D5: NCIt Discrepancy Assessment Summary Report

Date: September 29, 2017
Contributors: Nicole Vasilevsky, Jim Balhoff, Melissa Haendel, Chris Mungall

Introduction

A number of projects “cross-reference” NCIt classes, but these are often a) not vetted by NCIt; b) not kept current; c) not logically consistent; and d) not coordinated across communities. This fundamentally leads to difficulties in data integration and analytics, relegating the knowledge represented in NCIt to simple data tagging. To support data integration and cross-dataset analyses, and help the research community, we analyzed selected non-NCI sources that cross-reference NCIt. These included the Disease Ontology (DO), Online Mendelian Inheritance in Man (OMIM), Orphanet and MeSH. In addition, we worked with the NCI team to refine and update the ICD-O 3.1 mappings, with the goal of easing migration from ICD-O to NCIT, which is a more computable resource and more readily implemented in coding/metadata systems.

T2.1: Neoplasm Subset Assessment and T2.3: Algorithmic analysis logical inconsistencies

Methods

The kBOOM algorithm was run to generate a spreadsheet with precise mappings of the NCIT neoplasm branch in the Monarch Disease Ontology (MonDO). The spreadsheet is available in the Supplemental Folder (saved as T2.1Neoplasm_Subset_Assessment). A screenshot is displayed and description of the spreadsheet is described in Figure 1.

	A	B	C	D	E	F	G	H
1	**NCIt ID**	**NCIt class**	**relationship type**	**Ontology ID**	**Class**	**probability**	**Yes/No/Undecided**	**GitHub Link**
15	NCIT_C6017	Paranasal Sinus Adenocarcinoma	SubClassOf	DOID_0050619	paranasal sinus cancer	0.1963622082	No	e-ontology/issues/299
16	NCIT_C6018	Paranasal Sinus Mucoepidermoid Carcinoma	SubClassOf	DOID_0050619	paranasal sinus cancer	0.1963622082	No	e-ontology/issues/300
17	NCIT_C6019	Paranasal Sinus Adenoid Cystic Carcinoma	SubClassOf	DOID_0050619	paranasal sinus cancer	0.1963622082	No	e-ontology/issues/306
18	NCIT_C8193	Paranasal Sinus Squamous Cell Carcinoma	SubClassOf	DOID_0050619	paranasal sinus cancer	0.1963622082	No	e-ontology/issues/307

Figure 1: A screen capture of the kBOOM output, used to manually review the mappings. The spreadsheet is separated into four tabs, one for each disease ontology that was reviewed (Disease Ontology, DO; OMIM; Orphanet; and MeSH.) The columns represent the NCIT IDs (column A), classes (column B), the relationship type to the related ontology term (column C) in columns D and E, the ontology ID for the mapped ontology term (column D) and the label for the mapped ontology term (column E). kBOOM measures the probability that this relationship is precise (column F). For example, row 15 says NCIT_C6017 *Paranasal Sinus Adenocarcinoma* is a subclass of DOID_0050619 *paranasal sinus cancer* and kBOOM has assigned a low probability that this assertion is correct. We manually reviewed the mapped terms for accuracy and indicated in column G whether we agreed with the mapping (yes), disagreed (no) or could not make a judgement call (undecided). For the low probability mappings (under 0.25 probability, approximately 100 mappings), and mappings that were questionable, a ticket was created in our GitHub issue tracker (https://github.com/monarch-initiative/monarch-disease-ontology/issues) for further discussion and review.

Summary of Findings

In total, 3,224 mappings were reviewed from the neoplasm branch. Of these, the majority of terms (83%) were marked as correct (yes), less than 14% were marked as incorrect (no), and approximately 2% were undecided (Table 1).

Table 1: Summary of assignments for mapped terms from NCIt to other vocabularies. If a mapping of a NCIt term to another class was determined to be correct, it was marked "yes", if it was wrong, it was marked "no", and it was marked "undecided" if a decision could not be reached, and more discussion was required.

	Total # of terms	Percent of total
Yes	**2611**	80.99%
No	**555**	17.21%
Undecided	**58**	1.80%

Table 2: Summary of terms mapped from 4 terminologies to NCIt. Mappings from NCIt to four terminologies were reviewed. The total number of terms and the percent of total are listed below, as well as the percentage of correct ("yes"), incorrect ("no") and undecided mappings.

	Total # of terms	**Percent of Total**	**Percent Yes**	**Percent No**	**Percent Undecided**
DO	**2651**	82.2%	84.2%	13.8%	2.0%
OMIM	**522**	16.2%	79.3%	18.4%	2.3%
Orphanet	**48**	1.5%	97.9%	2.1%	0.0%
MESH	**3**	0.1%	100.0%	0.0%	0.0%

NCIT classifications

NCIT classifies benign and malignant neoplasms as subclasses of neoplasm. There appears to be some inconsistency with the way that NCIT uses the term tumor. In many cases, tumor is a subclass of neoplasm. For example, C4473 *Dermal Duct Tumor* is a subclass of C27273 *Poroma*, which is a subclass of C4879 *Benign Sweat Gland Neoplasm*. However, in some cases, tumor is used as an exact synonym for neoplasm, such as in C40178 *Mixed Endometrial Stromal and Smooth Muscle Neoplasm.* In the review of the mappings, if a DO *tumor* term was mapped to an NCIT *neoplasm* term, that was marked as correct. For example, see NCIT_C40178 *Mixed Endometrial Stromal and Smooth Muscle Neoplasm* is Equivalent To DOID_8302 *mixed endometrial stromal and smooth muscle tumor.* In addition, should *malignant neoplasm* and *carcinoma* be equivalent to *cancer?* These mappings were marked correct in this manual review, but this may be an open question for further discussion.

Disease Ontology classifications

In the review the of the DO mappings to NCIT, approximately 14% of terms were marked as incorrect. One issue that was frequently seen was DO cross references their cancer terms to NCIT neoplasm terms, but we do not consider these terms to be equivalent. For example, NCIT_C3361 *Salivary Gland Neoplasm* was mapped as equivalent to DOID_8850 *salivary gland cancer*, but neoplasm is not considered equivalent to cancer. In addition, NCIT tumor terms mapped DO cancer terms were marked as incorrect. For example, NCIT_C7113 *Endometrioid Tumor* is not equivalent to DOID_3001 *female reproductive endometrioid cancer.* Carcinoma terms mapped to cancer terms were also marked incorrect. For example, NCIT_C6014 *Paranasal Sinus Carcinoma* EquivalentTo DOID_0050619 *paranasal sinus cancer*.

kBOOM found regular examples where DO is cross referencing NCIT parent and child. The DO either included exact synonyms that cross referenced more than one NCIT class, or directly cross referenced more than one NCIT class, causing an incorrect mapping of NCIT to DO. For example, DOID_1660 *malignant pineal area germ cell neoplasm* has the exact synonym *malignant Pineal Parenchymal germ cell tumor,* which cross references NCI2004_11_17:C6767 *Malignant Pineal Region Germ Cell Tumor,* but the DO class also cross references NCIT_C4659 *Pineal Region Germ Cell Tumor,* which is the superclass of the malignant term. In these cases, the mappings were marked as incorrect in the spreadsheet.

NCIT and DO use different definitions for the term 'neoplasm'. For NCIT terms, the text definition states a neoplasm can be benign or malignant. However, DO cross references benign neoplasm terms to NCIT neoplasm terms. We view this as a significant error. For example, DOID_461 *muscle benign neoplasm* has the exact synonym *Myomatous tumor,* which is cross referenced to NCIT_C4063 *Myomatous Neoplasm*, which by definition is "A *benign or malignant* mesenchymal neoplasm arising from smooth, skeletal, or cardiac muscle."

Orphanet classifications

All of the Orphanet classifications were marked correct with the exception of one: Orphanet_99966 *Atypical teratoid rhabdoid tumor* is suggested to be a subclass of NCIT_C6906 *Atypical Teratoid/Rhabdoid Tumor*, however, I think the Orphanet cross reference to OMIM 609322 *RHABDOID TUMOR PREDISPOSITION SYNDROME 1; RTPS1* is incorrect, and the Orphanet and NCIT terms are equivalent.

OMIM classifications

As indicated in Table 2, approximately 20% of the terms in OMIM were marked incorrect. In some cases, OMIM combines two terms into one record, such as 'anal canal carcinoma' and 'cloacogenic carcinoma' (see record: https://www.omim.org/entry/105580). Another example is OMIM terms may be incorrectly classified as subclasses of NCIT terms, when in fact, they are likely equivalent. This may be due to incorrect cross references to OMIM in Orphanet. For example, NCIT_C3698 *Choroid Plexus Papilloma* is classified as a subclass of OMIM_260500 *Papilloma of Choroid Plexus*, but these terms are probably equivalent. A small percentage of OMIM classifications were 'undecided' (2.3%). In some of these cases, a 'susceptibility to' term was classified as a subclass of the cancer, such as OMIM_215400 *Chordoma, Susceptibility to* is a subclass of NCIT_C2947 *Chordoma*. These were classified as incorrect, as NCIT 'susceptibility' classes, are classified as children of NCIT_C3266 *Hereditary Neoplastic Syndrome* and not the cancer class to which they are susceptible.

T2.2: ICD-O Mapping

Methods

A spreadsheet containing post-composed IDC-O terms from version 3.1 was provided by Larry Wright. The spreadsheet is available in the Supplementary File folder, saved as T2.2ICD-O_Mapping. Rows 3 - 2965 contains mappings of the ICD-O terms to NCIt terms that were previously assigned by the NCIt team. Rows 2966 – 3853 did not contain mappings to NCIt terms. Our team manually reviewed these remaining terms and mapped the ICD-O terms to NCIt terms, where possible. If a NCIt term did not exist, a suggested parent was recommended in some cases.

Summary of Findings

A total of 86 NCIt terms were mapped to ICD-O terms (see column I (Label) and column K (NCIt ID)). New synonyms were recommended for 48 terms (see column J). A total of 342 new parents were suggested for the ICD-O terms that did not exist in NCIt (see column N). A parent class was suggested in case the NCIt team would like to add the ICD-O term as a new class in NCIt. Notes and comments were collected in column O (NV Notes).

Recommendations

For the purposes of the Monarch Disease Ontology, the incorrect mappings can be corrected in the curation file, where we can manually override the kBOOM classifications. We can share our findings with the other disease ontologies and encourage their review of their mappings to NCIT.

Conclusions and Future Directions

This report highlights some of the problematic cross-references to NCIT. We intend to write up these findings and submit a manuscript for publication. This work will help facilitate best practices for linking of NCIT for accurate data annotation, validation, mapping, and integration. The new ICD-O mappings can be used by communities wishing to migrate ICD-O to NCIt.