Connecting Medical Informatics and Bio-Informatics
R. Engelbrecht et al. (Eds.)
ENMI, 2005

Issues in the Classification of Disease Instances with Ontologies

Anita Burgun[a], Olivier Bodenreider[b], Christian Jacquelinet[c]

[a] *EA 3888, Laboratoire d'Informatique Médicale, Faculté de Médecine, IFR 140, Université de Rennes I, France*
[b] *U.S. National Library of Medicine, Bethesda, MD, USA*
[c] *Etablissement français des Greffes, Paris, France*

Abstract

Ontologies define classes of entities and their interrelations. They are used to organize data according to a theory of the domain. Towards that end, ontologies provide class definitions (i.e., the necessary and sufficient conditions for defining class membership). In medical ontologies, it is often difficult to establish such definitions for diseases. We use three examples (anemia, leukemia and schizophrenia) to illustrate the limitations of ontologies as classification resources. We show that eligibility criteria are often more useful than the Aristotelian definitions traditionally used in ontologies. Examples of eligibility criteria for diseases include complex predicates such as ' x is an instance of the class C when at least n criteria among m are verified' and 'symptoms must last at least one month if not treated, but less than one month, if effectively treated'. References to normality and abnormality are often found in disease definitions, but the operational definition of these references (i.e., the statistical and contextual information necessary to define them) is rarely provided. We conclude that knowledge bases that include probabilistic and statistical knowledge as well as rule-based criteria are more useful than Aristotelian definitions for representing the predicates defined by necessary and sufficient conditions. Rich knowledge bases are needed to clarify the relations between individuals and classes in various studies and applications. However, as ontologies represent relations among classes, they can play a supporting role in disease classification services built primarily on knowledge bases.

Keywords:
Ontologies; Medical domain; Knowledge bases; Knowledge representation; Classification; OpenGALEN; SNOMED CT; Diagnosis; Patient records.

1. Introduction

Biomedical ontology aims to study the kinds of entities (i.e., substances, qualities and processes) in reality which are of biomedical significance and the relations among them. One role played by ontologies is to organize data (instances) according to classificatory principles reflecting a theory of the domain. In the clinical domain, disease instances, i.e., the particular forms of diseases (as described in the records of patients suffering from these diseases) are expected to be associated with the relevant disease categories represented in biomedical ontologies [1]. One role of the classifiers (e.g., Racer, FaCT) developed for ontologies represented in Description Logics (DL)-

based systems is precisely the automatic classification of instances. From an operational viewpoint, an ontology can be seen as a set of concepts or types that are organized in such a way that knowledge can be processed automatically by computers. To achieve that goal, the underlying structure must be "well-formed" and based on formal criteria, and the semantics must be explicit and consistent. Definitions in ontologies often embrace the Aristotelian model of genus and differentiae. In practice, definitions rely on a set of primitive terms which are not defined but rather given as such, and a set of interconcept relationships whose nature must be explicitly stated. Clarity was already mentioned by Gruber [2]: "Definitions should be effective […]. Where possible, a complete definition (a predicate defined by necessary and sufficient conditions) is preferred over a partial definition (defined by only necessary or sufficient conditions)".

Besides terminological issues and clinical convention in naming diseases (some phrases do not literally mean what they say) [3,4], getting an effective definition can be difficult. Applied to diseases, the classificatory principles and properties represented in ontologies may not be sufficient to classify instances. In this paper, we discuss two major limitations of current biomedical ontologies, preventing them from effectively classifying disease instances. First, some properties are too general to be useful in an operational setting (e.g., "presence of abnormal cells", where "abnormal cells" refers to an abstraction or interpretation rather than a piece of information present in medical records). Second, the definition of medical conditions is not always sharp. In practice, diagnostic criteria often include probabilistic components rather than the binary (presence/absence) elements recorded in most ontologies. For these reasons, we argue that biomedical ontologies should not be expected to provide operational definitions of diseases. Rather, they can be used as supporting resources for disease classification services developed primarily on systems including probabilistic and statistical knowledge as well as rule-based criteria. We use three examples (anemia, leukemia and schizophrenia) to illustrate the limitations of ontologies as instance classification resources.

2. Anemia

Anemia is defined as "a reduction below normal in the concentration of erythrocytes or hemoglobin in the blood […]; it occurs when the equilibrium is disturbed between blood loss (through bleeding or destruction) and blood production" (Webster 30th ed.). This definition is not operational as it contains a reference to normal concentrations. The same source provides reference intervals for the interpretation of laboratory tests (p. 2182), including references for erythrocyte and hemoglobin concentration, with distinctions for several population groups, shown in Table 1. The definition, complemented by reference intervals, makes it clear that properties such as age and gender are required for the interpretation of blood concentrations of erythrocytes and hemoglobin. Interpreting "or" as a true alternative, a low concentration of either entity is sufficient to diagnose anemia.

Biomedical ontologies such as OpenGALEN[1] and SNOMED CT[2] differ from the dictionary definition in that they only refer to the concentration of either hemoglobin (OpenGALEN, see Figure 1) or erythrocytes (SNOMED CT, see Figure 2). However, all definitions have in common to refer to abnormally low concentrations.

[1] http://www.opengalen.org/
[2] http://www.snomed.org/

• Haemoglobin Concentration
• Low Concentration
• Low Biochemical Measurement
• Haematological Disorder
• Pathological Concentration

parent classes

(Haemoglobin Concentration
which < *has Quantity* (
Level which < *has Magnitude* low Level >
) >)

Figure 1 – Representation of Anaemia in OpenGALEN

* ***Is a***
Red blood cell disorder
* ***Finding site***
Hematopoietic system structure
* ***Finding site***
Erythrocyte
* ***Has definitional manifestation***
Erythropenia

Figure 2 – Representation of Anemia in SNOMED CT

Table 1 – Reference intervals for the interpretation of blood concentrations of erythrocytes and hemoglobin (conventional units)

Population group	Erythrocytes (million/mm^3)	Hemoglobin (g/dl)
Males	4.6-6.2	13.0-18.0
Females	4.2-5.4	12.0-16.0
Children *	4.5-5.1	11.2-16.5
Newborns	-	16.5-19.5

* (varies with age)

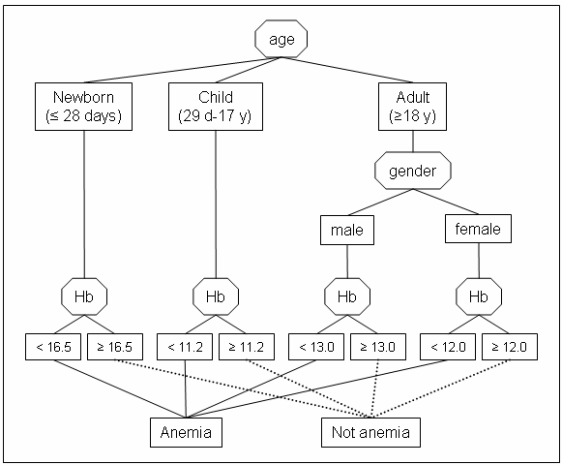

Figure 3 – Diagnostic tree for anemia based on age, gender and hemoglobin (Hb) concentration values

Not more than the textual definitions of dictionaries, the formal definitions of biomedical ontologies provide an operational definition of anemia, required for identifying anemic patients. A rule-based definition of anemia can be proposed instead, which would first consider *age* and then *gender* (for adults) before comparing *hemoglobin* (or erythrocytes) *concentration* values to the lowest bound of the reference interval for the corresponding population group. A diagnostic tree based on such rules (restricted to hemoglobin) is shown in Figure 3.

3. Leukemia

Leukemia is defined as 'a progressive, malignant disease of the blood-forming organs, characterized by distorted proliferation and development of leukocytes and their precursors in the blood and bone marrow' (Dorland, 28th ed). As for anemia, the textual definition uses references to abnormality but does not indicate precisely what kind of information is necessary to establish

the diagnosis of leukemia. Examples of eligibility criteria for *acute leukemia* include the presence of at least 30% blasts (immature hematopoietic cells) in the bone marrow[3].

In both OpenGALEN (Figure 4) and SNOMED CT (Figure 5), leukemia is represented as a neoplastic disease with location to the hematological/lymphatic system. Children of *leukemia* in SNOMED CT include kinds of leukemia by cell type (e.g., *Myeloid leukemia*), by the degree of cell differentiation (e.g., *Acute leukemia*) and by the existence of an active disease (e.g., *Leukemia in remission*). Like the textual definitions above, these formal definitions fail to provide sufficient informations for classifying instances of leukemias based on patient data. Moreover, none of the classificatory criteria are represented explicitly, which makes it difficult – if at all possible – for users to automatically process knowledge.

Additionally, the example of *leukemia* raises two interesting issues. First, biomedical knowledge is evolving rapidly, often leading to changes in the theory of the domain. For example, knowledge of gene mutations related to leukemia may change the classification of leukemia (e.g. MLL rearrangements are correlated with poor prognosis in childhood acute myeloid leukemia[4]). Second, the categorization of *leukemia in remission* as a kind of *leukemia* in SNOMED CT is problematic. Remission means that the disease seems under control, but *leukemia in remission* has few of the characteristics of the active disease of which it should therefore not be considered a subtype. For example, the percentage of blasts in the bone marrow required to assess remission is generally less than 5% (as opposed to more than 30% in the acute form of the disease). Moreover, if *remission* is a valid classificatory principle, it is also surprising that a *remission* subclass is defined for only a few diseases (including some cancers) in SNOMED CT.

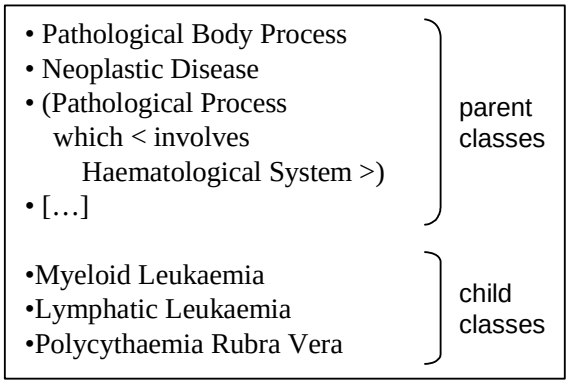

Figure 4 – Representation of Leukemia in OpenGALEN

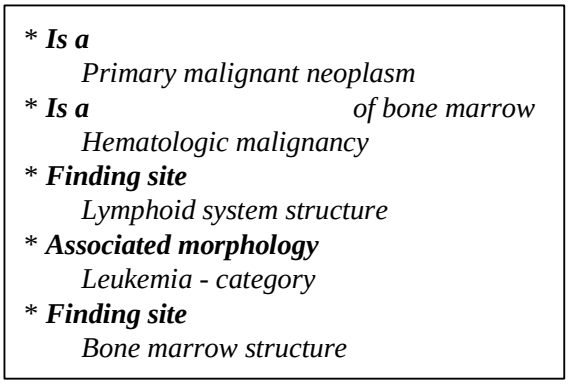

Figure 5 – Representation of Leukemia in SNOMED CT

4. Schizophrenia

The Diagnostic and Statistical Manual of Mental Disorders - 4th Edition (DSM-IV) is the main diagnostic reference thesaurus in psychiatry. For decades, psychiatrists have introduced in this classification diagnostic criteria for the most common mental disorders. Such criteria are needed to enable the consistent diagnosis of mental diseases in various contexts, including patient records, clinical trials and public health studies. Here is the definition provided for *schizophrenia* in DSM-IV: "a group of psychotic disorders characterized by disturbances in thought, perception, affect, behavior, and communication that last longer than 6 months". The following clinical

[3] http://www.cancer.org/docroot/cri/content/cri_2_4_3x_how_is_adult_acute_leukemia_diagnosed_57.asp?sitearea=&level=
[4] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=omim

criteria are added to this general definition in order to create an operational definition: "For a material part of at least one month (or less, if effectively treated) the patient has had two or more of: (i) Delusions; (ii) Hallucinations; (iii) Speech that shows incoherence, derailment or other disorganization; (iv) Severely disorganized or catatonic behavior." The example of *schizophrenia* illustrates the following three issues:

- The eligibility criteria for schizophrenia in DSM-IV correspond to complex predicates: (i) at least n criteria among m must be verified; (ii) criteria may differ according to some condition: symptoms must last at least one month if not treated, but less than one month, if treated; (iii) exclusion criteria are needed for differential diagnosis (schizophrenia is not directly caused by a general medical condition or the use of substances). Such complex predicates are more likely to be present in knowledge bases designed to assist in diagnosis than in ontologies. In fact, the definitions of *schizophrenia* in OpenGALEN and SNOMED CT only specify that it is a kind of *Psychotic disorder*. Possible qualifiers (e.g., *acute*) are also listed in SNOMED CT.
- As it is often the case for diseases, references to normality are used in DSM-IV definitions. For example, when referring to "bizarre delusions", a typical example is provided ("being abducted in a space ship from the sun") rather than a formal definition. The use of typical examples in lieu of formal definition is not specific to psychiatric disease. For example, typical instances of skin lesion provide effective operational definitions in dermatology. Such alternatives to Aristotelian definition are usually not found in ontologies.
- The onset of schizophrenia may include a schizophrenic prodrome, i.e., a temporal part of a disease in which the manifestations are milder, preceding the main phase of the disease. In medical ontologies, schizophrenic prodrome is expected to be represented as a temporal part of schizophrenia, not as its subclass as it is the case in SNOMED CT.

5. Discussion

5.1. Knowledge bases vs. ontologies

Biomedical ontologies represent classes of diseases based on broad classificatory principles such as etiology and location. The classification is refined using ad hoc principles such as organism for infectious diseases or pharmacologic action for drug poisoning. While useful for organizing biomedical knowledge and supporting inference (i.e., reasoning about classes), ontologies do not generally contain the detailed information required for classifying disease instances and reasoning about them. In contrast, the methods used in knowledge bases designed for solving problems (e.g., rule-based systems, probabilistic systems) are more suitable for storing and processing the diagnostic information required for the classification of disease instances. For example, rules for the definition of anemia are illustrated by the decision tree presented in Figure 3. On the other hand, knowledge bases are created to support specific tasks, such as diagnosis. The relation *manifestation of* between a finding and a disease differs from the "necessary and sufficient conditions" in operational definitions. For example, the presence of *fever* and *necrotizing tonsillitis* would suggest the diagnosis of leukemia but are unlikely to be mentioned in its definition. Furthermore, ontologies of diseases in decision-support systems are hierarchies designed for specific application rather than formal reference ontologies that can serve any purpose. Finally, part of the knowledge needed to classify diseases is not represented in standard clinical decision-support systems, but rather in textbooks of medicine or in clinical genomics resources such as OMIM.

5.2. Applications to data integration

Eligibility criteria complement Aristotelian definitions as they clarify the meaning of the relation "is an instance of" and, in particular, make it distinct from the "isa" relation. Such criteria are required for the consistent classification of disease instances in clinical information systems. As a consequence, they also contribute to data integration by enabling disease instances recorded in heterogeneous systems to be reliably linked to disease ontologies. In practice, eligibility criteria may be linked to clinical databases, enabling the corresponding predicates to be instantiated with the clinical information stored in patient records (e.g., *hemoglobin concentration*, *age* and *gender* to classify instances of *anemia*). Moreover, as illustrated by the notions of remission (existence of a class *disease in remission*) and prodrome (existence of a class *disease prodrome*), further efforts are needed to clarify how the temporal dimension can be used as an organizing principle in ontologies and connected with information about the course of diseases in patients. A better representation of the temporal dimension of diseases would also contribute to data integration, for example by helping to distinguish between active diseases and diseases in remission.

5.3. Toward instance classification services

Classifying disease instances, i.e., linking patient records to disease classes, is a complex process requiring several types of resources. This finding is consistent with the characterization of resources proposed in [5], especially the distinction between terminology models (ontologies) and inference models (knowledge bases). While knowledge bases contain the predicates necessary to establish a diagnosis from phenotypic and genotypic criteria, these criteria may not be directly applicable to the data found in patient records. For example, equivalent criteria may be expressed in several different ways and ontologies can help bridge the semantic gap between knowledge bases and patient data. Along the same lines, there is a need for data expressed in one unit system to be compared to references expressed in another unit system (e.g., between traditional and international unit systems). Rather than the product of one resource, classifying disease instances can be thought of as a service based on both ontologies and knowledge bases. This view is analogous to that, promoted by GALEN, of terminology services based on ontologies [6].

6. References

[1] Cimino JJ. From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. *J Am Med Inform Assoc*. 2000 May-Jun;7(3):288-97.

[2} Gruber, T.R Toward Principles for the Design of Ontologies Used for Knowledge Sharing, *Int. Journal of Human-Computer Studies*, 1995, Vol. 43, pp.907-928.

[3] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf Med*. 1998 Nov;37(4-5):394-403.

[4] Rector AL. Clinical terminology: why is it so hard? *Methods Inf Med*. 1999 Dec;38(4-5):239-52.

[5] Rector AL. The interface between information, terminology, and inference models. Medinfo. 2001;10(Pt 1):246-50.

[6] Rector AL, Solomon WD, Nowlan WA, Rush TW, Zanstra PE, Claassen WM. A Terminology Server for medical language and medical information systems. *Methods Inf Med*. 1995 Mar;34(1-2):147-57..

Address for correspondence

Anita Burgun, EA 3888, DIM, CHRU Pontchaillou, 2, rue H Le Guilloux F-35033 Rennes Cedex, anita.burgun-parenthoine@univ-rennes1.fr