Journal of King Saud University – Computer and Information Sciences (2016) 28, 311–329

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

An encoding methodology for medical knowledge using SNOMED CT ontology

Shaker El-Sappagh [a], Mohammed Elmogy [b,*]

[a] Faculty of Computers and Information, Minia University, Egypt
[b] Faculty of Computers and Information, Mansoura University, Egypt

Received 26 May 2015; revised 11 September 2015; accepted 9 October 2015
Available online 31 October 2015

KEYWORDS

Clinical decision support system (CDSS);
SNOMED CT (SCT) coding;
Semantic data retrieval;
Ontology;
Case-based reasoning (CBR);
Diabetes diagnosis

Abstract Knowledge-Intensive Case Based Reasoning (KI-CBR) systems mainly depend on ontology. Using ontology as domain knowledge supports the implementation of semantically-intelligent case retrieval algorithms. The case-based knowledge must be encoded with the same concepts of the domain ontology. Standard medical ontologies, such as SNOMED CT (SCT), can play the role of domain ontology to enhance case representation and retrieval. This study has three stages. First, we propose an encoding methodology using SCT. Second, this methodology is used to encode the case-based knowledge. Third, all the used SCT concepts are collected in a reference set, and an OWL2 ontology of 550 pre-coordinated concepts is proposed. A diabetes diagnosis is chosen as a case study of our proposed framework. SCT is used to provide a pre-coordination concept coverage of ~75% for diabetes diagnosis terms. Whereas, the uncovered concepts in SCT are proposed. The resulting OWL2 ontology will be used as domain knowledge representation in diabetes diagnosis CBR systems. The proposed framework is tested by using 60 real cases.

1. Introduction

Despite many studies aimed to improve the effectiveness of clinical decision support system (CDSS), several obstacles remain in applying it at the point of care. Case-based reasoning (CBR) is an artificial intelligence (AI) methodology for ill-formed and experience-based CDSS. The most important component of CBR system is the case-based that contains the previous experience. Building this knowledge is considered a challenging task. Electronic Health Record (EHR) is a complete source for cases in CBR (Branden et al., 2011) as it contains raw data for daily transactions. However, medical data are usually inappropriate for direct use in CBR. The data need some steps to be converted to CBR knowledge. Data preprocessing, encoding, and fuzzification are the most important steps for that task. The execution of these steps can convert EHR data to case-based knowledge.

This paper concentrates on the second step that standardizes or encodes the medical data using SNOMED CT (SCT) terminology. Standardization of case-based contents is critical for CBR systems because lack of standard knowledge impedes the CBR implementation (Ahmadian et al., 2011; González

* Corresponding author.
E-mail address: melmogy@mans.edu.eg (M. Elmogy).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2015.10.002

et al., 2013; Melton et al., 2006). It supports the seamless integration between CBR system and EHR system, the implementation of semantically intelligent case retrieval algorithms, and the building of complete case-bases from interoperable distributed EHR systems.

The encoding process consists of four main stages. The first stage is to collect all the patient encountered information. The second stage is to identify relevant concepts referring to diagnoses and procedures. The third stage is to map these concepts to a standard terminology. Finally, the fourth stage is to apply the encoding for case-based data using the previously determined concepts IDs.

EHR data and data models can be structured, semi-structured, or unstructured. Structured data storage, entry, and communication are the current trend in all implementations of the EHR systems (Kim and Park, 2012). The structured data facilitate the coding process because it is easier to find the mapping between structured data and terminology codes compared with raw or unstructured text. For semantic retrieval purpose in CBR system and the semantic interoperability of collected data from distributed EHR systems, a standard terminology is needed to encode case-based knowledge (Højen and Gøeg, 2012). The usage of unique identifiers and the conceptual structure for each concept in a medical terminology system allow an unambiguous interpretation of the concept meaning across systems. Standard medical terminologies play a significant role in health care by supporting recording, retrieval, and analysis of patient information (Lee et al., 2010).

To create a case-based diabetes diagnoses system based on EHR data, we need to perform the following steps in order:

1. Collect the EHR medical data from distributed systems and store them in a standard data model. El-Sappagh et al. (2015) proposed a standard relational data model for case-based diabetes diagnosis based on HL7 RIM.[1] They populated it with a set of 60 diabetic patient cases.
2. Perform data preparation steps to the collected data to enhance its quality and prepare it for CBR case-based structure. El-Sappagh et al. (2014) applied a set of machine learning algorithms on the collected data including feature selection, normalization, summarization, weighing, and discretization to process the tested cases.
3. Standardize the semantics of the resulting database contents by coding it with SCT terminology. This phase is the focus of the current study.
4. Make fuzzification for diabetes vague attributes. This phase is out of the scope.

Using ontologies in CBR systems facilitates the creation of KI-CBR that considerably reduces the knowledge acquisition bottleneck (Dendani et al., 2012). Formalization of ontologies is useful in CBR community for many reasons as listed below (Abou Assali et al., 2009; Dendani et al., 2012; El-Sappagh et al., 2014b):

1. As a place for *persistence storage of cases:* Individuals or concepts are used, which are embedded in the case-base ontology.
2. As a vocabulary to define *the case structure*: Cases can be embedded as individuals in the ontology itself, or they can be stored in a different persistence storage media as databases.
3. As a *terminology* to define the query vocabulary: A user can express his requirements better if he can use a richer vocabulary to define the query. During the similarity calculation, ontologies bridge the gap between the query terminology and the case-base terminology.
4. For *case retrieval* and similarity, for *case adaptation*, and for *case learning*.
5. For *reuse* of case knowledge between different CBR systems.
6. To facilitate the *integration* of CDSS in the EHR environment at the physician's point of care.
7. To facilitate the building of *distributed CBR systems* in distributed healthcare environments.
8. To facilitate the *collection of cases* from different EHR environments using a unified language.

The overarching aim of this study is to propose a data encoding methodology, apply it to diabetes diagnosis dataset, propose a domain ontology based on SCT, and implement a KI-CBR system. This CBR system will diagnose diabetes and determine the probability of having other complications in kidney and liver. For space restrictions, the last goal will be considered in future work. Fig. 1 shows the two main types of KI-CBR architectures. We focus on the type (a) where the case-based is a regular database, and some of the patient features are instances of domain ontology concepts. In type (b), cases are stored as instances in a case-based ontology.

The case-based features are either primitive or unstructured raw text. Primitive types are not encoded, and it has data types as numbers and ordinal. Raw text data need to be coded as concept IDs for SCT concepts. This coding facilitates the computation of clinical or semantic distances between medical concepts. Clinical distance measures the medical closeness between concepts in a medical ontology. Semantic similarity metrics quantifies similarity in meaning between two concepts (Melton et al., 2006). The semantic or clinical distances between concepts are used to calculate the enter-patient distance between two cases. The clinical distance is more accurate than the semantic one (Melton et al., 2006). To explain the idea, we will provide the following example.

As shown in Fig. 2, if we compare the two patients P1 and P2 with diseases D1 = "kidney disease" and D2 = "kidney disease", then the semantic similarity $Sim_{Semantic}$ (D1, D2) = 1. Another example, if D1 = "medullary cystic disease OS" and D2 = "medullary cystic disease OS" then $Sim_{Semantic}$ (D1, D2) = 1. However, the two patients in the second example are more similar than the first example because the clinical similarity $Sim_{Clinical}$ ("*kidney disease*","*kidney disease*") < $Sim_{Clinical}$ ("*medullary cystic disease OS*","*medullary cystic disease OS*"). The main reason is that: both patients with "*kidney disease*" have, from a semantic perspective, the same concept, and therefore the semantic distance is zero. When applying these concepts to the patient case, "*kidney disease*" could mean many other disease entities, including "*Medullary sponge kidney*", "*medullary cystic disease OS*", "*caliectasis*", "*amyloid nephropathy*", "*hypertensive renal disease*", and others. On the other hand, "*medullary cystic disease OS*" refers to the same particular disease entity. By comparing these

[1] http://www.hl7.org/implement/standards/rim.cfm.

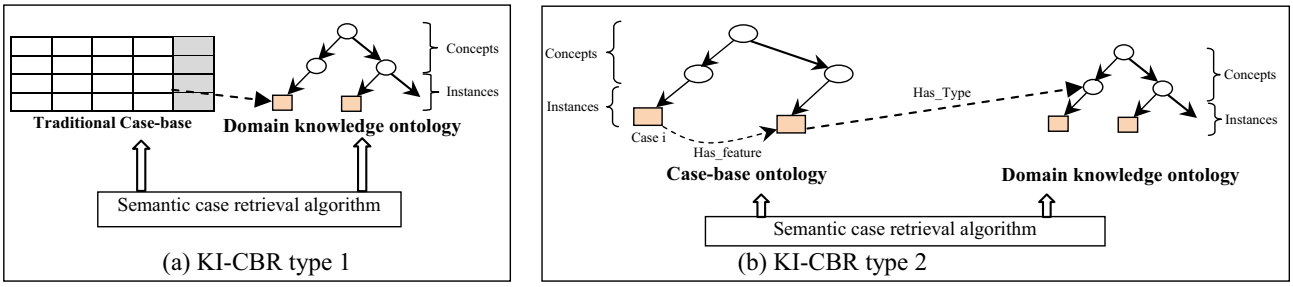


Figure 1 The main types of KI-CBR frameworks.

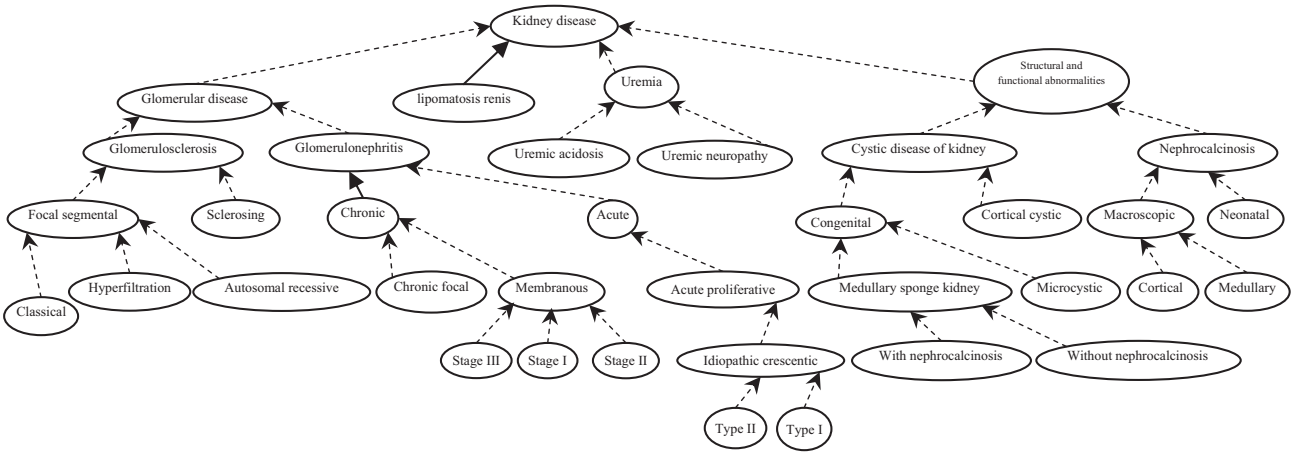


Figure 2 A sub-ontology from SCT for kidney disease.

similarities with the lexical similarity, the semantic and clinical similarity algorithms have higher accuracies (Harispe et al., 2014). For example $Sim_{Lexical}$ (“Medullary sponge kidney”, “Microcystic”) = 0, but their $Sim_{Clinical}$ and $Sim_{Semantic} \neq 0$, which is more accurate.

Case-based encoding based on SCT can use concept’s Fully Specified Name (FSN) or concept ID to encode this concept. Although the usage of FSN makes the case-based contents readable, the similarity will depend on the lexical similarity between FSNs. For improving similarity, we encode the case-based contents with concept-IDs. Concept-ID in SCT is the same as the primary key in the database. As a result, we can manage all descriptions and relationships of any concept. We expect that this decision will enhance the similarity calculations in CBR environment. The lexical similarity is not intelligent compared with the previously compared similarities. To achieve this level of semantic similarity, case-base must be encoded using an ontology. To get the optimum solution, we use the most popular and standard ontology (i.e., SCT).

The rest of this paper is organized as follows. Section 2 discusses the related work of SCT encoding problems and roles of ontology in CBR. Section 3 is the preliminaries and materials including a brief description of our dataset, the encoding options of EHR data, the matching algorithms, the normalization steps, and the encoding rules. Section 4 is the method used in this study including the proposed case-based preparation framework and the proposed case-based encoding methodology. Section 5 provides the results of our encoding process. Finally, conclusion and future work are discussed in Section 6.

2. Related work

Case retrieval is the most critical phase in CBR system (Lu et al., 2013). It depends totally on case-based structure and content. Calculating the similarity between two patients conditions based on simple string matching between clinical terms is insufficient (Harispe et al., 2014). On the other hand, the calculation of the similarity between ontology concepts by the clinical distance or the semantic distance increases the intelligence of the CBR system (Köhler et al., 2009; Melton et al., 2006). Intelligent case retrieval depends on the existence of a domain ontology and an encoded case-based (Zidi et al., 2014). The coding of data supports the CBR semantic case retrieval in many forms. For instance, concepts can be represented in different levels of granularity or abstraction (e.g., in SCT terminology, Type II diabetes mellitus can be represented as *Type II diabetes mellitus with neuropathic arthropathy IS_A Type II diabetes mellitus with arthropathy IS_A diabetes mellitus Type 2 IS_A diabetes mellitus*). This hierarchical knowledge enriches user queries. Moreover, the retrieval algorithm can handle different descriptions of the same concept (e.g., Myocardial infarction ≡ Heart attack ≡ Cardiac infarction). As a result, CBR system can think like domain expert, and it can integrate data from different systems like hospitals, doctors’ offices, and outpatient departments.

Subirats and Ceccaroni (2011) made encoding of a dataset for rehabilitation CDSS using international classification of functioning, disability and health (ICF), SCT, and ICD. However, their work is away from CBR, and their proposed

ontology is fragile. It contains only 77 concepts from 3 standard ontologies. Bichindaritz (2004) concluded that using standard ontologies with CBR enhances sharing and distribution of case knowledge. SCT is a complete clinical terminology in the world (Kooij et al., 2006). It contains more than 388,000 active concepts organized in 19 hierarchies, 1.14 million descriptions, and 1.38 million relationships (IHTSDO, 2015a).

Wasserman and Wang (2003) and Silva et al. (2011) concluded that SCT is the most suitable ontology for coding of problem lists and diagnosis. Moreover, SCT is better than ICD for encoding of diagnosis data because ICD is centralized on the classification of diseases. In 2005, Canada Health Infoway recommended the SCT as the terminology for coding patient data as part of the interoperable EHR (iEHR) initiative. Moreover, the US government and European Commission have suggested the usage of SCT as the standard clinical terminology (IHTSDO, 2015). As a result, SCT can be used in HER and CDSS (Kooij et al., 2006; Rasmussen and Rosenbeck, 2011).

However, SCT has many problems as redundancy, inconsistency, and improper ontological representation. Despite these limitations, a way to use SCT for coding EHR data must be discovered. There are few studies that describe how SCT is implemented in clinical settings, and they focused mostly on data capture, data retrieval, and decision support (Kim and Park, 2012; Liu et al., 2010). There are fewer details on how SCT concepts are stored and what methods are used to facilitate retrieval and decision support (Lee et al., 2013). As a result, implementing SCT is still relatively new and is a challenging proposition. There is still much work ahead to bring SCT into routine clinical use. Many studies propose mapping guidelines as a way to ensure consistent mapping procedures (Højen and Gøeg, 2012). Moreover, Lee et al. (2010) proposed an encoding methodology for EHR data by SCT.

Chiang et al. (2006) concluded that the reliability of SCT coding is imperfect, and may be a function of browsing methodology. The solution has two branches. *The first one* is the implementation of accurate SCT browsing software (Chiang et al., 2006). There are many browsing tools as CliniClue xplore, bioportal, and others (IHTSDO, 2015b). *The second one* is to create an efficient encoding scheme (Ryan et al., 2007). The existing methodologies for mapping clinical text in EHR to SCT concepts range from manual to semi-automatic and automatic methods (Barrett et al., 2012; Lamy et al., 2013; Lee et al., 2010). Most automatic methodologies use Natural Language Processing (NLP) tools, such as OpenNLP. However, they must have manual steps to verify the selected concepts. Most of these methods convert text to pre-coordinated concepts, and studies that have used post-coordination are abstract and did not include detailed descriptions of the approach used for constructing the post-coordinated expressions (Silva et al., 2011).

Lamy et al. (2013) presented a semi-automatic semantic method for the mapping of SCT concepts to VCM icons. Kim and Park (2012) proposed an EAV-based data model from CPGs for pressure ulcers wound assessment and to encode its data elements using SCT. It depends mainly on pre-coordination. Kooij et al. (2006) asserted that for standardization of EHR, it is required to use HL7 RIM data model and SCT code for every item. Lau et al. (2013) described a methodology for encoding problem lists used in general practice with SCT. This method has been complemented by Lee et al. (2010). These two methods have encoded raw data sets, and pre-coordination has gotten the highest priority. The Lee's methodology is a complete method. However, it has concentrated on the data cleaning, normalization, and matching steps, and it has not mentioned the physical storage structure of the data, such as EAV. It had not defined if the used EHR database used a standardized model as RIM or not, as asserted by Kooij et al. (2006). In addition, the usage of clinical terminology needs decisions on how the terminology concepts can fit into the system's data structures. For instances, decisions regarding how to use terminologies like SCT with information standards like HL7 RIM must be taken. This process is often called terminology binding. Moreover, the Lee's methodology has not discussed how the codes and its values are semantically stored.

In our paper, the proposed encoding methodology enhances the usage of SCT for encoding medical data by specifying the used physical data model and the unique semantic of used SCT codes. The study is done using a diabetes diagnosis dataset as the case study. First, we collect SCT concepts that match the clinical terms of our dataset. Second, for unmatched terms, we add our proposed custom codes. Finally, we encode our dataset using the collected SCT concept IDs. SCT is a huge ontology and covers so many domains. If case retrieval algorithms depend on the whole SCT, its performance will suffer. The diabetes diagnosis CBR only utilizes a subset of SCT concepts related to diabetes. As a result, we build an OWL 2 ontology for only the collected SCT concept IDs. Some clinical terms in our dataset have no match in SCT. A custom concept-ID is proposed for all of these terms, and the proposed OWL 2 ontology is enriched with these concept-IDs too. The encoded knowledge base is used as CBR's case-base, and the ontology is used as domain knowledge to build an intelligent KI-CBR system.

3. Preliminaries and materials

3.1. Dataset description

To test the feasibility of our proposal, this paper uses 60 EHRs as a case study. The data are obtained and managed by the hospitals of Mansoura University, Mansoura, Egypt. Our diabetes diagnosis features are collected by our domain experts. Some data are collected from a diagnostic biochemical lab (AutoLab, Mansoura, Egypt). The used dataset was collected from January 2010 through August 2013. There are 67 eligible patients, who enrolled in this study. However, seven control subjects were excluded due to limited blood samples for testing AFP. Our dataset contains 70 features for describing diabetic patients and for linking diabetes with other disorders, such as cancer, kidney diseases, and liver diseases. The dataset is distributed as 33.3% pre-diabetic patients, 53% diabetic patients, and 13.7% normal patients. Table 1 shows the description of the considered features in this study.

The data type column in Table 1 defines the data type of each feature. The features with types "I = Instance" of SCT concept initially contains free text. After the encoding process, it contains concept-IDs of the most suitable SCT concepts. This feature type is the focus of this study. Fig. 3 shows an ER model for all entities and used attributes in our dataset. These entities and attributes are enriched from diabetes

Table 1 Patient's features for describing cases.

Feature type	Feature name		Data type	Normal range	UoM	Min–mean–max	F. No.
Demographics	Residence		P, C	{Urban, Rural}	–	–	1
	Occupation		P, C	{Farmer, Police...}	–	–	2
	Gender		P, C	{Male, Female}	–	–	3
	Age		P, N	–	Year	29–48.117–74	4
	BMI		P, N	18.5–25	kg/m^2	20–33.117–45	5
Diabetes lab tests	HbA1C		P, N	$< = 5$	mmol/L	5–6.373–7.4	6
	2h PG		P, N	< 139	mg/dl	165–202.733–235	7
	FPG		P, N	< 99	mg/dl	96–129.633–156	8
Hematological profile	Prothrombin INR		P, N	0–1	%	1–1.16–1.4	9
	Red cell count		P, N	4.2–5.4	10^6/cmm	3.8–5.194–5.88	10
	Hbg		P, N	12–16	g/dL	9.8–12.332–13.4	11
	Hematocrit (PCV)		P, N	37–47	vol%	31.1–35.215–36.8	12
	MCV		P, N	80–90	fl	26.8–71.908–76.4	13
	MCH		P, N	27–32	pg	3.3–25.47–29.4	14
	MCHC		P, N	30–37	%	1.8–35.465–41.7	15
	Platelet count		P, N	150–400	10^3/cmm	135–316.183–2000	16
	White cell count		P, N	4–11	10^3/cmm	6–8.055–9.2	17
	Basophils		P, N	0–1	%	0–1.013–5	18
	Lymphocytes		P, N	20–45	%	21.2–25.768–29	19
	Monocytes		P, N	2–10	%	1.7–2.942–4	20
	Eosinophils		P, N	1–4	%	1–1.897–3.4	21
Symptoms	Urination frequency		I	–	–	–	22
	Vision		I	–	–	–	23
	Thirst		I	–	–	–	24
	Hunger		I	–	–	–	25
	Fatigue		I	–	–	–	26
Kidney function lab tests	Serum potassium		P, N	3.5–5.3	mEq/L	2.4–3.767–4.3	27
	Serum Urea		P, N	5–50	mg/dL	17–31.56–67	28
	Serum Uric acid		P, N	3.0–7.0	mg/dL	3–4.237–7.9	29
	Serum Creatinine		P, N	0.7–1.4	mg/dL	0.9–1.35–3.6	30
	Serum Sodium		P, N	135–150	mEq/L	134–137.833–158	31
Lipid profile	LDL cholesterol		P, N	0–130	mg/dL	50–94.917–170	32
	Total cholesterol		P, N	0–200	mg/dL	158–209.367–275	33
	Triglycerides		P, N	60–160	mg/dL	78–144.767–189	34
	HDL cholesterol		P, N	45–65	mg/dL	30–55.533–65	35
Tumor markers	FERRITIN		P, N	28–397	ng/mL	–	36
	AFP Serum		P, N	0.5–5.5	IU/ml	–	37
	CA-125		P, N	1.9–16.3	U/mL	–	38
Urine analysis	Chemical examination	Protein	I	–	–	–	39
		Blood	I	–	–	–	40
		Bilirubin	I	–	–	–	41
		Glucose	I	–	–	–	42
		Ketones	I	–	–	–	43
		Urobilinogen	I	–	–	–	44
	Microscopic examination	Pus	I	–	–	–	45
		RBcs	I	–	–	–	46
		Crystals	I	–	–	–	47
Liver function tests	S. Albumin		P, N	3.5–5.0	g/dL	1.9–4.082–5.4	48
	Total Bilirubin		P, N	0.0–1.0	mg/dL	0.8–1.317–3	49
	Direct Bilirubin		P, N	0.0–0.3	mg/dL	0.3–0.533–1.6	50
	SGOT (AST)		P, N	0–40	U/L	35–54.567–165	51
	SGPT (ALT)		P, N	0–45	U/L	35–57.317–183	52
	Alk. phosphatase		P, N	64–306	U/L	170–214.2–360	53
	γ GT		P, N	7–32	U/L	18–35.833–98	54
	Total protein		P, N	6.0–8.7	g/dL	3.1–4.858–8.7	55
Females history	Amenorrhea		I	–	–	–	56
	Birth		I	–	–	–	57
	Dysmenorrhea		I	–	–	–	58

(*continued on next page*)

Table 1 (*continued*)

Feature type	Feature name	Data type	Normal range	UoM	Min–mean–max	F. No.
Diagnosis	Diabetes type	P, C	–	–	–	59
Nephropathy	Nephropathy check	I	–	–	–	60
Lipid disease	Hypercholesterolemia check	I	–	–	–	61
Cancer type	Tumor markers	I	–	–	–	62
Liver disease	Liver problem	I	–	–	–	63
Radiological examination	Radiological examination	I	–	–	–	64

Data types = {P = Primitive, I = Instance of SCT concept, N = Numerical, C = Categorical, O = Ordinal}.

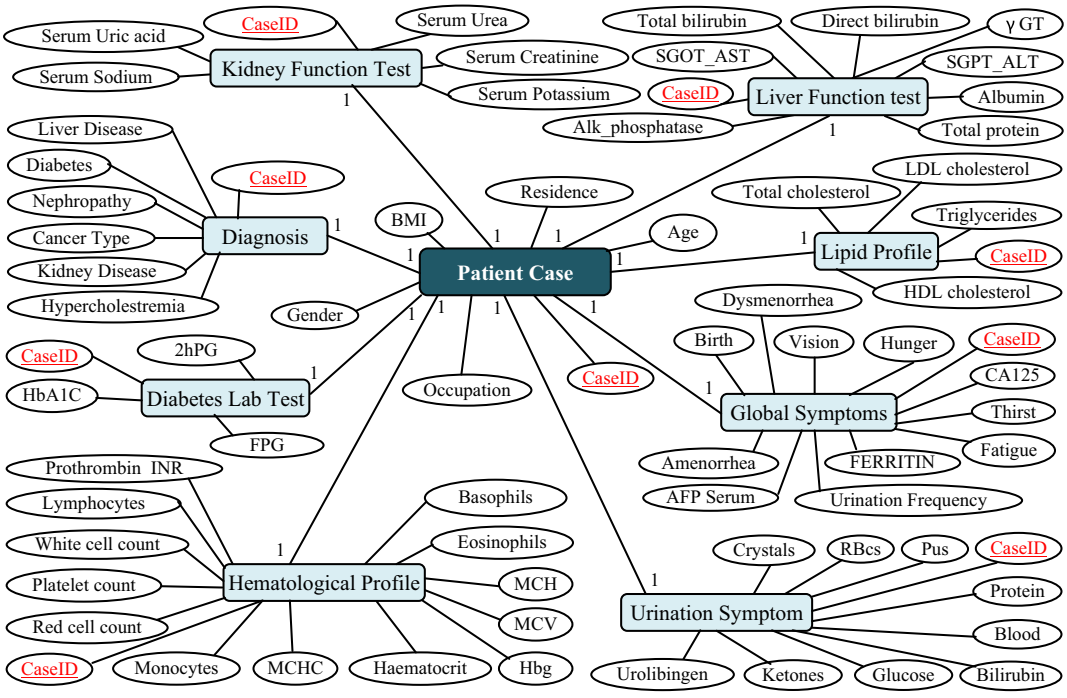


Figure 3 Diabetes diagnosis and other related complaints case-base data model.

diagnosis Clinical Practice Guidelines (CPG) as in National Guidelines Clearing House. Entities and features related to diabetes treatment, medications and drugs are out of scope. The used diabetes diagnosis features are divided into two groups namely structured and semi-structured. The structured or numerical features, such as age and lab tests, are not encoded in SCT concepts because the coding does not enhance the semantic retrieval algorithm of CBR. The semi-structured or texture features (e.g., features in Global_Symptoms table in Fig. 3) are mapped to standard SCT concepts. Our concentration is on CBR semantic retrieval aspect not the sharing and interoperability issues. For example, if feature *HbA1c* = 6.4 is encoded in SCT as |*43396009*: *Hemoglobin A1c measurement*| = 6.4, this code enhances the semantic interoperability. However, it does not enhance semantic retrieval process in CBR. Moreover, the case solution features are not encoded because these features do not participate in measuring the similarity between cases.

3.2. EHR data encoding options

Encoded case-based and SCT-based domain ontologies support the selective retrieval, which is mainly based on subsumption testing of the recorded concepts (Dolin et al., 2002). SCT can code several types of data, such as disorders, diagnoses, symptoms, procedures, examination findings, and laboratory tests. In our study, many of these types are used for diabetes mellitus diagnosis. However, the capabilities for SCT implementation depend on the EHR's logical model (e.g., HL7 v3 RIM). EHR supports the physical storage of pre-coordinated and/or post-coordinated concepts (El-Sappagh et al., 2014c). In our case, the EHR relational database implementation depends on the HL7 v3 Reference Information Model RIM (HL7, 2015). The data are encoded using SCT standardized terminology. Moreover, we have created a standardized relational data model for case-based that is based on RIM to store

all patient data (encoded and not encoded data) (El-Sappagh et al., 2015).

Our focus is on the encoding of patient's clinical observations that are specialization from *Act* class in RIM. To use terminology in an information system, the terminology binding issues must be solved (Benson, 2009). It requires the determination of how the terminology can fit into the information structure. The objective is to enable the efficient and unambiguous combination of HL7 v3 RIM semantic with SCT concept model. The HL7 *Observation* class has two attributes (Code and Value). SCT concepts are stored in either or both of these attributes. The *Code attribute* represents the nature of the action in the patient care (e.g., using concepts from an SCT hierarchy). The *Value attribute* can take a non-terminological value (such as numerical value) or a terminological value (such as SCT concept). Preserving the semantic of clinical information requires the usage of a single approach to encoding all the clinical data. There are two types of patient's observation data:

1. Related data to actions are taken or requested as part of the provision of care as procedures and encounters. In this case, SCT expressions, which focus on procedures concepts, provide appropriate content for the "Code" attribute.
2. Data about clinical findings: In this case, the "*Code*" and "*Value*" attributes are used to represent statements. This type has two categories depending on the ability to divide the finding into Codes and Values.
   - 2.1 The findings have two clearly distinct facets (code and value). For this case, the Code attribute will be: Code = the action was taken to make the finding. The Value attribute will be: Value = the result of the observation, such as Measurement of serum hemoglobin = 14 g/dl.
   - 2.2 The findings are captured in a single "nominalized" expression. In this case, SCT supports the encoding of these assertions in a unique expression by using concepts from the 404684003|clinical finding and 413350009|finding with explicit context hierarchies. For instance, the finding "has a fracture of her left femur" is a single statement. This expression can be stored in Code or Value attributes.

If a clear separation cannot be defined between the action (i.e., Code) and the result (i.e., Value), the normalized concept can be represented by only one of the Code or Value attributes. In our work, we assert the separation of code and value to facilitate the selective retrieval of SCT concept. This method supports the implementation of semantic retrieval algorithms in CBR system.

The problem of standardizing clinical data is not completely solved even if implementers are using common terminologies. Therefore, it is possible to represent the same information in multiple ways while using standard terminologies and information models. The same information can be represented using one or several concepts. It is critical to be rigorous in the selection of the concepts and the way of their representation. In other words, the coding of data can be achieved by using pre-coordination or post-coordination (Andrews et al., 2008). These methodologies have advantages and disadvantages (Dolin et al., 2002). In our RIM-base relational data model, the pre-coordinated concept is stored in a single field even if it represents a compound concept (e.g., *190419001|diabetes mellitus, adult onset, with other specified manifestation*). However, the post-coordinated concept is stored in multiple fields and represents more complex concept expressions and relevant qualifiers to be expressed where pre-coordinated concepts are not available or sufficient. The two forms can be used together for coding EHR data.

Andrews et al. (2008) concluded that pre-coordination is easier and ensures consistency. For post-coordination, rules must exist for the consistent use of SCT. Moreover, transforming SCT concepts into normal forms can achieve consistency and support selective retrieval (Dolin et al., 2002). The IHTSDO implementation guide is limited. IHTSDO suggests that each hierarchy has a particular purpose. However, Lee et al. (2010) found overlaps between "clinical finding" and "morphologic abnormality" hierarchies, when mapping a palliative care dataset. Højen and Gøeg (2012) suggested a set of guidelines for consistent coding using SCT. However, Rasmussen and Rosenbeck (2011) concluded that local guidelines cannot handle the inconsistency usage of terminology between organizations.

As a result, the coding by using post-coordination has many problems. In addition, there is no complete and uniform methodology for achieving it. For preserving the consistency between CBR case-base, user query, and SCT-based domain ontology, this paper concentrates on pre-coordinated concepts. For post-coordination that is stored as free text, the case retrieval algorithm will be more complex as the short and long normal forms need to be generated when using structural subsumption (Lee, 2014). Testing the equivalence or subsumption between post-coordinated concepts or between pre- and post-coordinated concepts has not solved the problem yet (Lee et al., 2013). Moreover, SCT pre-coordination has been proved sufficient for coding clinical data, and local concepts can extend its coverage (Wasserman and Wang, 2003). Moreover, we encode a diabetes diagnosis dataset based on our created SCT reference set. The paper mainly depends on *404684003|Clinical finding* and *71388002|Procedure* as the main hierarchies.

3.3. Matching algorithms

Table 2 shows the four matching algorithms used in SCT to ICD-10 mapping project (Lee, 2007). We utilize these algorithms to find matches between SCT concepts and diabetes diagnosis terms. The first three techniques are lexical matching, and the fourth one is semantic matching.

3.4. Normalization steps

Matching algorithms are applied to both original and normalized terms. The normalization of diabetes terms in EHR and SCT terms is done by removing "noise" using the UMLS normalization steps (UMLS, 2015). These steps include:

- Remove genitive, e.g., Kidney's Diseases → kidney diseases.
- Remove words that do not affect meaning (e.g., stop words, exclude words, and SCT prefixes). *Stop words* are frequent short words that do not affect the phrase: and, by, for, in,

Table 2 The used matching algorithms.

Algorithm	Explanation
Exact match	Exact string match where all words are same and in the same sequence
Match All	String match where all words are same but not necessary in the same order; additional words allowed
Partial match	String match where one or more words is found
Semantic match	For inactive concepts use historical relationships Was-A, Same-As, May-Be-A, Replaced-By to find current concepts
Unmatched	Assigned when no match is found

of, on, the, to, with, no, and (nos). *Exclude words* are words that may change meaning of the word but if ignored help to locate a term otherwise missed, such as: about, alongside, an, anything, around, as, at, because, before, being, both, cannot, chronically, consists, covered, does, during, every, find, from, instead, into, more, must, no, not, only, or, properly, side, sided, some, something, specific, than, that, things, this, throughout, up, using, usually, when, while, and without. *SCT prefixes* are letters with special meaning in SCT: [X], [D], [M], [SO], [Q], [V], and so on. For example, Laceration of the kidney with open wound into the abdominal cavity → Laceration kidney open wound abdominal cavity.

- Convert to lowercase, e.g., **K**idney **D**iseases → kidney diseases.
- Strip punctuation, e.g., Volume depletion, renal, due to effector loss **(**hormonal deficit**)** → Volume depletion renal due to the effector loss hormonal deficit.
- The uninfected phrase, e.g., Kidney Diseases → kidney disease.
- Sort words, e.g., Kidney Disease → disease kidney.

These normalization steps help to improve the results as well as the time it takes to search for matches. The process of matching involves the checking of these matching algorithms one at a time to find the best candidate SCT concepts. For each matching algorithm, we begin with the original terms, then the UMLS normalized terms. The type of match is based on the algorithm applied. For example, we group the terms matched with SCT concepts using *Exact Match* algorithm, using *Match All* algorithm, etc.

3.5. Coding rules

To manage the coding process and to ensure consistency, some coding rules were set up for the process. Examples of these coding rules include:

(1) Default context is assumed for all concepts (IHTSDO, 2015), so *243796009|situation with explicit context* hierarchy is not utilized.
(2) The 404684003|Clinical finding (finding)|, 363787002|Observable entity (observable entity)|, and then 71388002|Procedure (procedure)| hierarchies are the prioritized hierarchies in this order.
(3) Clinical findings hierarchy is more beneficial for retrieval and reuse purposes than for observable entities.
(4) Consistency is checked when selecting the correct hierarchies and the correct concepts in these hierarchies (Højen and Gøeg, 2012).
(5) Appropriateness check for the type of the found concepts, which is done as there are 19 types of concepts in SCT.
(6) The most specific concepts are tried first.
(7) To enhance the semantic retrieval process of CBR system, pre-coordinated concepts are the only choice, and for not matched terms, a new custom concept is proposed to extend SCT.

3.5.1. Types of concepts

We concentrate on CBR functionality, especially on case retrieval algorithms. Semantic retrieval is improved by using domain ontology and encoded case-base. As a result, we use our developed OWL 2 ontology from SCT terminology as a domain ontology (El-Sappagh et al., 2014c). This ontology contains all concepts related to a diabetes diagnosis. Not all SCT concepts can improve semantic retrieval. The concepts that have siblings and exist in IS-A hierarchies are suitable. In our case, concepts with primitive data types, such as numerical and dates, are not encoded. Although, numerical clinical terms as HbA1c level = 11% can be encoded as *43396009|HbA1c – Hemoglobin A1c level = 444751005|high hemoglobin A1c level|*. These types of codes do not improve the semantic retrieval algorithm of CBR because the semantic retrieval algorithms depend on the calculation of semantic or clinical distances between concepts. These types of concepts have no concept hierarchy that is suitable for semantic retrieval. On the other hand, these concepts have synonyms that are suitable for improving interoperability. However, interoperability is out of scope. Moreover, data fields of our database schema are not encoded.

To illustrate the idea, Fig. 4 is a snapshot from CliniClue SCT browser for the term "*HbA1c*". We have searched the used hierarchies (i.e., Clinical finding, Procedure, and Observable Entity), and the most suitable hierarchy is the Procedure. As shown in Fig. 4, the concept *43396009|Hemoglobin A1c measurement* has no semantically equivalent concepts, which can replace it in the CBR query. Moreover, the concepts in its *IS_A* hierarchy cannot be used for calculating clinical distance in semantic queries.

The only reason for using these types of concepts is to support interoperability. On the other hand, Fig. 5 is a snapshot of *90708001|kidney disease* (*disorder*) concept. This *IS_A* hierarchy contains the different kidney diseases concepts and their relationships. This hierarchy is richer than hierarchy in Fig. 4, and it can enhance the calculation of clinical similarity and representation of the physician query.

4. The preparation framework

In this section, we discuss the proposed case-based overall preparation framework. In addition, we discuss the proposed encoding methodology in detail.

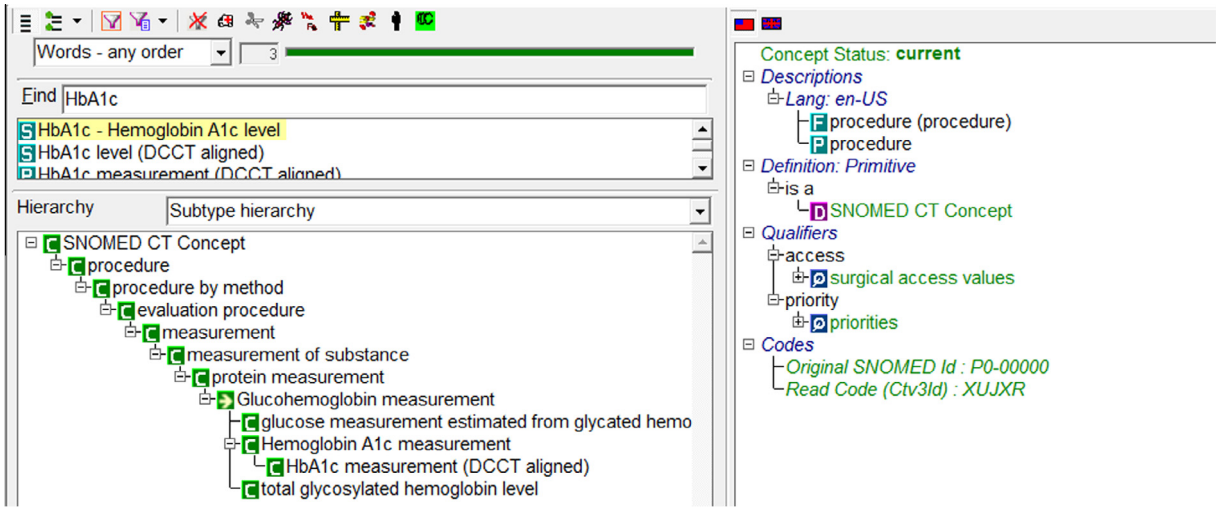


Figure 4 HbA1c sub-tree in the SCT *Procedure* hierarchy.

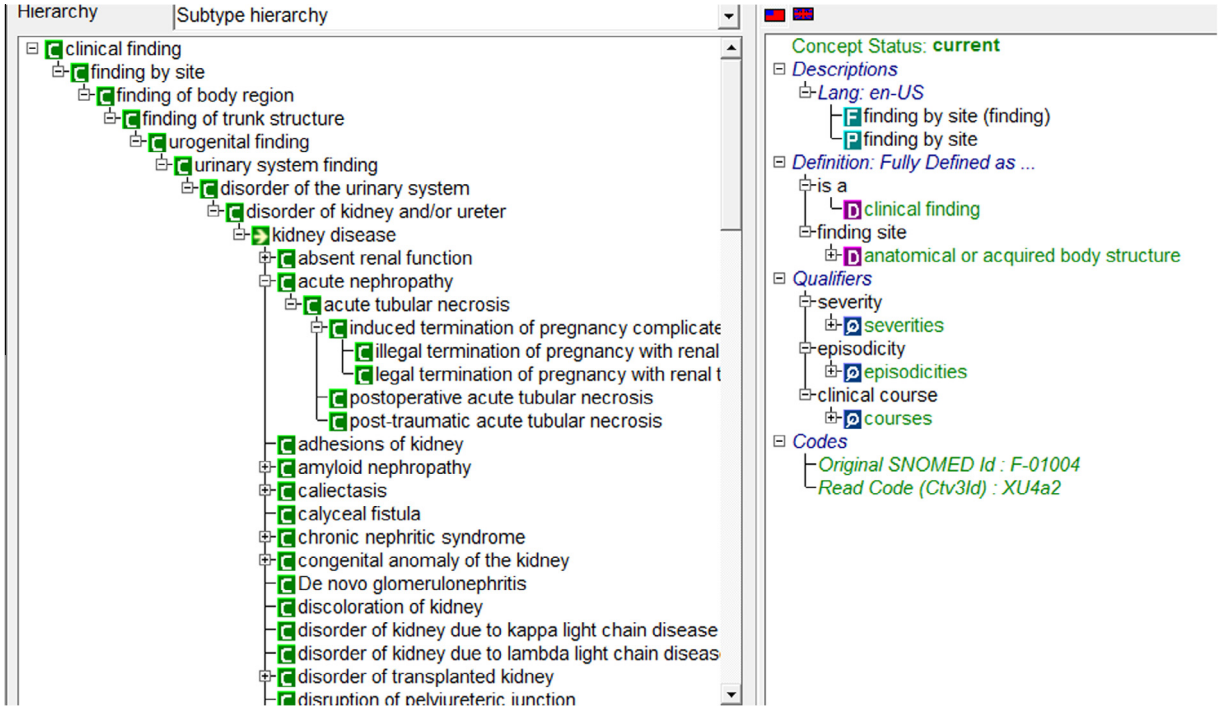


Figure 5 Kidney disease sub-tree in the SCT *clinical finding* hierarchy.

4.1. The proposed case-based preparation framework

Our proposed case-based preparation framework consists of three sequential phases including data pre-processing, data encoding, and data fuzzification, as shown in Fig. 6. It creates a case-based knowledge from EHR schema and contents. The framework maps EHR structure and content into a standard case-base structure and content, respectively. We collect all diabetes features from the distributed EHR systems. *For data preparation phase*, a set of data pre-processing techniques are applied to these data to create a high-quality dataset (El-Sappagh et al., 2014a). Next, a standard case-base structure based on HL7 RIM is designed to allow the terminology binding into the database structure (El-Sappagh et al., 2015). The resulting case-base structure is formed in the Problem–Solution–Outcome form. *For the encoding phase*, the encoding process based on standard terminology (i.e., SNOMED CT) is performed on the dataset. This phase is the focus of this paper, and it will be described next in detail. After coding the case-based unstructured contents, there are two properties of resulting data. The textual data are represented with formal concept-IDs, which have semantic and clinical meanings. No synonyms of a clinical term appear in the dataset. For example, all the terms: kidney disease, a disorder of kidney, renal disorder, nephropathy, renal disease, and nephrosis are coded using the unique identifier *90708001|kidney disease* (*disorder*).

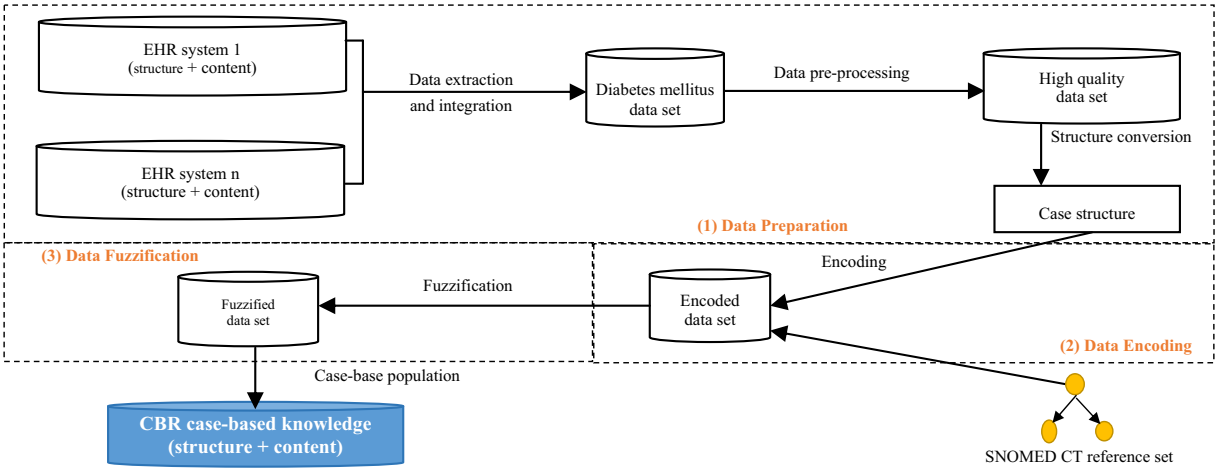


Figure 6 The case-base preparation phases.

Finally, *for the fuzzification phase*, it is used to handle vague knowledge. Physicians always describe patients using vague terms, such as the sugar level is high; the patient is obese, and so on. Moreover, patients often describe their conditions using imprecise terms. As Zadeh (2003) argued, knowledge acquired through experience is perception-based. As a result, it is subject to imprecision and vagueness. If this knowledge is not treated in a suitable way that can consider and convey its inherent imprecision, usually leads to the poor effectiveness of the knowledge-based systems that use it. Vagueness can be handled using fuzzy logic (Zadeh, 2003), which has been used in diabetes diagnosis rule-based systems (Lee and Wang, 2011). For space restrictions, fuzziness will be handled in another work. The powerful case retrieval algorithms can be created for the resulting case-base. These algorithms can benefit from the semantics of the domain ontology, the vague data, and the crisp data. This study focuses on the encoding phase where the case-based textual or unstructured features are converted to a formal and precise SCT concept IDs.

4.2. The proposed case-base encoding methodology

Implementing SCT is relatively still new, and it is considered as a challenging proposition. In this section, we propose a methodology for encoding EHR data, and we use the diabetes diagnosis dataset as a case study. This dataset is encoded to play the role of a case-base in a CBR-based CDSS for diabetes diagnosis. This methodology consists of five sequential steps. An overview of this method is shown in Fig. 7.

Step 1: *Determine the implementation strategy used to represent SCT concepts in HL7 RIM fields*.As discussed in Section 5, the most suitable strategy is the two-faceted form (Code = Value) where Code and/or Value can store a pre-coordinated concept-IDs. Pre-coordination can be used to represent concepts in lab tests, symptoms, diagnosis, physical examination, procedures, and so on (Ryan et al., 2007).Data types need to be filtered. There are two reasons for not encoding numerical features. First, SCT is not designed to encode data of primitive types as numbers (Lee et al., 2010). Second, there are no concept hierarchies for any of these concepts. As a result, all numerical features are not encoded in our study, which include: age, BMI, 2 hPG, Total Bilirubin, Direct Bilirubin, SGOT (AST), SGPT (ALT), γ GT, total protein, albumin, FPG, HbA1C, prothrombin (INR), basophils, eosinophils, monocytes, lymphocytes, white cell count, platelet count, MCHC, MCH, MCV, hematocrit, hemoglobin, red cell count, serum potassium, serum creatinine, serum uric acid, serum urea, etc. The coding of these features does not enhance the semantic retrieval algorithm of CBR. As a result, the features of these data types should be filtered out and should not be a part of the potential list of terms to be encoded. On the contrary, the unstructured or textual concepts (e.g., urination frequency, urine glucose, thirst, hunger, vision, kidney disease, hypercholesterolemia diagnose, cancer type, etc.) are encoded and linked to our created SCT ontology. The semantic case retrieval algorithm measures the semantic distance between coded values of both query and stored cases.

Step 2: *Collect a list of diabetic description data elements*.Two types of elements are available, i.e., organizing elements and resulting elements. The organizing elements are the selected tables and selected data elements (columns) names from the EHR database. *Organizing elements* names can be manually extracted by viewing the database schema and copying each data element name, or using a DBMS application to export the schema into a text file or spreadsheet. *The resulting elements* are the list of free text, and coded values (data items) contained in these columns. Domain expert and diabetes CPGs (Rodríguez et al., 2011) can enhance the creation of the complete list of data items.Our SCT reference set ontology contains 1161 concepts, 3306 descriptions, and 1713 relationships (El-Sappagh et al., 2014c). In this paper, we concentrate on the concepts appeared in our dataset, which participate in the case-based construction. These data are used by a physician when describing patient conditions in clinical encounters. These terms include: *diabetes symptoms* (e.g., Vision, Fatigue,

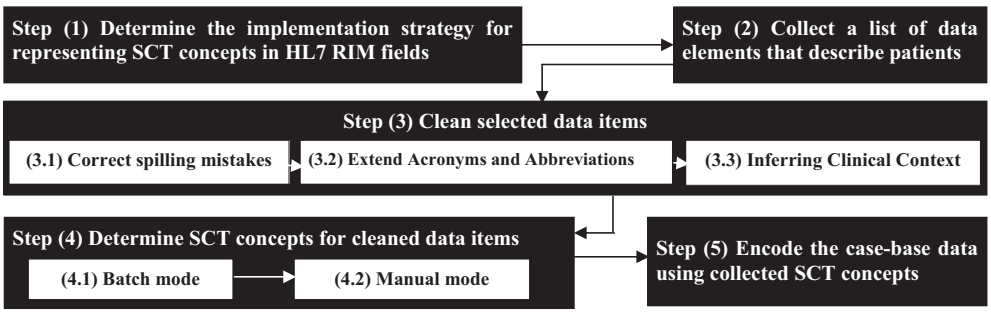


Figure 7 The proposed case-base encoding methodology.

Hunger, Thirst, urination frequency, Protein, Blood, Bilirubin, Glucose, Ketones, Urobilinogen, Pus, RBcs, Crystals, etc.), *disorders* (e.g., kidney diseases or nephropathy, cancer types, liver diseases, Hypercholesteremia, etc.), and *medical history* (e.g., Dysmenorrhea, Amenorrhea, AFP Serum, Ferritin, Birth Type, etc.) Our OWL 2 ontology for diabetes diagnosis (El-Sappagh et al., 2014c) does not support some diabetes disorders such as heart disorders, eye disorders (i.e., retinopathy), hypertension, and nervous system disorders (i.e., neuropathy). This enhancement will be considered in a future work. Moreover, no medication terms exist, symptoms related to physical activity and smoking do not exist, and some history risk factors as a first-degree relative with diabetes do not exist. As a result, the OWL 2 ontology that is created in this study enhances the previously created ontology in El-Sappagh et al. (2014c).All diabetes symptoms are categorized or codified in the pre-processing phase, as shown in our framework in Fig. 6 (El-Sappagh et al., 2014a). Table 3 provides a sample of these coded values. These codes are stored in lookup tables, which are usually external to the EHR databases and are separate from the EHR systems. Generally, the coded values may be recorded in one table or the coded values for each data element may be in individual tables. In our case, each element is stored in a separate table. Encoding the coded values with SCT can be considered as a form of mapping.

Step 3: *Cleaning the Selected Data Items*.Poor quality text requires cleaning of each term in the identified list of terms. The data cleaning process ensures the data items are consistent and accurate. NLP algorithms and tools can automate data cleaning and encoding to extract relevant concepts from free-text data. We do not split any concept because our data elements represent atomic concepts, and splitting of such terms can cause loss of semantics. The following steps are applied to the processed terms:

- *Correct spelling mistakes*: Lexical matching requires correction of spelling mistakes. For example, Hypercholestremia becomes Hypercholesterolemia; Glomerulonephrinitis becomes Glomerulonephritis; Urobilinogen becomes Urobilinogen, and so on.
- *Extend Acronyms and Abbreviations to their full-form*: Because they caused mismatches in the string matching process, acronyms, and abbreviation are removed, such as Alk. Phosphatase → alkaline phosphatase, CA-125 → cancer antigen 125 or carbohydrate antigen 125, AFP Serum → serum alpha-fetoprotein, RBcs → red blood cells in urine, HCV → Hepatitis C Virus, HCC → Hepatocellular carcinoma. The developer toolkit of SNOMED CT version 2013 contains a large number of abbreviations as synonyms. It provides a tool for mapping abbreviations to concepts in the form of a "word equivalents" table (IHTSDO, 2015c). This table contains 10185 equivalent words and abbreviations to enhance the searching process by providing semantic and syntactic information about the term. For instance, the abbreviation *DM* = *Diabetes Mellitus*, and the words *Kidney*, *nephric*, *nephritic*, and *renal* are all equivalent. Moreover, SCT provides extended text definitions for 704 concepts.
- *Inferring Clinical Context*: When searching for an SCT concept that matches a particular clinical term, we search three hierarchies in the following order: Clinical finding, observable entity, and then procedure.

Table 3 Examples of codified diabetes features.

Symptom	Code	Symptom	Code
Urination frequency	*Normal* = 3–5 times urination per day + = 6–8 times urination per day ++ = 9–10 times urination per day +++ = More than 10 times urination per day	Females history	*Dysmenorrhea*: Normal, +, ++ *Amenorrhea*: Normal, Amenorrhea *Birth*: Normal, Twins, Sterile, overweight baby
Urine Analysis	Nil, +, ++, +++		...

Step 4: *Determine SCT concepts for cleaned data items*.After collecting the list of data items to be coded from case-based dataset, we use batch and manual lexical matching to determine the matched SCT pre-coordinated concepts.

- *Batch mode*: For perfect automation of this step, NLP algorithms and tools are required, but NLP is out of scope. In our case, we have created a relational database for SCT core tables and batch lexical matching is done by using a set of SQL queries on SCT *Description* table (Wasserman and Wang, 2003). It is based on syntactic string similarity (Jiaheng et al., 2013). Moreover, SCT version 2013 developer toolkit provides table named *WordKeyIndex_Concepts* where each row in this table is a word followed by a reference to a Concept. A Concept is referenced if the word appears anywhere in the combination of the Fully Specified Name, Preferred Term, or Synonyms. It provides another table named *WordKeyIndex_Description* where each row in this table is a word followed by a reference to a Description in which this word appears.These tables are checked if no match found. We depend on SCT as a whole concentrating on clinical finding, observable entity, and procedure hierarchies. Our previously created SCT subset is checked if it contains these concepts or not. If concepts do not exist in our subset, then these concepts and all concepts in their paths up and down are added to our subset. Another ontology version is created for this subset to be used for semantic retrieval algorithm in CBR system. The inputs to the batch-matching step include a list of data items, a list of normalized data items, original SCT descriptions, and a normalized set of SCT descriptions.Moreover, the lexical similarity is not enough, and we depend on three concept matchings (Agrawal and Elhanan, 2014). The first is the lexical information conveyed by a concept's unique descriptor and other possible synonyms. The second level of information is the hierarchical positioning of the concept within SCT's directed acyclic graph by utilizing is-a relationship type to create a subsumption structure. The third level is the formal logical definition of each concept represented by a set of defining relationships to other concepts. The following steps are executed in batch matching:
- *Normalization steps* (*Wang et al., 2014*): As discussed in Section 3.4, before comparing strings, normalization steps for both patient terms and SCT concepts are essential (UMLS, 2015). These steps include: Words within parentheses are removed for SCT; terms are tokenized into atomic forms and converted into lowercase; stop words such as "a", "the", "of", and "NOS" are removed (NLM, 2015). Moreover, punctuation is removed from multi-word expressions; lexical variations of terms using equivalent words can be created (IHTSDO, 2015c); finally sorting of the result words in alphabetical order is performed. For example, the concept "*190390000| Type II diabetes mellitus with gangrene* (*disorder*)" is normalized to "*Diabetes gangrene Mellitus Type II*."
- *Matching step*: A cycle through matching algorithms in Table 2 (using SQL queries) on normalized and not normalized terms is performed. To achieve the context-based matching, we search clinical finding, observable entity, and procedure hierarchies in order. Remove any successful exact-match and match-all terms from further matching cycles.
- *Reviewing step*: Manually review and verify the matched concepts, one term at a time, with a domain expert to check accuracy.
- *Specification step*: Manually navigate the SCT hierarchies down to find a more specific concept for each found match. The most specific concepts are required. In this step, our domain experts in cooperation with SCT coder search in the SCT hierarchies down to more specifically refine the selected concepts.
- *Manual mode*: The manual matching examines the remaining concepts from batch matching. In this step, we use SCT browsers. The search may be limited to a specific hierarchy, a specific reference set, or the whole SCT. We navigate down the SCT hierarchies to find specific concepts after finding initial matches. Some studies prefer some hierarchies than others. For example, Højen and Gøeg (2012) and Rasmussen and Rosenbeck (2011) preferred *Clinical Finding* than *Observable Entity*. Ryan et al. (2007) preferred *Observable Entity* than *Clinical Finding*. Lee et al. (2010) preferred *Clinical Finding* than *Morphologic Abnormality*. The priority measures are varying, such as depending on content coverage, the level of granularity or possibility to express data in the form of Code = Value. Højen and Gøeg (2012) provided the guidelines for selecting the appropriate SCT hierarchy according to the type of data. The searching process is done as follows:
  1. Our two domain experts with extensive training in the methodology search *independently* for SCT concepts that match clinical terms by using some browsers including CliniClue (Clinical Information Consultancy Ltd, 2015), SNOW OWL (B2i Healthcare, 2015), BioPortal (*BioPortal, 2015*), IHDSDO[2], and WordNet (WordNet, 2015). The first two browsers support the category-specific search for a term and some linguistic techniques in searching process. WordNet browser is used to find equivalent words. The selected concepts from both experts are compared. If they are equal, then go to the next step. If they are not equal, the searching is re-evaluated.
  2. If a complete or exact match in one hierarchy is found, then use it. If there are complete matches in more than one hierarchy, then Clinical Finding, Observable Entity, and then Procedure are given the highest priority in this order.
  3. If there is no perfect concept, then search for synonyms.

[2] http://browser.ihtsdotools.org/.

4. If there is no perfect match, then search for a concept on the SCT hierarchical levels above. In this case, the EHR data are more precise than SCT concept.
5. If still no perfect match can be found, navigate using the SCT hierarchy from the top down. In this case, the EHR data is more general than SCT concept.
6. If there is no match then search for concepts where both EHR data and SCT concepts are partially overlapped.
7. If selected concept is inactive, an attempt is made to locate an active concept (i.e., concept status = 0 or current) through the historical relationships such as "149016008|MAY BE A (attribute)|", "384598002|MOVED FROM (attribute)|", "370125004|MOVED TO (attribute)|", "370124000|REPLACED BY (attribute)|", "168666000|SAME AS (attribute)|" and "159083000|WAS A (attribute)|".
8. If no match is found, then create a new concept to be used in an SCT extension.

Step 5: *Encode the case-based data using collected SCT concepts*.We have created a table with two columns of case-base clinical terms and their selected SCT concept-IDs, and we have used this table to replace each term with its corresponding SCT concept-ID. We have built a JAVA application, which connect to our case-base database and replace the unstructured clinical values by their equivalent SCT concept-IDs. For each selected concept ID, we have collected all its parents up to the root concept (i.e., *138875005|SNOMED CT*), and we have collected all of its children. We have created an OWL 2 ontology for all selected SCT concepts, and the used concept IDs in case-base are connected to our created OWL 2 ontology. In the next section, we propose and test a semantic retrieval algorithm, which utilizes our proposed OWL 2 ontology.

5. Results

In this study, we proposed a case-base data encoding methodology based on SCT. It has the potential to become a general-purpose terminology encoding approach. It can be used in different clinical systems because it depends on standard technologies, such as SCT and HL7. There are three other results of this study. In the following subsections, we will discuss these results in detail.

5.1. Encoded elements

We collected a list of SCT concepts for our case-based textual clinical data elements. The original diabetes diagnosis case-based contained 64 features. There are 38 numerical features, which are ineligible for encoding in SCT. The remaining 26 features have 82 free text values. Fig. 8 presents the results of the mapping process. The partial matching represents the highest percent (i.e., 57.3%). Fig. 9 and Table 4 present a comparison between domain coverage for our methodology and some other methodologies. Table 5 shows some concept matching in our case-based terms. The pre-coordinated concept coverage of SCT of our case-based dataset is ~75%.

The unmatched terms values include Bilirubin, Urine Pus, Urine Crystals, and Urine RBcs. Most of these terms are in urine analysis. An expression is categorized as "No match" when the exact meaning of its clinical expression is not represented in SCT. The post-coordination has not been utilized to avoid the complexity of semantic similarity algorithm implementation. Table 6 shows a sample of encoded data that cannot be coded using SCT. SCT handles other urine analysis concepts as Urine Protein, Urine Urobilinogen, Urine Blood, Urine Glucose, Urine Ketones, and others in a different and more consistent way. It provides concepts for all of the expected values for Urine Blood, as shown in Table 5. As a result, our study has uncovered these significant inconsistencies in SCT concept coverage.

Some not matched concepts as Shrink kidney can be post-coordinated as *181414000|entire kidney* (*body structure*): *246115007|size* (*attribute*) = *260371004|decreasing* (*qualifier value*). However, the paper's case-base data model is not designed for post-coordination, and this action will complicate the case retrieval algorithm. Therefore, for all unmatched concepts, we add custom concepts, and we will propose these concepts as an extension set for future releases of SCT.

5.2. Encoded case-based contents

We have coded our case-based contents with the selected SCT concepts. The resulting case-based has a standard structure according to HL7 RIM (El-Sappagh, 2015). Moreover,

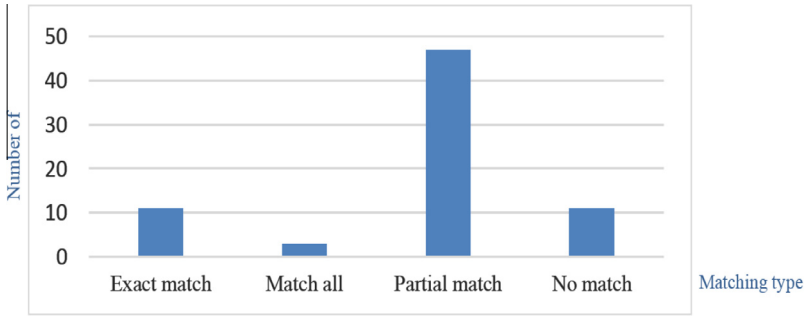


Figure 8 The matching results of SCT concepts.

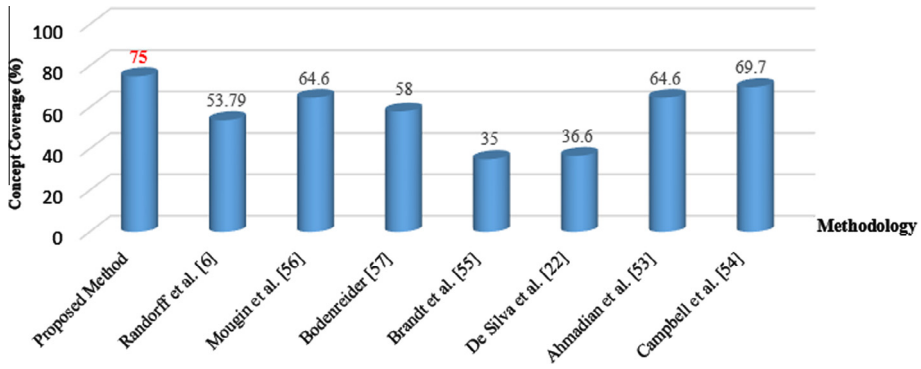


Figure 9 A comparison of concept coverage between the proposed methodology and some other ones.

Table 4 An evaluation of proposed encoding methodology.

Methodology	The proposed method	Højen and Gøeg (2012)	Mougin et al. (2011)	Bodenreider (2009)	Brandt et al. (2011)	Silva et al. (2011)	Ahmadian et al. (2010)	Campbell et al. (1997)
Domain	Diabetes diagnosis	Hybrid	Adverse drugs	MedDRA	Orphanet rare diseases	Tomography	Pre-operative assessment	Hybrid medical and nursing data
Concept coverage (%)	**75**	53.79	64.6	58	35	36.6	64.6	69.7

according to the results of this study, it becomes standard contents according to SCT standard ontology. The resulting case-based supports:

1. Using it as a knowledge source for building a KI-CBR for diabetes diagnosis. This system is more intelligent than traditional ones because it depends on a smart semantic retrieval algorithm. This algorithm can use the created OWL 2 ontology to calculate the clinical distance between cases not just geometric distances. This system will be implemented in a future work.
2. Because semantic interoperability is achieved by encoded medical data in EHR and medical knowledge in case-based, CBR system can easily be integrated with the distributed healthcare environment. Moreover, the case-based can collect cases from the distributed EHR systems, and finally building distributed CDSS becomes applicable.

Fig. 10 shows a snapshot of our resulting case-based content. It contains only eight cases, and each case is represented by only 24 features plus a solution feature of diabetes conditions. As said before, no numerical features have been encoded. The unstructured features have been encoded using SCT Concept-IDs. We selected to encode the data using the Concept-ID not the Fully Specified Name to maximize the accuracy of encoded data. Each concept in SCT has a specific Concept-ID, which is the primary key of the concept. However, the readability of the resulting encoded case-base is not human-oriented. For space restrictions, this issue will be handled in a future work. The numerical features will also be fuzzified in a future work, and the resulting case-base will support the fuzzy and semantic retrieval algorithms. Building a CBR for diabetes diagnosis that handle these two critical issues will have an accurate performance and flexible querying capabilities (Alexopoulos et al., 2010; Melton et al., 2006).

5.3. SCT reference set OWL 2 ontology

Finally, we built an OWL 2 ontology of the selected SCT concepts using Protégé 4.3.[3] These concepts include the concepts' complete paths from SCT root concept (i.e., *138875005| SNOMED CT Concept*) to the most specific leaf concepts. The resulting ontology contains 550 unique concept IDs. Fig. 11 shows a snapshot of the ontology. This ontology will be used as domain knowledge in the KI-CBR system for diabetes management to improve the semantic retrieval process. It allows CBR system to think like domain experts when searching for similar cases. Our ontology is mainly dependent on the Concept IDs and IS_A relationship.

The overall architecture of our encoding process and KI-KBR system is shown in Fig. 12. There are six steps ranging from extracting medical terms from our dataset to the implementation of the semantic case retrieval algorithm. Steps 1, 2, 3, and 4 are included in our encoding methodology. Step 5 results in the ontology in Fig. 11. Step 6 will be handled in the following section.

5.4. Semantic case retrieval algorithm

To implement a KI-CBR system, the case retrieval algorithm based on similarity measures is a fundamental component.

[3] http://protege.stanford.edu/.

Table 5 Examples of matches between case-base and SCT terms (e.g., exact, all, partial and semantic matches).

Table	Feature	Textual value	SCT concept	SCT descriptions	Description type	Match type	Status
Global_Symptom	Urination frequency	Normal	162115004	543222010|micturition frequency normal (finding)	F	Partial match	0
				252774015|micturition frequency normal	P		
		+	249291007	371976012|infrequent urination (finding)	F	Partial match	0
				371977015|does not urinate often enough	S		
		++	162116003	252775019|increased frequency of urination (finding)	F	Partial match	0
				252780011|passes water too often	S		
				252776018|increased frequency of micturition	S		
Urination_Symptom	Urine blood	Normal	167297006	259827017|urine blood test = negative (finding)	F	Partial match	0
		+	167300001	259831011|urine blood test = + (finding)	F	Partial match	0
		++	167301002	548904018|urine blood test = ++ (finding)	F	Partial match	0
		+++	167302009	548905017|urine blood test = +++ (finding)	F	Partial match	0
Diagnosis	Cancer	Liver cirrhosis	19943007	748614010|cirrhosis of liver (disorder)	F	Exact match	0
				33568015|cirrhosis of liver	P		
				33572016|hepatic cirrhosis	S		
Global_Symptom	Birth	Twin birth	28030000	758665012|twin birth (finding)	F	Exact match	0
		Overweight baby	276613009	2965943019|high birth weight (disorder)	F	Match All	0
				412838015|high birth weight baby	S		
		Normal birth	45723005	782994017|normal female reproductive function (finding)	F	Partial match	0
		Sterile	10114008	526803011|female sterility (finding)	F	Exact match	0

Description type (F = Fully Specified Name, P = Preferred term, S = Synonym), status is the concept status (0 = current or active concept).

Semantic similarity functions can measure the clinical similarity between concepts. It is based on ontology. In Table 1, all features with *datatype* = *I* store SCT concept IDs and require semantic similarity measures. We propose a new hybrid measure based on path length and concept features. *First*, for path length, our similarity is based on the *depth* of the Least Common Ancestor (LCA) of the two concepts and the *closeness level* of concepts to their LCA. In other words, (1) the deeper the LCA, the more specific it is considered and, thus, the more similar the compared concepts are assumed; (2) the closer the two concepts are to their LCA, the more similar they are. *Second*, to quantify similarity for concept features, the commonalities and differences between concepts must be considered (Harispe et al., 2014). JCOLIBRI API[4] uses four semantic similarity measures: path-based such as *fdeep_basic* and *fdeep*, and feature-based such as *cosine*, and *detail* (Recio-García et al., 2014). These measures have been tested (Recio-García et al., 2014); however, these algorithms are not accurate compared to our algorithm. Their limitations come from: path-based measures do not take into account the depth of concepts from their LCA, and feature-based measures depend only on the commonalities between compared concepts.

Our proposed measure overcomes these limitations and integrates path based and feature-based approaches. The proposed composite similarity measure $SIM_{Semantic}(u, v)$ uses Eq. (1):

$$SIM_{Semantic}(u, v) = w_1 sim_{Path}(u, v) + w_2 sim_{feature}(u, v) \quad (1)$$

where $w_1, w_2 \in (0, 1]$ are weights for $w_1 + w_2 = 1$, u and v are the compared medical concepts, and $sim_{Path}(u, v)$ (Eq. (2)) is an adapted version of Wu and Palmer 1994 (Eq. (3)) because $sim_{Wuandpalmer}(u, u) < 1$ which violates the Identity Of the Indiscernibles (IOI) property (Harispe et al., 2014).

$$sim_{Path}(u, v) = \begin{cases} 1 & \text{if} \quad u = v \\ sim_{Wuandpalmer} & \text{otherwise} \end{cases} \quad (2)$$

[4] http://gaia.fdi.ucm.es/research/colibri/jcolibri.

$$sim_{Wuandpalmer}(u,v) = \frac{2 * depth(lca(u,v))}{shortest\_path(u, lca(u,v)) + shortest\_path(v, lca(u,v)) + 2 * depth(lca(u,v))} \quad (3)$$

In addition, $sim_{Feature}(u,v)$ is based on Sánchez et al. (2012), Eqs. (4) and (5):

$$sim_{Feature}(u,v) = 1 - Dist_{Batet}(u,v) \quad (4)$$

$$Dist_{Batet}(u,v) = \log_2\left(1 + \frac{|A(u) \setminus A(v)| + |A(v) \setminus A(u)|}{|A(u) \setminus A(v)| + |A(v) \setminus A(u)| + |A(u) \cap A(v)|}\right) \quad (5)$$

where $A(u)$ is the set of ancestors of u, i.e., $A(u) = \{v | u \Delta v\}$, $A(u) \setminus A(v)$ is specificity of u, and $A(u) \cap A(v)$ is the commonality between u and v.

After measure formalization, it is evaluated by comparing it with the most popular semantic similarity algorithms in CBR (i.e., with JCOLIBRI2 (Recio-García et al., 2014)). As shown in Fig. 2, this is done by doing experiments using a sub-ontology from our SCT ontology for kidney diseases, assuming that w1 and w2 are 0.5 in equation 1. The results of the evaluation process are shown in Table 7.

We argue that there must be a difference between the lexical, semantic, and clinical similarity. Lexical similarity depends on the level of textual similarity between the two concepts. Therefore, the lexical similarity $SIM_{lexical}$ (“*Chronic focal*”, “*Membranous*”) is equal to 0, and this is not accurate because both *197618004|chronic focal glomerulonephritis* and *77182004|membranous glomerulonephritis* are both derived from *20917003|chronic glomerulonephritis*. The semantic similarity adds some intelligence to this process; however, it has many limitations. We propose to handle these limitations by using the clinical similarity measurement. In clinical similarity, the $Sim_{Clinical}$ (“*kidney disease*”, “*kidney disease*”) < $Sim_{Clinical}$ (“*autosomal dominant focal segmental glomerulosclerosis*”, “*hyperfiltration focal segmental glomerulosclerosis*”). As a result, the three similarities are not equal regarding the accuracy, i.e., $Sim_{Lexical} \neq Sim_{Clinical} \neq Sim_{Semantic}$. Our proposed similarity measure takes into account the level of specificity of a concept that subsumes the two compared concepts and the level of commonality between the compared concepts. As a result, as shown in Table 7, the similarity Sim (*Type I*, *Type I*) = 1 because Type I and Type I are very specific in the ontology. The similarity Sim (*Acute*, *Chronic*) = 0.889 because these concepts are not specific; they contain many sub-concepts. Our algorithm is very sensitive to the level of similarity between the compared concepts.

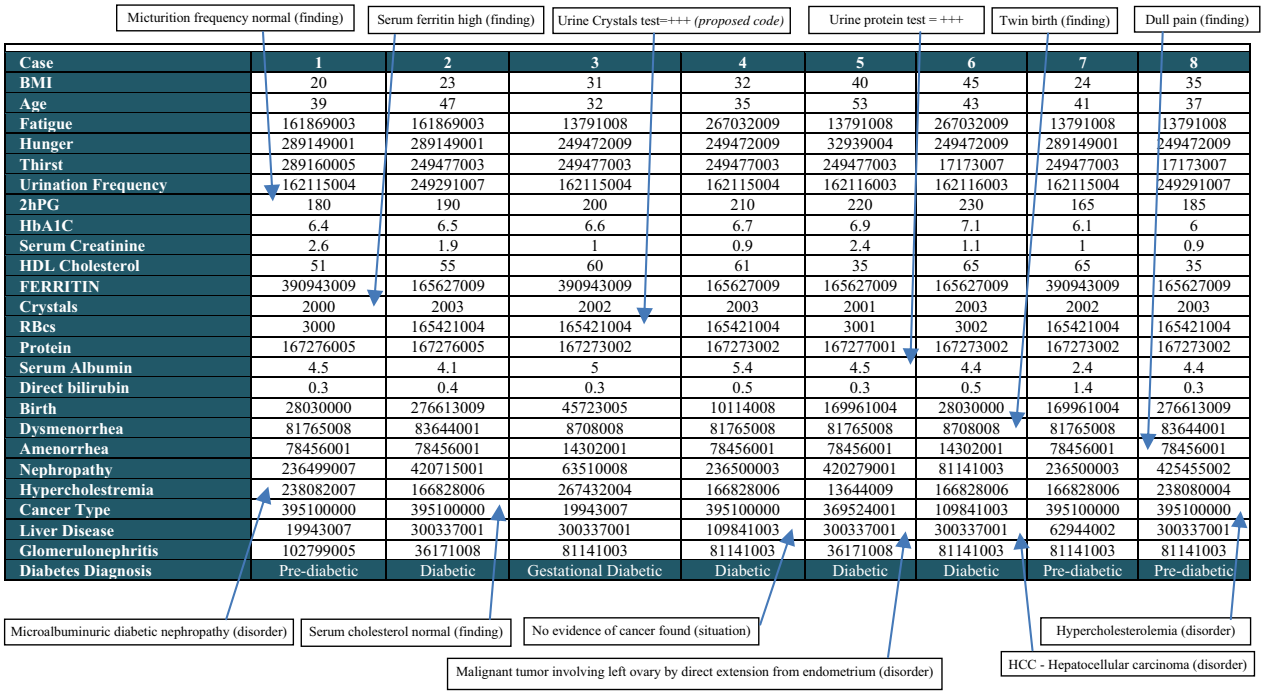
As shown in Table 7, *Fdeep_basic* and *Fdeep* do not take into account the depth of concepts from their LCA (i.e., the closeness between concepts) as in cases 7, 8. Moreover, *Cosine* and *Detail* do not account for the differences between concepts such as cases 5, 6. What is more, there are distributed inefficiencies as *Detail (Type I, Type I)* ≠ 1, *Cosine* (“*lipomatosis renis*”, *Uremia*) = 1, etc. On the other hand, the proposed similarity measure provides logically consistent results for all types of problems because it takes into account the depth of the compared concepts from their LCA, the differences between compared concepts, and their commonalities. As a result, the proposed OWL 2 ontology will add a great value for implementing an intelligent CDSS system for diabetes diagnosis.

Table 6 Sample coded data values that could not be encoded with SNOMED CT.

Term	Source Table	Comments
Vision (redness allergy)	Global symptoms	A concept is found to represent the normal vision as 45089002| normal vision (finding)|, blurred vision as 246636008|blurred vision – hazy|, Redness as 386713009| red color (finding)|, and diabetic retinopathy as 4855003| diabetic retinopathy (disorder)|. However, no codes are found for redness allergy
CA-125 (normal level)	Global symptoms	A concept is found to represent the abnormal value of CA-125 432519008| increased cancer antigen 125 (finding)|, but there is no concept of representing its normal value
AFP Serum (level normal of serum alpha-fetoprotein)	Global symptoms	A concept is found to represent the abnormal value of AFP Serum 399643001| serum alpha-fetoprotein level elevated (finding)|, but there is no concept of representing its normal value
Bilirubin (Nil, +, + +, + + +)	Urination symptom	The only concept we could find is 79706000|bilirubin (substance)|, which has meaning far from needed one
Pus (Nil, +, + +, + + +)	Urination symptom	The two found concepts are 11311000|pus (substance)| and 367646009|pus (morphologic abnormality)| and all have different meanings
Shrink kidney	Diagnosis	A possible SNOMED CT concept is 236448000| small kidney (disorder), but “small” does not convey the same meaning as “Shrink”
		...

6. Conclusion

This paper has tried to enhance the application and implementation of KI-CBR systems for diabetes diagnosis. As KI-CBR mainly relays on ontology, we utilized SCT standard for building domain background-knowledge ontology and encoded case-based knowledge. This paper proposed a clinical data encoding methodology. It applied this methodology to a diabetes diagnosis case-base dataset. The used standard



Case	1	2	3	4	5	6	7	8
BMI	20	23	31	32	40	45	24	35
Age	39	47	32	35	53	43	41	37
Fatigue	161869003	161869003	13791008	267032009	13791008	267032009	13791008	13791008
Hunger	289149001	289149001	249472009	249472009	32939004	249472009	289149001	249472009
Thirst	289160005	249477003	249477003	249477003	249477003	17173007	249477003	17173007
Urination Frequency	162115004	249291007	162115004	162115004	162116003	162116003	162115004	249291007
2hPG	180	190	200	210	220	230	165	185
HbA1C	6.4	6.5	6.6	6.7	6.9	7.1	6.1	6
Serum Creatinine	2.6	1.9	1	0.9	2.4	1.1	1	0.9
HDL Cholesterol	51	55	60	61	35	65	65	35
FERRITIN	390943009	165627009	390943009	165627009	165627009	165627009	390943009	165627009
Crystals	2000	2003	2002	2003	2001	2003	2002	2003
RBcs	3000	165421004	165421004	165421004	3001	3002	165421004	165421004
Protein	167276005	167276005	167273002	167273002	167277001	167273002	167273002	167273002
Serum Albumin	4.5	4.1	5	5.4	4.5	4.4	2.4	4.4
Direct bilirubin	0.3	0.4	0.3	0.5	0.3	0.5	1.4	0.3
Birth	28030000	276613009	45723005	10114008	169961004	28030000	169961004	276613009
Dysmenorrhea	81765008	83644001	8708008	81765008	81765008	8708008	81765008	83644001
Amenorrhea	78456001	78456001	14302001	78456001	78456001	14302001	78456001	78456001
Nephropathy	236499007	420715001	63510008	236500003	420279001	81141003	236500003	425455002
Hypercholestremia	238082007	166828006	267432004	166828006	13644009	166828006	166828006	238080004
Cancer Type	395100000	395100000	19943007	395100000	369524001	109841003	395100000	395100000
Liver Disease	19943007	300337001	300337001	109841003	300337001	300337001	62944002	300337001
Glomerulonephritis	102799005	36171008	81141003	81141003	36171008	81141003	81141003	81141003
Diabetes Diagnosis	Pre-diabetic	Diabetic	Gestational Diabetic	Diabetic	Diabetic	Diabetic	Pre-diabetic	Pre-diabetic

Figure 10 A small fragment of the encoded case base.

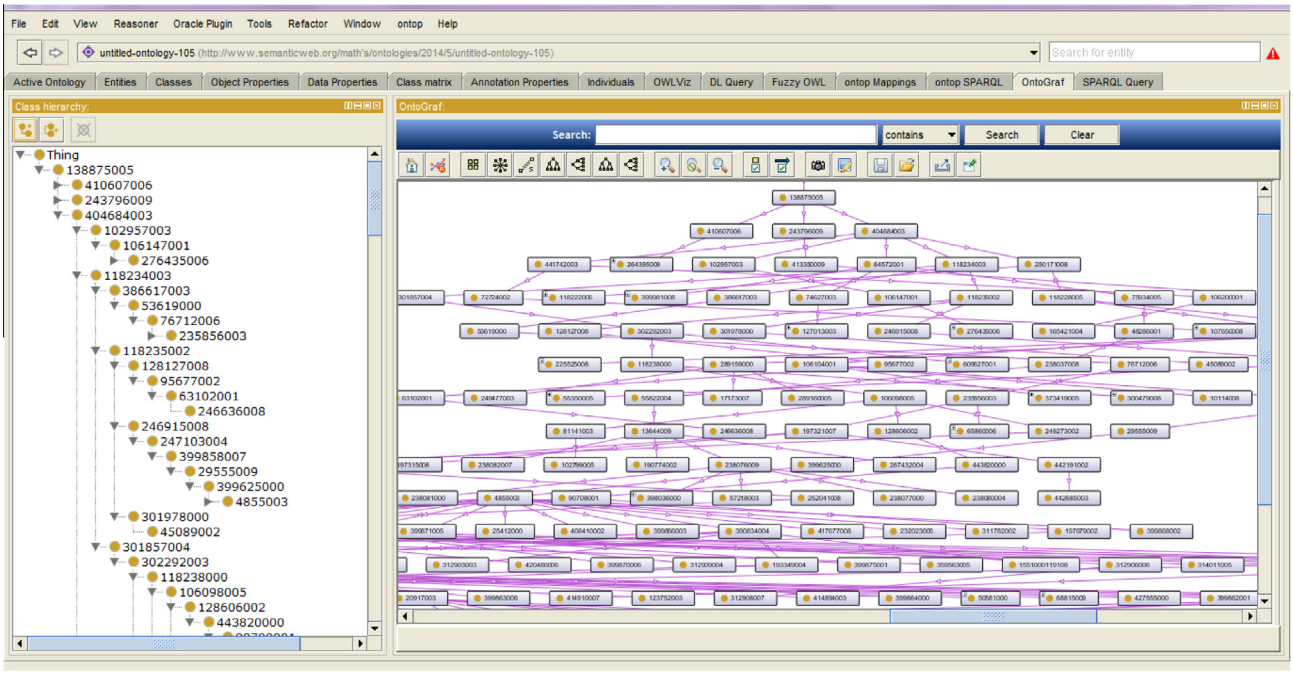


Figure 11 The resulting OWL 2 domain knowledge ontology.

ontology was SNOMED CT, which provides a concept coverage of ~75% for concepts in our dataset. The proposed methodology is suitable for encoding of any other clinical data in medical domains. Case-based textual medical terms have been encoded using the selected concepts. Other data types as numerical data have not been encoded because they will not enhance the retrieval algorithm. Moreover, the collected concepts IDs have been used to build an OWL 2 ontology using protégé 4.3. A set of customs codes has been proposed for the unmatched clinical terms. These codes have been added to the resulting ontology and proposed as extensions in the future releases of SNOMED CT. This ontology will be used as domain knowledge in a KI-CBR system for diabetes diagnosis. In the future, we will use the built ontology and the encoded case-based to build a CDSS for diabetes diagnosis using KI-CBR paradigm. The resulting case-based and

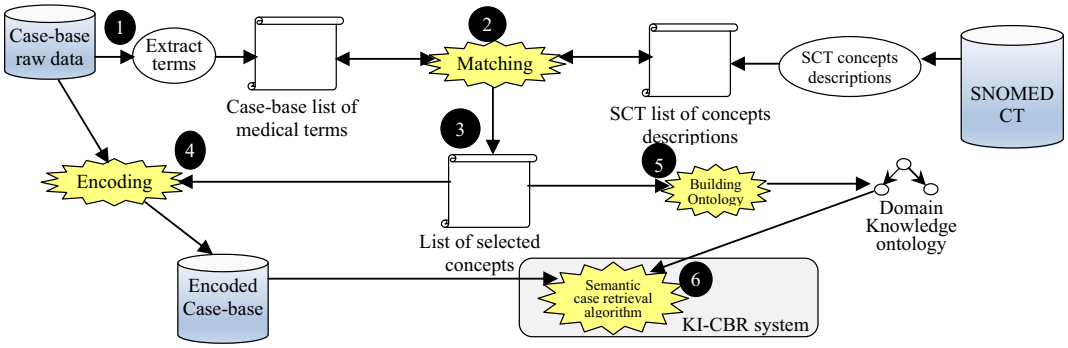


Figure 12 The overall encoding architecture.

Table 7 The comparison between JCOLIBRI semantic similarity methods and our proposed one.

Method	Fdeep_basic	Fdeep	Cosine	Detail	Proposed method	Case No.
Similarity						
Sim (Type I, Type I)	1	1	1	0.928	**1**	1
Sim ("lipomatosis renis", Uremia)	0	0	1	0.5	**0.11**	2
Sim (Cortical, Classical)	0	0	0.2	0.5	**0.04**	3
Sim (Chronic, Stage II)	0.57	0.66	0.82	0.875	**0.66**	4
Sim (Glomerulosclerosis, Type I)	0.286	0.286	0.436	0.75	**0.47**	5
Sim (Glomerulosclerosis, Acute)	0.286	0.5	0.577	0.75	**0.383**	6
Sim (Acute, Chronic)	0.429	0.75	0.75	0.833	**0.889**	7
Sim (Type I, Chronic)	0.429	0.429	0.654	0.833	**0.423**	8

ontology will enhance the case retrieval algorithm by making it more semantically intelligent.

Acknowledgment

The authors would like to thank Dr. Farid Badria, Prof. of Pharmacognosy, Department and head of Liver Research Lab, Mansoura University, Egypt; and Dr. Hosam Zaghloul, Prof. at Clinical Pathology Department, Faculty of Medicine, Mansoura University, Egypt, for their efforts in this work.

References

Abou Assali, A., Lenne, D., Debray, B., 2009. Case retrieval in ontology-based CBR systems. Adv. Artif. Intell. 5803, 564–571.

Agrawal, A., Elhanan, G., 2014. Contrasting lexical similarity and formal definitions in SNOMED CT: consistency and implications. J. Biomed. Inf. 47, 192–198.

Ahmadian, L., Cornet, R., de Keizer, N., 2010. Facilitating pre-operative assessment guidelines representation using SNOMED CT. J. Biomed. Inf. 43, 883–890.

Ahmadian, L. et al, 2011. The role of standardized data and terminological systems in computerized clinical decision support systems: literature review and survey. Int. J. Med. Inf. 80, 81–93.

Alexopoulos, P., Wallace, M., Kafentzis, K., Askounis, D., 2010. Utilizing imprecise knowledge in ontology-based CBR systems by means of fuzzy algebra. Int. J. Fuzzy Syst. 12 (1).

Andrews, J., Patrick, T., Richesson, R., Brown, H., Krischer, J., 2008. Comparing heterogeneous SNOMED CT coding of clinical research concepts by examining normalized expressions. J. Biomed. Inf. 41 (6), 1062–1069.

B2i Healthcare, 2015. SNOW OWL Browser. <http://b2i.sg> (Accessed 5 May 2015).

Barrett, N., Weber, J., Thai, V., 2012. Automated clinical coding using semantic atoms and topology. Proc. Comput. Based Med. Syst. (CBMS), 1–6.

Benson, T., 2009. Using SNOMED CT and HL7 together. In: Principles of Health Interoperability HL7 and SNOMED, 1st ed. Springer, pp. 267–280.

Bichindaritz, I., 2004. Mémoire: case based reasoning meets the semantic web in biology and medicine. Adv. Case Based Reasoning 3155, 47–61.

BioPortal, 2015. College of American Pathologists – National Health Service, SNOMED CT. <http://bioportal.bioontology.org/ontologies/SNOMEDCT?p=classes> (Accessed 15 May 2015).

Bodenreider, O., 2009. Using SNOMED CT in combination with MedDRA for reporting signal detection and adverse drug reactions reporting. AMIA Annu. Symp. Proc., 45–49.

Branden, M., Wiratunga, N., Burton, D., Craw, S., 2011. Integrating case-based reasoning with an electronic patient record system. Artif. Intell. Med. 51, 117–123.

Brandt, M., Rath, A., Devereau, A., Ayme, S., 2011. Mapping orphanet terminology to UMLS. Artif. Intell. Med. 6747, 194–203.

Campbell, J. et al, 1997. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity CPRI work group on codes and structures. J. Am. Med. Inf. Assoc. 4, 238–251.

Chiang, M., Hwang, J., Yu, A., Casper, D., Cimino, J., Justin, J., 2006. Reliability of SNOMED-CT coding by three physicians using two terminology browsers. AMIA Annu. Symp. Proc., 131–135.

Clinical Information Consultancy Ltd, 2015. CliniClue Xplore. <www.cliniclue.com> (Accessed 10 May 2015).

Dendani, N., Khadir, M., Guessoum, S., 2012. Use a domain ontology to develop knowledge intensive CBR systems for fault diagnosis. In: IEEE International Conference on Information Technology and e-Services, pp. 1–6.

Dolin, R.H., Spackman, K.A., Markwell, D., 2002. Selective retrieval of pre- and post-coordinated SNOMED concepts. Proc. AMIA Symp., 210–214.

El-Sappagh, S.H., Elmogy, M., Riad, A., Badria, F., Zaghlol, M., 2014a. EHR data preparation for case based reasoning construction. Adv. Mach. Learn. Technol. Appl. 488, 483–497.

El-Sappagh, S.H., El-Masri, S., Elmogy, M., Riad, R., Saddik, B., 2014b. An ontological case-base engineering methodology for diabetes management. J. Med. Syst. 38 (8), 1–14.

El-Sappagh, S.H., Elmogy, M., El-Masri, S., Riad, A., 2014c. A diabetes diagnostic domain ontology for CBR system from the conceptual model of SNOMED CT. In: The Proceeding of the IEEE Second International Conference on Engineering and Technology (ICET 2014), pp. 1–7.

El-Sappagh, S.H., Elmogy, M., Riad, A., 2015. A CBR system for diabetes mellitus diagnosis: case-base standard data model. Int. J. Med. Eng. Inf. 7 (3).

González, C., López, D.M., Blobel, B., 2013. Case-based reasoning in intelligent health decision support systems. Stud. Health Technol. Inf. 189, 44–49.

Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., Montmain, J., 2014. A framework for unifying ontology-based semantic similarity measures: a study in the biomedical domain. J. Biomed. Inf. 48, 38–53.

HL7, Health Level Seven International, 2015. HL7 Reference Information Model. <http://www.hl7.org /implement/standards/product_brief.cfm?product_id = 77> (Accessed 10 May 2015).

Højen, A., Gøeg, K., 2012. SNOMED CT implementation mapping guidelines facilitating reuse of data. Methods Inf. Med. 51 (6), 529–538.

IHTSDO, International Health Terminology Standards Development Organisation 2015a. SNOMED CT Technical Implementation Guide, January 2015 International Release.

IHTSDO, SNOMED CT Browsers, 2015b. <http://ihtsdo.org/fileadmin/user_upload/doc/browsers/browsers.html> (Accessed 23 May 2015).

IHTSDO, 2015c. SNOMED Clinical Terms: Developer Toolkit Guide, January 2013 International Release.

Jiaheng, L., Lin, C., Wang, W., Li, C., Wang, H., 2013. String similarity measures and joins with synonyms. In: Proceedings of the 2013 ACM International Conference on Management of Data, pp. 373–384.

Kim, H., Park, H., 2012. Development and evaluation of data entry templates based on the entity-attribute-value model for clinical decision support of pressure ulcer wound management. Int. J. Med. Inf. 81 (7), 485–492.

Köhler, S. et al, 2009. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am. J. Hum. Genet. 85 (4), 457–464.

Kooij, J., Goossen, W.T., Goossen, A.T., Jong, M., Beek, L., 2006. Using SNOMED CT codes for coding information in electronic health records for stroke patients. Stud. Health Technol. Inf. 124, 815–823.

Lamy, J., Tsopra, R., Venot, A., Duclos, C., 2013. A semi-automatic semantic method for mapping SNOMED CT concepts to VCM Icons. Stud. Health Technol. Inf. 192, 42–46.

Lau, F., Simkus, R., Lee, D. 2013. A methodology for encoding problem lists with SNOMED CT in general practice. In: Proceedings of the Third International Conference on Knowledge Representation in Medicine, pp. 97–103.

Lee, 2007. DHK. Reverse Mapping ICD-10-CA to SNOMED CT, UVic Master of Science Research Project Report.

Lee, D., 2014. The Science and Practice of SNOMED CT Implementation (Ph.D. thesis), University of Victoria.

Lee, C., Wang, M., 2011. A fuzzy expert system for diabetes decision support application. IEEE Trans. Syst. Man Cybern. B Cybern. 41 (1), 139–153.

Lee, D., Lau, F., Quan, H., 2010. A method for encoding clinical datasets with SNOMED CT. BMC Med. Inf. Decis. Making 10, 53.

Lee, D., Cornet, R., Lau, F., Keizer, N., 2013. A survey of SNOMED CT implementations. J. Biomed. Inf. 46, 87–96.

Liu, J., Lane, K., Lo, E., Lam, M., Truong, T., Veillette, C., 2010. Addressing SNOMED CT implementation challenges through multi-disciplinary collaboration. Stud. Health Technol. Inf. 160 (Pt 2), 981–985.

Lu, Y., Li, Q., Xiao, W., 2013. Case-based reasoning for automated safety risk analysis on subway operation: case representation and retrieval. Saf. Sci. 57, 75–81.

Melton, G., Parsons, S., Morrison, F., Rothschild, A., Markatou, M., Hripcsak, G., 2006. Inter-patient distance metrics using SNOMED CT defining relationships. J. Biomed. Inf. 39 (6), 697–705.

Mougin, F., Dupuch, M., Grabar, N., 2011. Improving the mapping between MedDRA and SNOMED CT. Artif. Intell. Med. 6747, 220–224.

NLM, 2015. U.S. National Library of Medicine. Stopwords, <http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/ 020_170.html> (Accessed 10 May 2015).

Rasmussen, A., Rosenbeck, K., 2011. SNOMED CT implementation: implications of choosing clinical findings or observable entities. Stud. Health Technol. Inf. 169, 809–813.

Recio-García, J., Díaz-Agudo, B., González-Calero, P., 2014. The COLIBRI platform: tools, features and working examples. In: Montani, Stefania, Jain, Lakhmi (Eds.), . In: Successful Case-based Reasoning Applications-2, vol. 494. Springer Berlin Heidelberg, Heidelberg, Germany, pp. 55–85.

Rodríguez, C., Cáceres, J., Sicilia, M., 2011. Generating SNOMED CT subsets from clinical glossaries: an exploration using clinical guidelines. ENTERprise Inf. Syst. 221, 117–127.

Ryan, A., Eklund, P., Esler, B., 2007. Toward the interoperability of HL7 v3 and SNOMED CT: a case study modelling mobile clinical treatment. Stud. Health Technol. Inf. 129 (Pt 1), 626–630.

Sánchez, D., Batet, M., Isern, D., Valls, A., 2012. Ontology-based semantic similarity: a new feature-based approach. Expert Syst. Appl. 39, 7718–7728.

Silva, T., MacDonald, D., Paterson, G., Sikdar, K., Cochrane, B., 2011. Systematized nomenclature of medicine clinical terms (SNOMED CT) to represent computed tomography procedures. Comput. Methods Programs Biomed. 101 (3), 324–329.

Subirats, L., Ceccaroni, L., 2011. An ontology for computer-based decision support in rehabilitation. Adv. Artif. Intell. 7094, 549–559.

UMLS, 2015. Unified Medical Language System. UMLS Overview–Tutorial. <http://www.nlm.nih.gov /research/umls> (Accessed 3 May 2015).

Wang, Y., Patrick, J., Miller, G., O'Halloran, J., 2014. Linguistic mapping of terminologies to SNOMED CT. In: First European Conference on SNOMED CT organized by the Network of Excellence Semantic Mining. <http://www.hiww.org/smcs2006/proceedings/6WangSMCS2006final.pdf> (Accessed 22 November 2014).

Wasserman, H., Wang, J., 2003. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. In: AMIA 2003: Annual Symposium Procedure, pp. 699–703.

WordNet, 2015. Princeton University, WordNet a Lexical Database for English. <http://wordnet.princeton.edu/wordnet> (Accessed 10 May 2015).

Wu, Z., Palmer, M., 1994. Verb semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. Morgan Kaufmann Publishers, San Francisco, California, pp. 133–138.

Zadeh, L., 2003. From search engines to question-answering systems the need for new tools. In: The 12th IEEE International Conference of Fuzzy Systems, vol. 2, pp. 1107–1109.

Zidi, A., Bouhana, A., Abed, M., Fekih, A., 2014. An ontology-based personalized retrieval model using case base reasoning. Proc. Comput. Sci. 35, 213–222.