Issues in Integrating Epidemiology and Research Information in Oncology: Experience with ICD-O3 and the NCI Thesaurus

Anita Burgun[1], M.D., Ph.D., Olivier Bodenreider[2], M.D., Ph.D.
[1]EA 3888, School of Medicine, IFR 140, Rennes University, Avenue Pr Léon Bernard, 35043 Rennes Cedex, France
[2]U.S. National Library of Medicine, NIH, Bethesda, Maryland, USA

[1]anita.burgun@univ-rennes1.fr, [2]olivier@nlm.nih.gov

The integration of the International Classification of Diseases for Oncology (ICD-O) and the NCI Thesaurus (NCIT) is expected to facilitate the integration of epidemiology data (cancer registries) with basic and clinical research data. We evaluated the degree to which ICD-O and NCIT provide consistent representations of neoplasms. 1,550 concepts (515 for topography and 1,035 for morphology) are shared by ICD-O and NCIT. Only 366 relations (about 1%) between these topography and morphology concepts are shared between ICD-O and NCIT. Two relationships – Disease Has Primary Anatomic Site and Disease Has Associated Anatomic Site – representing the anatomical site of a disease account for about 78% of the 1,376 relations between shared topography and morphology concepts in ICD-O and NCIT. In addition to these two roles, nine other NCIT relationships are found between topography and morphology concepts. Several issues are discussed, including incomplete representations in NCIT, mapping issues, systematic polysemy, and the use of post vs. pre-coordinated terms. The methods proposed provide a framework for analyzing inconsistencies.

INTRODUCTION

Many countries operate and maintain population-based cancer reporting systems for epidemiological studies. For 25 years, the International Classification of Diseases for Oncology (ICD-O) has been the major standard for coding neoplasms. On the other hand, vocabularies such as the NCI Thesaurus (NCIT) play an important role in cancer research, especially with initiatives such as caBIG[1], in which large volumes of information are shared. The integration of epidemiology and research data is needed to correlate the results from clinical trials with the characteristics of diseases in populations (prevalence, survival, etc.). It presupposes compatibility between terminologies [1]. Both ICD-O and the NCIT are integrated in the NCI Metathesaurus in which concepts common to both terminologies are identified by the same unique identifier. The alignment of concepts from ICD-O and NCIT through the NCI Metathesaurus largely facilitates integration studies. However, in addition to sharing concepts, terminologies are also expected to provide a consistent representation of the domain. The two major elements characterizing neoplasms in ICD-O are topology and morphology. For example, as shown in Figure 1, the morphology concept *Renal cell carcinoma* is present in both ICD-O and NCIT. The corresponding topography in ICD-O is *Kidney*. Similarly, NCIT asserts the relation *Renal cell carcinoma Disease Has Primary Anatomic Site Kidney*. *Renal cell carcinoma* is thus represented consistently in ICD-O and NCIT.

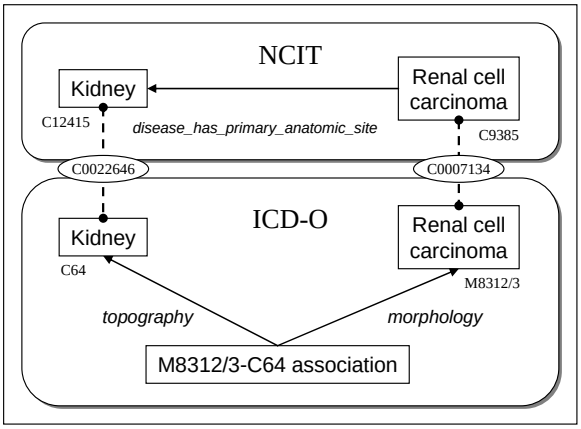

Figure 1. Renal cell carcinoma in ICD-O and NCIT

The objective of this study is to evaluate the degree to which ICD-O and NCIT provide consistent representations of neoplasms. More precisely, we investigate the two following issues: 1) *Concepts*: The topography and morphology concepts from ICD-O are all expected to be present in NCIT. 2) *Relations*: The associations between topography and morphology codes provided for some neoplasms in ICD-O are expected to be present in NCIT, the topography concept representing the anatomical site of the morphology concept. A secondary objective of this study is to identify issues in one representation based on the information provided by the other. While many studies have investigated ICD-O [e.g., 2, 3] and NCIT [e.g., 4, 5], this paper is, to our knowledge, the first attempt to analyze their consistency.

[1] http://cabig.cancer.gov/

BACKGROUND

The **International Classification of Diseases for Oncology** (ICD-O) is a dual classification for neoplasms[2], with coding systems for both topography and morphology [6].

The *topography axis* uses the same three- and four-character categories as ICD-10 for tumors (C00-C80), thereby providing greater site detail for nonmalignant tumors than is provided in ICD-10. Additional (non-ICD-10) topography codes are provided to identify the site of haematopoietic and reticuloendothelial tumors.

The *morphology axis* describes specific histologic cell types and their behavior. It provides 5-digit codes ranging from 8000/0 to 9989/3. The first four digits indicate the specific histological type. The fifth digit (after the slash) is a behavior code which indicates whether a tumor is malignant (/3), benign (/0), in situ (/2), or uncertain whether benign or malignant (/1).

Topography codes and morphology codes can be freely associated – post-coordinated, in terminology parlance – to describe neoplasms along these two dimensions. Additionally, ICD-O provides a list of pre-coordinated terms, i.e., fixed associations between topography and morphology terms. The version used in this study is ICD-O3, extracted from the NCI Metathesaurus. (ICD-O associations are represented in the file MRSAT.RRF, under the attribute "SAC").

The **NCI Thesaurus** (NCIT) is a public domain Description Logic-based terminology produced by the NCI that includes broad coverage of the cancer domain[3] [7, 8]. It has been designed to be used in systems supporting basic, translational, and clinical research. Its characterization of neoplasms is based on several roles, including *Disease Has Primary Anatomic Site*, and *Disease Has Associated Anatomic Site* (between diseases and anatomical entities), and *Disease Has Abnormal Cell* (between diseases and histologic entities). The version of the NCIT used in this study is 2006 10D, also extracted from the NCI Metathesaurus.

The **NCI Metathesaurus** integrates the public domain vocabularies of the UMLS Metathesaurus, of which it shares the basic organization. Specific vocabularies such as ICD-O are also present in the NCI Metathesaurus. As in the UMLS Metathesaurus, terms from different source vocabularies naming the same entity are given the same identifier, allowing for easy mapping between vocabularies. The version used in this study is 2006 10D.

METHODS AND RESULTS

Computing shared concepts between ICD-O and NCIT

In what follows, the NCI Metathesaurus is used to identify common concepts between ICD-O and NCIT. For example, as shown in Figure 1, *Kidney* in ICD-O (C64) and NCIT (C12415) are represented by the same NCI Metathesaurus concept (C0022646). Of note, one given ICD-O code can be associated with several concepts in the NCI Metathesaurus. For example, the morphology code 8201/2 is associated with both *Cribriform carcinoma in situ* (CL017913) and *Ductal carcinoma in situ, cribriform type* (CL053323).

There are 409 distinct topography codes and 1,091 morphology codes in ICD-O, represented in the NCI Metathesaurus by 1,085 and 1,419 concepts, respectively. Of these, 515 topography concepts (48%) and 1,035 morphology concepts (73%) are also present in the NCI Thesaurus.

However, since a given ICD-O code can be associated with several concepts, an alternative measure of overlap between ICD-O and NCIT would consider the number of ICD-O codes for which there is at least one associated concept present in NCIT. From this perspective, it appears that only 106 topography codes (10%) and 221 morphology codes (16%) from ICD-O are not associated with any NCIT concepts. For example, the topography code C75.4 *Carotid body* (C0007277) and the morphology code 8632/1 *Gynandroblastoma* (C0018413) are not present in the NCIT.

Assessing shared descriptions between ICD-O and NCIT

ICD-O associations. The associations provided by ICD-O generally link one morphology code to one topography code. Occasionally, morphology codes can be associated with several topography codes, or with a code whose fourth digit is left unspecified (e.g., C70._), indicating the existence of a code for a more specific site. In this case, we generated the associations between the morphology code and all 4-digit topography codes listed under this 3-digit code (here, under C70: C70.0, C70.1 and C70.9). We also generated the association between the morphology code and the 3-digit topography code. Analogously, topography codes are sometimes associated not with a 5-digit morphology code, but with a 3-digit morphology code (e.g., 926) or with a range of codes (e.g., 927-934). In this case, we generated the associations between the topography code and all 5-digit

[2] http://www.who.int/classifications/icd/adaptations/oncology/en/
[3] http://nciterms.nci.nih.gov/NCIBrowser/

morphology codes listed under the 3-digit code or range (e.g., 9260/0, 9261/3 and 9262/0 for 926). We also generated the association between the topography code and the 3-digit morphology code or range.
A total of 3,297 such associations were generated. We then associated topography and morphology codes with the corresponding NCI Metathesaurus concepts – often several concepts for a given code. 22,881 pairs of concepts were generated.

NCIT associations. In ICD-O, topography codes generally correspond to anatomical entities and morphology codes to neoplasms. For this reason, in NCIT, we used the roles defined between anatomical entities and diseases, restricting diseases to neoplasms by selecting only those concepts whose semantic type is *Neoplastic Process*. In practice, we used the roles *Disease Has Primary Anatomic Site* and *Disease Has Associated Anatomic Site*. We found 19,028 such relations involving 369 distinct anatomical entities and 6,330 neoplasms (among the 8,295 neoplasm concepts in NCIT).

Table 1. Characterization of the pairs of topography and morphology concepts from ICD-O (a) and NCIT (b). The three letters refer to the topography concept, the morphology concept, and the association, respectively. (I: specific to ICD-O, N: specific to NCIT, B: common to both)

(a) ICD-O			(b) NCIT		
B B B	366	(1.6%)	B B B	366	(1.9%)
B B I	8,725	(38.1%)	B B N	486	(2.6%)
B I I	2,663	(11.6%)	B N N	8,472	(44.5%)
I B I	8,591	(37.5%)	N B N	1,096	(5.8%)
I I I	2,536	(11.1%)	N N N	8,608	(45.2%)
Total	22,881	(100%)	Total	19,028	(100%)

Shared associations. After transforming the associations between topography and morphology into a common representation, i.e., pairs of NCI Metathesaurus concepts, we simply compute the intersection between the two sets of pairs of concepts. Quite surprisingly, only 366 pairs of concepts are shared between ICD-O and NCIT, i.e., roughly 1% of the pairs for ICD-O or NCIT. In fact, 495 of the 3,297 associations (15%) between topography and morphology codes in ICD-O are represented by (at least) one pair of concepts associated in NCIT.
We characterized every pair in ICD-O and in NCIT with information about the origin of the topography concept, the morphology concept and the association. In each case, we examined whether these elements were present in ICD-O only (I), NCIT only (N) or both (B). For example, the 366 pairs common to ICD-O and NCIT were characterized as "BBB", because both concepts and the association were shared by both terminologies. Our findings are summarized in Table 1.

Characterizing relations between shared concepts

We analyzed the roles (relationships) in NCIT between the 515 topography concepts and 1,035 morphology concepts from ICD-O also present in NCIT.

NCIT relations. We first considered only those relations asserted in NCIT. The two relationships examined earlier – *Disease Has Primary Anatomic Site* and *Disease Has Associated Anatomic Site* – representing the anatomical site of a disease account for about 78% of the 1,376 relations. Of note, some ICD-O codes are associated with a large number of disease terms. For example C42.1 *Bone marrow* (C0005953) is related to 70 different concepts through *Disease Has Primary Anatomic Site*, including *Chronic Myeloid Leukemia*, *Aggressive NK-Cell Leukemia*, and *Refractory Anemia with Ringed Sideroblasts*. In addition to these two roles, nine other NCIT relationships are found between topography and morphology concepts. Of particular interest are the roles *Disease Has Abnormal Cell* and *Disease Excludes Abnormal Cell*. Only a small number of concepts from ICD-O are involved in these relations, primarily M-8001/3 *Tumor cells, malignant* (C0334227) and M8001/1 *Tumor cells, NOS* (C0431085).

Relations from other source vocabularies. We extracted all the relations between topography and morphology concepts shared by ICD-O and NCIT, regardless of their semantics or origin. Twenty-three distinct semantic relationships (forty six when inverse relations are taken into account) are represented between the concepts shared by ICD-O and NCIT. Eleven of these are NCIT roles, the other being mostly SNOMED CT relationships. For example, 8500/2 *Intraductal adenocarcinoma, noninfiltrating, NOS* (C0007124) is associated with *Breast* (C50 in ICD-O) (C0006141) through the role *Finding Site*, the relation coming from SNOMED CT.

DISCUSSION

A framework for identifying and analyzing inconsistencies between ICD-O and NCIT

The analysis of concepts and relations shared between ICD-O and NCIT provides a framework for identifying and characterizing potential inconsistencies. In this section, we review some examples from the major categories of ICD-O concepts and relations not found in NCIT. (Of course, the same framework could be used to review those concepts and relations present in NCIT and not in ICD-O, using "BBN"-"NNN" instead of "BBI"-"III").

Incomplete representations in NCIT are revealed by ICD-O associations for which both concepts are present in NCIT, but not associated (listed as "BBI"

in our classification). For example, the topography associated with *Chromophobe adenoma* (8270/0) in ICD-O is *Pituitary gland* (C75.1). While both concepts are present in NCIT, no relation is asserted between them. In fact, no anatomical site is specified in NCIT for *Chromophobe adenoma* (C2857). Of note, the role *Disease Has Abnormal Cell* is filled with *Neoplastic cell*, confirming the underspecification of this representation.

Impedance mismatch for topography can be revealed by associations classified as "BBI" also. For example, the morphology code *Adenolymphoma* (8561/0) is associated with the topography codes C07._ and C08._, referring to the instantiation of *Parotid gland* (C07) and *Other and unspecified major salivary glands* (C08). In NCIT, *Adenolymphoma* (C2854) is appropriately associated with *Salivary gland*. While *Salivary gland* in NCIT corresponds to the group of entities referred to in ICD-O by C07 and C08, they are represented by different concepts in the NCI Metathesaurus and no shared associations can be found. In other words, unlike ICD-O, NCIT does not represent this neoplasm in specific salivary glands, such as the parotid gland. Rather than a contradiction, this example illustrates impedance mismatch between ICD-O and NCIT. The existence of hierarchical relations in the NCIT could help bridge this difference. Implicitly, the association of *Adenolymphoma* with *Parotid gland* can be inherited from the association between *Adenolymphoma* and *Salivary gland* through the following hierarchical relations in NCIT: *Parotid gland isa Major salivary gland* and *Major salivary gland isa Salivary gland*.

Missing (mapping for) morphology concepts ("BII" and "III"). The morphology concepts 8632/1 *Gynandroblastoma* (C0018413) and 8580/0 *Benign thymoma* (C0040101) cannot be directly mapped to NCIT through the NCI Metathesaurus, because the corresponding concepts in NCIT map to different concepts in the Metathesaurus. In fact, for these neoplasms, NCIT represents either more specific concepts (e.g., *Ovarian gynandroblastoma*) or more generic concepts (e.g., *Thymoma*). In both cases, the morphology concept in NCIT is associated with the same topography concept as its equivalent in ICD-O.

The NCI experts consider that *Gynandroblastoma* is a synonym for *Ovarian gynandroblastoma*, as it only occurs in the ovary, and record these two terms as synonymous names for the NCIT concept C3072. Conversely, like SNOMED CT, the NCI Metathesaurus distinguishes between *Gynandroblastoma* and *Ovarian Gynandroblastoma* and records them as distinct concepts. *Benign thymoma* simply is an obsolete term in NCIT.

Finding a mapping between morphology concepts in ICD-O and NCIT is possible in these two cases, but the hierarchical relations between *Ovarian gynandroblastoma* and *Gynandroblastoma* and between *Benign thymoma* and *Thymoma* come from other source vocabularies of the NCI Metathesaurus than ICD-O and NCIT.

Missing (mapping for) topography concepts ("IBI" and "III"). With only 409 topography codes (1,085 concepts), the ICD-O vocabulary for topography is significantly smaller than NCIT's (about 2,400 anatomical entities excluding the subcellular level). Yet, many anatomical entities from ICD-O cannot be found in NCIT, including *Skin of forehead*, *Sublingual gland duct* and *Anterior wall of stomach*. In some cases, the missing concepts are groupings of anatomical entities specific to ICD-O (e.g., *"Long bones of upper limb, scapula and associated joints"*). And while there are only 106 topography codes (26%) from ICD-O for which no concept is found in NCIT, the proportion of topography concepts from ICD-O shared by NCIT is less than 50% (515 concepts). Moreover, many of these topography concepts are associated with a large number of morphology codes (e.g., 95 different codes for *Skin of forehead*), contributing to the limited number of shared relations observed.

Terminology and knowledge representation issues

Pre-coordination vs. post-coordination. 1,550 concepts (515 for topography and 1,035 for morphology) are shared by ICD-O and NCIT. While these concepts can be post-coordinated in ICD-O to describe neoplasms, ICD-O also provides pre-coordinated terms (i.e., specific associations). Only 366 of these associations are present in NCIT. One potential use of these associations would be to identify all possible morphology concepts for a given topography concept (or the other way around) though the role *Disease Has Primary Anatomic Site*. In fact, only 632 such associations are present (and only 119 associations through *Disease Has Abnormal Cell*). In practice, it would not be possible to get a complete list of neoplasms for a given topography or morphology.

Morphology vs. disease. The ICD-O axes topography and morphology are in essence not independent. The morphology properties of a neoplasm, including its cell type, depend on the origin of the tumor, which is coded as topography. Moreover, the word senses associated with cancer diseases and tumors are related in systematic and predictable ways, leading to systematic polysemy. For example, *Non invasive ductal breast carcinoma* may be understood as either morphology or disease. Similarly, because of systematic polysemy, a relation *has associated morphology* (from SNOMED CT) is asserted between *Non invasive ductal breast carcinoma* and itself (reflexive relation). Moreover, *Intraductal adenocarcinoma*,

noninfiltrating, NOS is a synonym of *Non invasive ductal breast carcinoma* in the NCI Metathesaurus. As a result, it is associated with *Breast* through the role *Finding Site* in SNOMED CT

Inconsistent categorization. Surprisingly, 21 morphology concepts from ICD-O are not categorized as *Neoplastic Process* in the NCI Metathesaurus. While this is justified in a few cases (e.g., *Neoplastic Cell*), most seem to be errors (e.g., *Malignant Lymphoma, Non-Cleaved Cell Type*). Several concepts also have a different semantic type in NCIT and the NCI Metathesaurus (e.g., *Refractory Anemia with Excess Blasts*, categorized as *Disease or Syndrome* in the NCI Metathesaurus and *Neoplastic Process* in NCIT). These discrepancies can result in integration issues and cause errors in applications. For example, the number of neoplasms associated with *Bone marrow* (C0005953) through *Disease Has Primary Anatomic Site* in NCIT is not 70 as shown earlier, but 67 when the disease concepts are restricted to the semantic type *Neoplastic Process* in the NCI Metathesaurus.

Applications

Quality assurance in terminologies. The framework presented above can be used for quality assurance purposes. Inconsistency is generally indicative of inaccurate or missing relations in either terminology or both. It can also reveal misalignment between the terminologies (here, inaccurate association with a NCI Metathesaurus concept). In addition to helping identify problems, this framework can also suggest solutions. For example, the topography concept *Pituitary gland* in ICD-O can be automatically proposed as the filler for the role *Disease Has Primary Anatomic Site* in NCIT in the case of the incomplete representation of *Chromophobe adenoma*. presented earlier.

Generalization: Knowledge triangulation. While some guidelines for ontology development recommend the creation of "orthogonal", nonoverlapping ontologies, the availability of multiple representations of the same domain enables the comparison of these representations. This comparison is facilitated when the concepts are already aligned in a system such as the NCI (or UMLS) Metathesaurus. Comparing shared relations between concepts can be seen as examining these relations from the perspective of several terminologies, i.e., knowledge "triangulation". The mapping of NCIT to ICD-O recently created by NCI can also be exploited for triangulating knowledge about neoplasms.

Because they play a central role in information integration, reference ontologies such as NCIT, developed as an authoritative, current classification of cancers, must be tested for consistency with the other terminologies used in the same domain, including legacy terminologies, such as ICD-O. Differences must be analyzed to distinguish between obsolete or inaccurate representations, on the one hand, and alternative, consistent views on reality, on the other.

Integrating epidemiology, research and clinical practice. In future work, we will also study the representation of neoplasms in SNOMED CT. Like ICD-O and NCIT, SNOMED CT terminology is part of the NCI Metathesaurus. Moreover, the morphology codes in ICD-O come from earlier versions of SNOMED, which guarantees a tight integration between these two sources. The integration of ICD-O, NCIT and SNOMED CT is expected to facilitate the integration of the domains covered by these terminologies: epidemiology (cancer registries), research (annotated data), and clinical practice (patient records).

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. Sherertz DD, Tuttle MS, Olson NE, Hsu GT, Carlson RW, Fagan LM, et al. Accessing oncology information at the point of care: experience using speech, pen, and 3-D interfaces with a knowledge server. Medinfo 1995;8 Pt 1:792-5
2. Clarke CA, Undurraga DM, Harasty PJ, Glaser SL, Morton LM, Holly EA. Changes in cancer registry coding for lymphoma subtypes: reliability over time and relevance for surveillance and study. Cancer Epidemiol Biomarkers Prev 2006;15(4):630-8
3. van der Sanden GA, Wesseling P, Schouten LJ, Teepen HL, Coebergh J. A uniform histological cluster scheme for ICD-O-coded primary central nervous system tumors. Neuroepidemiology 1998;17(5):233-46
4. Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods of Information in Medicine 2005;44(4):498-507
5. Kumar A, Smith B. Oncology ontology in the NCI thesaurus. In: Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science); 2005. p. 213-220
6. Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin M, et al., editors. International classification of diseases for oncology. 3rd ed. ed. Geneva: World Health Organization; 2000
7. de Coronado S, Haber MW, Sioutos N, Tuttle MS, Wright LW. NCI Thesaurus: using science-based terminology to integrate cancer research results. Medinfo 2004;11(Pt 1):33-37
8. Sioutos N, Coronado Sd, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. Journal of Biomedical Informatics 2007;40(1):30-43