted to the denormalization of hierarchies is reported in the following:

```
<xsl:template name=”denormalizeHierarchy”>
<xsl:param name=”node” />
<xsl:param name=”multiplicity” />
<xsl:element name=”{name($node)}”>
<!–inserts all the attributes–>
<xsl:for-each select=”@*”>
<xsl:attribute name=”{name(.)}”>
<xsl:value-of select=”.”/>
</xsl:attribute>
</xsl:for-each>
<!–examines each child of the relevant node–>
<xsl:for-each select=”$node/*”>
<!–checks if the node is a Role–>
<xsl:if test=”my:isStereotype(., name(./parent::*), ’Role’)=’true’”>
<!–if is a 1-to-many relationship adds the role tag–>
<xsl:if test=”compare($multiplicity, ’true’)=0”>
<xsl:element name=”{name(.)}”>
<!–resolves and collapses Role and Entity attributes–>
<xsl:call-template name=”collapsingAttributes”>
<xsl:with-param name=”role” select=”.” />
</xsl:call-template>
</xsl:element>
</xsl:if>
<!–checks the multiplicity of the node–>
<xsl:if test=”compare($multiplicity, ’true’) != 0”>
<xsl:call-template name=”collapsingAttributes”>
<xsl:with-param name=”role” select=”.” />
</xsl:call-template>
</xsl:if>
</xsl:if>
<!–checks if the node is a Role–>
<xsl:if test=”my:isStereotype(., name(./parent::*), ’Role’) != ’true’”>
<xsl:call-template name=”resolveDataType”>
<xsl:with-param name=”node” select=”.” />
</xsl:call-template>
</xsl:if>
</xsl:for-each>
</xsl:element>
</xsl:template>
```

Moreover, the following functions have been implemented and used in the above-described templates, identified in ►Fig. 10 by a rectangle with the plus sign against the bottom line:

- Resolve Data Type: It analyzes a complex attribute and stores each property in a single column of the relevant table on the basis of the data type schema. However, attributes that assume multiple values (e.g., value of the Observation class) are modeled creating a bridge table to associate each attribute instance to the relevant node.
- Node is a participation: It checks if a relevant node belongs to a participation stereotype of the HL7 RIM on the basis of the CDA schema.
- 1-to-many relationship: It examines whether the multiplicity of the relationship between the relevant node and its father is 1-to-many on the basis of the CDA schema.
- Collapse attributes: Starting from the participation node, this function collects the attributes of both role and entity nodes and collapse them in a single node after resolving data types.

The result of this process is a XML document that can be subsequently pruned and grafted considering the specifications of the user with a particular attention on nodes considered unnecessary for the purpose of the business process analysis.

An Example of the Transformation of a CDA Document

In this article, the proposed approach is tested on a case study that analyses current and historically relevant vital signs of an individual. These data are collected in different specifications of the CDA schema produced by different organizations during different events, depending also on the national implementations. For instance, in Italy these data are stored and exchanged using the Report that collects results based on observations generated by laboratories and the Discharge letter that gathers information relative to the patient’s hospitalization. At international level, HL7 has released an implementation guide, the CCD,[25] to share patient clinical data specifying the structure and semantics of a patient summary clinical document. In this article, the attention will be focused on the vital signs section of the CCD that models individual’s clinical findings, such as blood pressure, heart rate, respiratory rate, height, weight, body mass index, head circumference, crown-to-rump length, and pulse oximetry.

For the purpose of our case study, we choose the class Observation as a fact of the dimensional model given that it describes an “action performed in order to determine an answer or a result value.” This is the starting point to transform the CDA document in a XML document to be loaded in the data warehouse as reported in the example depicted in ►Fig. 8, where the main template that implements the function to visit the XML tree is based on the proposed methodology. Navigating the tree in a child–parent direction each Observation node will include its ancestors with relevant attributes, such as organizer, section, and ClinicalDocument. Moreover, both children of the ClinicalDocument node (i.e., recordTarget and documentationOf) are included in the model as children of the Observation node, along with their children. Subsequently, the tree is parsed in a parent–child direction and the unique child of the Observation node (i.e., referenceRange) is included in the model. During these activities, each attribute is analyzed and resolved through the template Resolve Data Type taking into account the HL7 data type they belong to and also considering if they are multi- or single-valued attributes.

►Fig. 8 reports an example of the denormalization of a HL7 hierarchy highlighting the template devoted to this

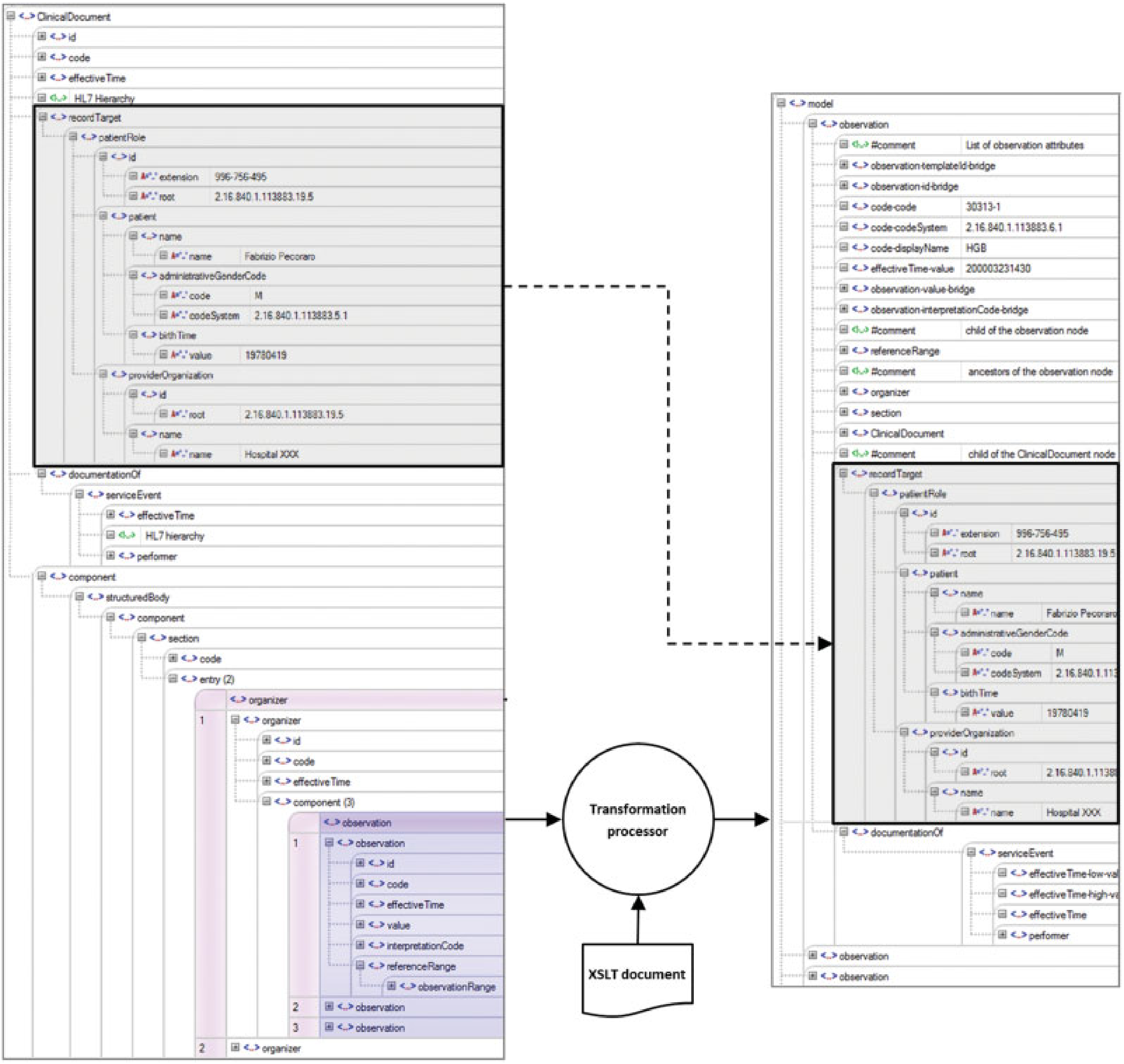


Fig. 8 Example of the Clinical Document Architecture (CDA) transformation in a dimensional model oriented XML document based on the eXtensible Stylesheet Language Transformations (XSLT) document. The denormalization of the recordTarget hierarchy is also highlighted.

activity. In particular the <recordTarget, patientRole, patient, providerOrganization> hierarchy is denormalized by collapsing all the attributes of the four classes in the <recordTarget> tag, given that it is related in a 1-to-1 association with the ClinicalDocument Act. Moreover, complex attributes are resolved taking into account the HL7 data type they belong to and also considering if they are multi- or single-valued attributes. For instance, the value of the attribute birth time of the patient Entity class is directly inserted in the recordTarget element. Differently, the identity of a patient requires a bridge element (<patientRole-id-bridge>) to associate it with the recordTarget considering that each patient can be identified using codes provided by different agencies such as the National Healthcare Service, insurance, and hospital.

Representing a Business Process Using Different CDA Specifications

The conceptual framework proposed in this article simplifies the design and implementation of ETL procedures given that information collected in the EHR are structured using a common schema based on the HL7 CDA that codes data using standard nomenclatures and dictionaries. However, given that each business process can be described using information collected in different documents, this task requires the analysis and the harmonization of the different CDA specifications moving the integration issue from a source system to a document template point of view. In particular, the integration of clinical information captured in different CDA implementations requires: (1) to disambiguate the meaning of the same concept that may change depending on the type of document where it is contained; and (2) to implement specific procedures to extract the same information from different types of documents.

In this section, similarities and differences in the representation of the main CDA concepts (e.g., attributes and classes to model information, cardinalities of relationships and attributes, vocabularies adopted to represent data) are highlighted taking into account the results obtained in the business process already introduced in the previous paragraph. In particular, information describing the chosen business process is collected in the CCD, in the Report that collects information on laboratory results, as well as in the Discharge letter that gathers information relative to the patient's hospitalization. These documents have been analyzed to develop the logical data map for the identification, for each document template, of the CDA component that specifies the dimensional model concept to be mapped. This step has to take into account: (1) the class that describes it; (2) attributes that map with the dimensional model columns; and (3) the path to reach the identified class. ►Table 4 shows the results of this mapping highlighting how the columns of each table of the data warehouse model are mapped with the attributes of the relevant CDA class.

At the body level, information describing the clinical event is organized in different perspectives using the Section, Organizer, and Act classes. For instance, the discharge letter structures data in different Sections, each one containing a set of interrelated Observations, while a report can have multiple self-associated Sections, each one further structured in different Organizers to subgroup the events performed during the health

Table 4 Logical data map highlighting how the data warehouse model primitives (target table and column) are mapped with the CDA concepts (source document, path, class, and attribute)

Target		Source			
Table	**Column**	**Document**	**Path**	**Class**	**Attribute**
Patient	ID	All	recordTarget.PatientRole	Patient	id.extension
	Age				birthTime.value
	Gender				genderCode.code
	ZIP				addr.postalCode
	city				addr.city
Facility	ID	All	ServiceEvent.performer.AssignedEntity	Organization	id.extension
	ZIP				addr.postalCode
	City				addr.city
Performer Bridge	Function	All	ServiceEvent.performer	Performer	functionCode
	Time				time
Performer	ID	All	ServiceEvent.performer.AssignedEntity	Person	id.extension
	ZIP				addr.postalCode
	City				addr.city
Laboratory Test	Value	Report	Section*.Organizer^	Observation	value.value
		Letter	Section		
		CCD	Section.Organizer		
	Unit	Report	Section*.Organizer^		value.unit
		Letter	Section		
		CCD	Section.Organizer		
	Date	Report	Section*.Organizer^		effectiveTime.value
		Letter	Section		
		CCD	Section.Organizer		
	minValue	Report	Section*.Organizer^.Observation	ObservRange	value.minValue
		Letter	Section.Observation		
		CCD	Section.Organizer		
	maxValue	Report	Section*.Organizer^.Observation		value.maxValue
		Letter	Section.Observation		
		CCD	Section.Organizer		

Abbreviation: CDA, Clinical Document Architecture.
Note: Symbol * specifies a recursive association, while ^ identifies an optional element.

care service provision. Such diversification requires the definition of different methods to extract the same information from each document template. This aspect is highlighted in the path column of the logical data map (see ►Table 4) that specifies how to navigate the XML document to extract the specific attribute collected in the relevant class. Conversely, the specifications of the two documents model the clinical event using an Act class of the CDA ClinicalStatement choice. In particular, both the report and the discharge letter describe a laboratory test through an Observation and the related ObservationRange: the former class is used to model both the measures of the fact table (value and unit) and the date degenerate dimension.

This similarity simplifies the design of transformation procedures as highlighted both in the class and attribute columns of the logical data map reported in ►Table 4. In terms of framework implementation, the same ETL can be adopted taking into account that the data warehouse model primitives are equally mapped with concepts of the different CDA specifications considered. However, this tool should take into account the path adopted by each CDA implementations to extract the same data and map it in the relevant attribute of the dimensional model.

Case Study

The methodology proposed in this article has been tested on a native XML data warehouse developed with BaseX using a set of 682 CCDs available on a public repository. The analysis is focused on the following indicators to assess the quality of care of patients with diabetes: top 10 diseases, distribution of patients per age group, and sex; number of comorbidities; proportion of patients who have performed at least three checks of hemoglobin glycated in the last year; proportion of patients who have the last-taken value of hemoglobin glycated under 7%; and incidence of diabetic retinopathy.

The dashboard reporting the overall result of this analysis is summarized in ►Fig. 9. In particular, indicators related to the hemoglobin glycated show a high level of compliance of patients with planned activities (70%) as well as an impact of care processes (66%). Differently, a high level of incidence of retinopathy (40%) reveals an unsatisfactory patient health status. To compile these indicators, documents are queried using the XQuery processor. An example of this query is presented in the BaseX screenshot shown in ►Fig. 10 where the indicator provides the proportion of patients who have the last-taken value of hemoglobin glycated under 7%.

Discussion

This article proposes a methodology to use the information collected in the EHR system for secondary purposes. This analysis has been performed from the architectural and data modeling perspectives.

From the architectural point of view, this article demonstrates the feasibility of EHR systems to develop an enterprise clinical data warehouse architecture in a clinical governance framework. To our knowledge, this is a novel approach that intends to exploit the entire EHR infrastructure to develop a

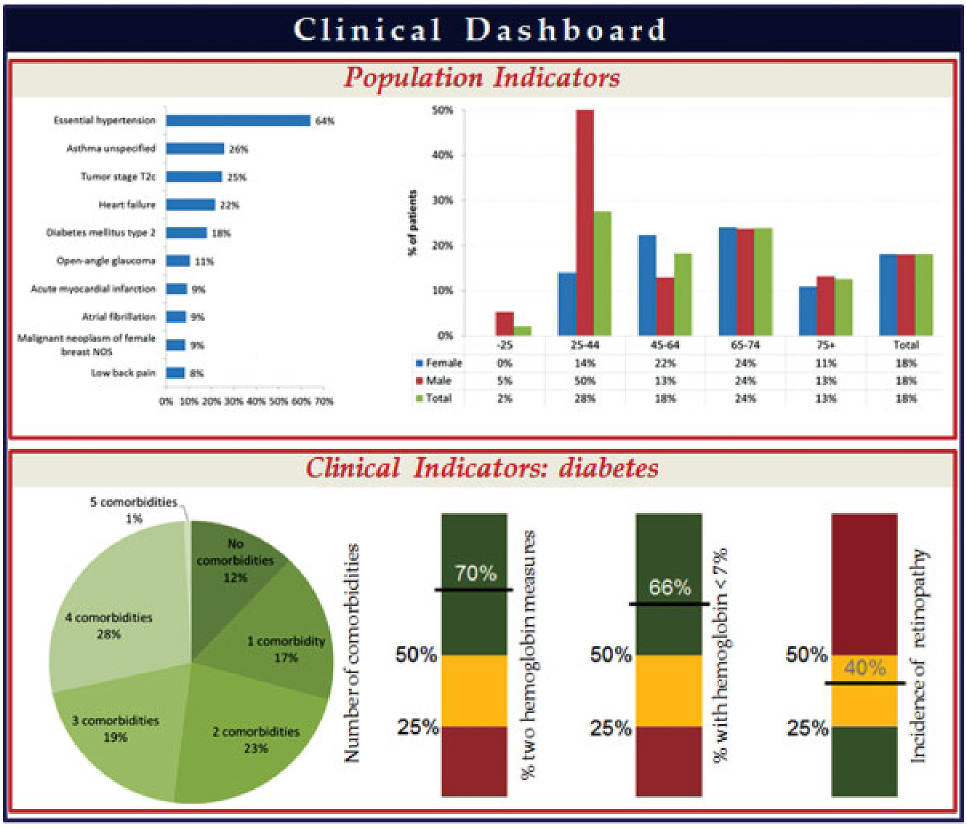
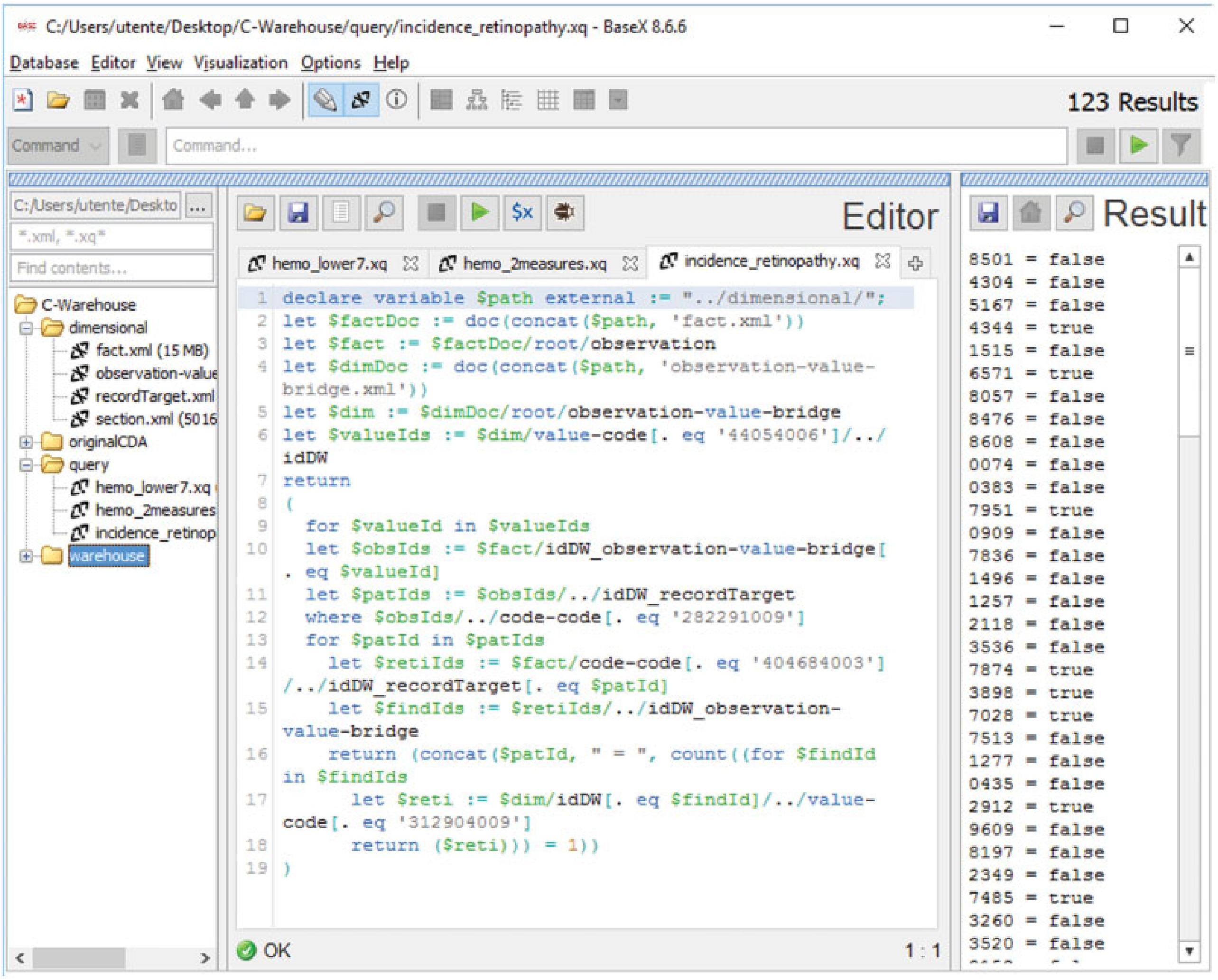


Fig. 9 Dashboard reporting the results of the computed indicators.



```
declare variable $path external := "../dimensional/";
let $factDoc := doc(concat($path, 'fact.xml'))
let $fact := $factDoc/root/observation
let $dimDoc := doc(concat($path, 'observation-value-bridge.xml'))
let $dim := $dimDoc/root/observation-value-bridge
let $valueIds := $dim/value-code[. eq '44054006']/../idDW
return
(
  for $valueId in $valueIds
  let $obsIds := $fact/idDW_observation-value-bridge[. eq $valueId]
  let $patIds := $obsIds/../idDW_recordTarget
  where $obsIds/../code-code[. eq '282291009']
  for $patId in $patIds
    let $retiIds := $fact/code-code[. eq '404684003']/../idDW_recordTarget[. eq $patId]
    let $findIds := $retiIds/../idDW_observation-value-bridge
    return (concat($patId, " = ", count((for $findId in $findIds
      let $reti := $dim/idDW[. eq $findId]/../value-code[. eq '312904009']
      return ($reti))) = 1))
)
```

```
8501 = false
4304 = false
5167 = false
4344 = true
1515 = false
6571 = true
8057 = false
8476 = false
8608 = false
0074 = false
0383 = false
7951 = true
0909 = false
7836 = false
1496 = false
1257 = false
2118 = false
3536 = false
7874 = true
3898 = true
7028 = true
7513 = false
1277 = false
0435 = false
2912 = true
9609 = false
8197 = false
2349 = false
7485 = true
3260 = false
3520 = false
```

Fig. 10 Screenshot of the BaseX environment reporting the structure of the data warehouse, the query to identify patients who have the last value of hemoglobin glycated under 7%, and the relevant result.

Business Intelligence tool that supports the evaluation of health care activities under different points of view. In fact, the implementation of tools such as clinical dashboards represents an opportunity to increase not only effectiveness, efficiency, and quality of health care services, but also the transparency of economic and clinical activities as well as the availability of real-time information to decision makers.[37] The use of the EHR as an ODS represents the core of the proposed architecture that guarantees the separation between the transactional and the analytical processing entailing different advantages. First, it ensures data quality using data and documents already integrated in a health infrastructure. Information are provided using a publish–subscribe paradigm that guarantees health data to be promptly exchanged at the moment an event is published, with an automatic detection of relevant information directly from source systems. This ensures a timely and continuously updated information flow as well as data integrity and consistency based on a sample size that covers a broad target population.

Moreover, this approach facilitates the gradual integration of other applications within the data warehouse architecture given that rules and parsing procedures are standardized in a common framework. This can provide additional values also in different health care-related sectors including education, clinical research, public health, security, and policy support.[38] For instance, the integration of clinical information with community data collected in geographical or sociodemographic information systems can maximize the potential of secondary use of EHR data to study the impact of prevention and other significant public health issues in the management of a specific disease over the territory.

Finally, the use of EHR as a source of information in a data warehouse architecture makes it easier to design and develop ETL tools given that clinical information provided by source systems is stored in a unique infrastructure and structured using standardized documents independently from the features of the source legacy systems. In particular, the semantic interoperability is ensured by using the HL7 CDA standard developed to exchange information between the source information systems and the EHR infrastructure based on common vocabularies.

From the data modeling point of view, HL7 CDA has been the basis to develop a conceptual framework to design a data warehouse dimensional model based on the CDA concepts.

The use of standardized source on information such as CDA documents has the advantage of simplifying the ETL procedures, considering that the involved information systems already share a common schema and has to implement a set of transformation tools to improve interoperability. This approach simplifies the identification of the dimensional model primitives that describe the business process to be analyzed and makes it easier to implement transformation tools to load HL7 CDA messages in the data warehouse. This reduces the resources to be invested to implement the ETL tool that is considered the most time-consuming and expensive activity in the data warehouse developing process.[39]

However, given that in the EHR information describing the same business process can be extracted from different document templates, it is important to take into account the structure of each document in representing the relevant business process. To accomplish this task, the proposed conceptual framework integrates data captured in different CDA implementations applying specific procedures to extract the same information from different types of document considering that the meaning of the same concept may change depending on the type of document where it is contained.

Conclusion

The use of CDA collected in a structured EHR to define a data warehouse architecture is justified by the availability of a shared repository of documents that represent the complete lifetime medical history of the subject of care provided by various health professionals. The data extracted from this type of systems provide useful information for clinical governance purposes as well as for scientific and epidemiological research.[40]

To analyze data from a statistical point of view, they have to be extracted from EHR documents using an ETL tool that transforms them to be loaded in a specialized data warehouse and then processed for secondary purposes. To facilitate these transformation procedures, a semiautomatic procedure was defined mapping the primitives of the data warehouse dimensional model with the HL7 CDA classes. The feasibility of this approach was demonstrated implementing the conceptual framework, as shown in the example provided. Moreover, the conceptual framework has been designed to be also applied to every type of repository based on CDA documents.

This semiautomatic procedure will be tested on a wider set of clinical documents based on different CDA specifications (e.g., discharge report forms, prescription of pharmacological products and specialist visits, patient summary) with the aim of developing a dashboard to assess the quality of health care services provided in the framework of continuity of care.

What was already known on the topic?

- The use of clinical and administrative data for secondary purposes is considered an important challenge to be accomplished to support clinical research and decision support.
- This makes it necessary to integrate data provided by standalone heterogeneous information systems developed using different technologies and that are not implemented to be interoperable with each other.
- Different recent experiences have recognized the successful use of electronic health record (EHR) systems as the basis to model, transform, and store EHR data to create a data warehouse environment to improve the use of clinical data for secondary purposes.

What does this study add to our knowledge?

- The studies reported in the literature devoted to implement data warehouses based on EHR system are generally limited to a single institution or provider and/or on a specific target population. In our approach, we consider a longitudinal EHR (L-EHR) system as an infrastructure that provides a more comprehensive description of the patient's health status with a complete lifetime medical history described by the different types of clinical documents generated by different providers and available across multiple health care organizations.
- To integrate L-EHR data in a data warehouse, a conceptual framework has been designed and implemented to facilitate the information flows. L-EHR data are semiautomatically extracted from the CDA collected in the L-EHR and transformed to be loaded in the target data warehouse system.

Highlights

- A data warehouse architecture is proposed based on the longitudinal electronic health record (L-EHR) as an operational data store.
- An extract, transform, and load (ETL) tool was designed and developed to extract information from the L-EHR and transform them to be loaded in the target data warehouse.
- The transformation procedures are based on a semiautomatic framework that maps the primitives of the HL7 (Health Level 7) CDA (Clinical Document Architecture) schema with the concepts of the dimensional model.
- The framework has been formalized based on first-order logic, while from the implementation perspective a source code is proposed on the basis of the eXtensible Stylesheet Language Transformations (XSLT) standard.

Clinical Relevance Statement

Data warehouse architecture is proposed based on the electronic health record (EHR) as an operational data store. An extract, transform, and load (ETL) tool was designed and developed to extract data from EHR and transform them to be loaded in the target data warehouse. The transformation procedures are based on a semiautomatic framework that maps the primitives of the HL7 (Health Level 7) CDA (Clinical Document Architecture) schema with the concepts of the dimensional model.

Authors' Contributions

All authors contributed in the conception of the study as well as in the design of the conceptual framework that was subsequently implemented by F.P. All authors contributed equally in drafting, critically revising, and writing the final version of the article.

Funding

None.

Conflict of Interest

None declared.

References

1 Safran C, Bloomrosen M, Hammond WE, et al; Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007;14(01):1–9

2 Rea S, Pathak J, Savova G, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. J Biomed Inform 2012;45(04):763–771

3 Danciu I, Cowan JD, Basford M, et al. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform 2014;52:28–35

4 Diomidous M, Zimeras S, Mantas J. Spatial electronic health record for the epidemiological clinical data. Travel Health Informat Telehealth 2009;1:66–72

5 Abhyankar S, Demner-Fushman D, McDonald CJ. Standardizing clinical laboratory data for secondary use. J Biomed Inform 2012; 45(04):642–650

6 Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc 2009;16(03):328–337

7 Roque FS, Jensen PB, Schmock H, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLOS Comput Biol 2011;7(08):e1002141

8 Lurio J, Morrison FP, Pichardo M, et al. Using electronic health record alerts to provide public health situational awareness to clinicians. J Am Med Inform Assoc 2010;17(02):217–219

9 Zhou X, Chen S, Liu B, et al. Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Artif Intell Med 2010;48(2-3):139–152

10 Sahama TR, Croll PR. A data warehouse architecture for clinical data warehousing. Proceedings of the fifth Australasian Symposium on ACSW Frontiers, Ballarat, Victoria, Australia; 2007:227–232

11 de Mul M, Alons P, van der Velde P, Konings I, Bakker J, Hazelzet J. Development of a clinical data warehouse from an intensive care clinical information system. Comput Methods Programs Biomed 2012;105(01):22–30

12 Stow PJ, Hart GK, Higlett T, et al; ANZICS Database Management Committee. Development and implementation of a high-quality clinical database: the Australian and New Zealand Intensive Care Society Adult Patient Database. J Crit Care 2006;21(02):133–141

13 Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. Radiother Oncol 2013;108(01): 174–179

14 Pecoraro F, Luzi D, Ricci FL. Secondary uses of EHR systems: a feasibility study. Proceedings of IEEE International Conference on E-Health and Bioengineering (EHB), Iasi, Romania; 2013:1–6

15 Selby JV, Krumholz HM, Kuntz RE, Collins FS. Network news: powering clinical research. Sci Transl Med 2013;5(182):182fs13

16 Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61–81

17 Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med 2009;48(01):38–44

18 Hu H, Correll M, Kvecher L, et al. DW4TR: a data warehouse for translational research. J Biomed Inform 2011;44(06):1004–1019

19 Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005; 330(7494):765

20 Schubart JR, Einbinder JS. Evaluation of a data warehouse in an academic health sciences center. Int J Med Inform 2000;60(03): 319–333

21 Rubin DL, Desser TS. A data warehouse for integrating radiologic and pathologic data. J Am Coll Radiol 2008;5(03):210–217

22 Watson HJ, Goodhue DL, Wixom BH. The benefits of data warehousing: why some organizations realize exceptional payoffs. Inf Manage 2002;39:491–502

23 Serbanati LD, Contenti M, Mercurio G, Ricci FL. LUMIR: a region-wide virtual longitudinal EHR. Proceedings of the 9th International HL7 Interoperability Conference (IHIC), Crete, Greece; 2008

24 Roth MT, Schwarz P. Don't scrap it, wrap it! A wrapper architecture for legacy data sources. Proceeding of the Conference on Very Large Data Bases (VLDB), Athens, Greece; 1997:266–275

25 HL7 Implementation GuideCDA Release 2–Continuity of Care Document (CCD). Ann Arbor, MI: Health Level Seven, Inc.; 2007

26 Lupse O, Vida O, Stoicu-Tivadar L, Stoicu-Tivadar V. Using HL7 CDA and CCD standards to improve communication between healthcare information systems. IEEE 9th International Symposium on Intelligent Systems and Informatics (SISY); 2011:453–457

27 Kimball R, Ross M. The Data Warehouse Toolkit Second Edition: The Complete Guide to Dimensional Modelling. New York: Wiley Computing Publishing; 2002

28 Schadow G, Mead CN, Walker DM. The HL7 Reference Information Model under scrutiny. Stud Health Technol Inform 2006;124:151–156

29 Luzi D, Pecoraro F, Ricci FL, Mercurio G. A medical device Domain Analysis Model based on HL7 Reference Information Model. Proceeding of Medical Informatics in a United and Healthy Europe (MIE); 2009:162–166

30 Inmon WH. Building the Data Warehouse. 4th ed. New York: John Wiley & Sons; 2005

31 Inmon WH, Zachman JA, Geiger JG. Data Stores, Data Warehousing, and the Zachman Framework: Managing Enterprise Knowledge. New York: McGraw-Hill; 1997

32 Eggebraaten TJ, Tenner JW, Dubbels JC. A health-care data model based on the HL7 Reference Information Model. IBM J Res Develop 2006;46:5–18

33 Hümmer W, Bauer A, Harde G. XCube: XML for data warehouses. Proceedings of the 6th ACM International Workshop on Data Warehousing and OLAP; 2003:33–40

34 Boussaid O, Messaoud RB, Choquet R, Anthoard S. X-warehousing: an XML-based approach for warehousing complex data. Advances in Databases and Information Systems. Berlin, Germany: Springer; 2006:39–54

35 Park BK, Han H, Song IY. XML-OLAP: a multidimensional analysis framework for XML warehouses. Data Warehousing and Knowledge Discovery. Berlin, Germany: Springer; 2005:32–42

36 Mahboubi H, Ralaivao JC, Loudcher S, Boussaïd O, Bentayeb F. X-wacoda: an XML-based approach for warehousing and analyzing complex data. Data Warehousing Design and Advanced Engineering Applications: Methods for Complex Construction. Pennsylvania, United States: IGI Global; 2009:38–54

37 Metller T, Rohner P. Supplier relationship management: a case study in the context of health care. J Theor Appl Electron Commer Res 2009;4:58–71

38 Committee on Data Standards for Patient Safety. Key Capabilities of an Electronic Health Record System. Institute of Medicine Report, 5; 2003

39 Kimball R, Caserta J. The Data Warehouse ETL Toolkit. Indianapolis: Wiley; 2006

40 Serbanati LD, Ricci FL. EHR-centric integration of Health Information Systems. E-Health and Bioengineering Conference (EHB). Iaşi, Romania 2013:1–4