Towards an Ontology for Representing Malignant Neoplasms

William D. Duncan[1,*] Carmelo Gaudioso[2] and Alexander D. Diehl[3]
[1] Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute, Buffalo, NY, 14203, USA
[2] Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute, Buffalo, NY, 14203, USA
[3] Department of Biomedical Informatics, University at Buffalo, Buffalo, NY, 14203, USA

ABSTRACT

Oncology research produces data about a wide variety of entities such as tumor types, locations, pathology, and staging, patient treatments and outcomes, and experimental systems such as mouse models and cell lines. In order to conduct effective cancer research, terminologies, classification systems, and ontologies are needed that can integrate these various datasets and provide standards for consistently representing entities.

In this paper, we discuss our ongoing efforts to address these difficulties by developing a realism-based ontology for representing instances of malignant neoplasms, disease progression, treatments, and outcomes. This ontology is being built using the principles of the OBO Foundry, and makes use of other OBO Foundry ontologies, such as the Ontology for General Medical Sciences, Uberon, and the Cell Ontology. As a result of our efforts, we have made worthwhile progress towards developing a robust ontological framework for representing malignant neoplasms.

1 INTRODUCTION

Oncology research produces data about a wide variety of entities such as tumor types, locations, pathology, and staging, patient treatments and outcomes, and experimental systems such as mouse models and cell lines. In order to conduct effective cancer research, terminologies, classification systems, and ontologies are needed that can integrate these various datasets and provide standards for consistently representing entities. These standards facilitate the meaningful linking, sharing, and analysis of disparate datasets between researchers and across institutions. However, the incomplete and inconsistent representation of cancer-related data makes it difficult to perform these activities.

In this paper, we discuss our ongoing efforts to address these difficulties by developing a realism-based ontology for representing instances of malignant neoplasms, disease progression, treatments, and outcomes. This ontology is being built using the principles of the OBO Foundry (Smith et al. 2007), and makes use of other OBO Foundry ontologies, such as the Ontology for General Medical Sciences and the Cell Ontology. We chose to focus on these entities because they are key elements driving accurate cohort selection based on diagnosis, stage, and treatment; and clinical decision support. As a result of our efforts, we have made worthwhile progress towards developing a robust ontological framework for representing malignant neoplasms.

* To whom correspondence should be addressed:
william.duncan@roswellpark.org

2 PROJECT MOTIVATION

This research developed out of a number of interests. The first is that we recognized a need to connect cancer data from multiple sources with differing levels of granularity. Some important levels include: (1) diagnosis and treatment information about the patient and how the patient responses to treatment; (2) anatomical information about the organs in which the cancer originates; (3) pathology information about the tissues removed during procedures, such as tumor tissues and lymph nodes; (4) cellular information, such as data obtained from flow cytometry and immunohistochemistry; (5) and molecular information, such as genomic sequencing. Providing a framework for tying these kinds of data together is essential for cancer research by providing the basis for the use of advanced ontology-based querying and analytical methods that allow for data integration across multiple sources and scales.

3 CURRENT CLASSIFICATION SYSTEMS, TERMINOLOGIES, AND ONTOLOGIES

A number of existing classification systems, ontologies, and terminologies have terms for representing malignant neoplasms. Prominent examples include the International Statistical Classification of Diseases 10th Revision (ICD-10), the International Statistical Classification of Diseases for Oncology (ICD-O), the National Cancer Institute Thesaurus (NCIT), and the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT). However, since many of these information organization systems do not share a common upper-level framework, it is not easy to leverage information contained in other terminologies and ontologies. For instance, SNOMED CT does not have terms for checkpoint inhibitors, whereas the NCI Thesaurus does. Ideally, we would like to use terms from each system (i.e., SNOMED CT and NCIT), but due to differences in their relations and hierarchical structures, it is difficult to do so. For example, the term 'metastasis' denotes a disorder in SNOMED CT, but denotes the spread of cancer (i.e., a process) in the NCIT. OBO Foundry ontologies, in contrast, are generally designed using the Basic Formal Ontology (BFO) as their upper-level framework, and this enables the creation of domain specific ontologies whose terms can be reused by other OBO Foundry ontologies. For instance, the Drug On-

tology (Hanna et al. 2013) uses terms from the Chemical Entities of Biological Interest (ChEBI) (Hastings et al. 2013) ontology to represent drug ingredients.

3.1 International Statistical Classification of Diseases

Due to its long history and widespread adoption, the International Statistical Classification of Diseases (ICD)[1] is perhaps the most relevant system for classifying diseases. Maintained by the World Health Organization (WHO), ICD is a globally recognized healthcare classification system consisting of hierarchically structured codes that represent diseases, disorders, and other health related issues.[2] In relation to the current topic, the International Statistical Classification of Diseases for Oncology (ICD-O)[3] has codes for representing a number of pertinent characteristics of a malignant neoplasm, such as the anatomical site of the neoplasm, the neoplasm's histology (e.g., small cell, clear cell), and behavior (e.g., if it has metastasized). For example, an ovarian adenocarcinoma is represented using the following combination of codes:

- C56 – the site code for an ovary
- 8140/3 – 8140 is the code for a neoplasm arising from glandular epithelial tissue, and '/3' represents that the neoplasm is malignant

The advantage of ICD's coding system is that allows diseases to be easily grouped and counted for statistical and reporting purposes. For instance, to find all patients who have an adenocarcinoma, you only have to look for patients whose histological code begins with '814' and has a behavior code greater than 3. However, there are two related noteworthy drawbacks to implementing ICD as an ontology. First, ICD does not contain codes for many of the important cancer related entities that need to be represented, such as treatments and molecular disorders. This shortcoming is compounded by ICD's lack of formal relations that would allow codes to be linked to other information. Thus, even if we created code lists for the missing entities, we would still be faced with the task of creating well-defined relations that would allow this information to be linked to ICD codes.

3.2 National Cancer Institute Thesaurus

The National Cancer Institute Thesaurus (NCIT) is a reference terminology developed by the National Cancer Institute (Sioutos et al. 2007). It contains over 100,000 concepts with textual definitions and 400,000 cross links between its concepts.[4]

In our examination of the NCIT, we found that many of the definitions in the malignant neoplasms branch were sufficiently defined and the hierarchy was rich enough to suit our purposes. However, when we examined other branches of the NCIT, certain problems became apparent. In particular, we found the definitions for cell types related to cancer to be inadequate. Consider the following NCIT concepts and definitions:

- *Abnormal Cell (C12913)*: An abnormal human cell type which can occur in either disease states or disease models.
- *Neoplastic Cell (C12922)*: Cells of, or derived from, a tumor.
- *Malignant Cell (C12917)*: Cells of, or derived from, a malignant tumor.

The definition for *Abnormal Cell* suffers from its being circular (i.e., an abnormal cell is defined as being an abnormal cell type), and thus the definition does not provide any new information. Furthermore, the definition specifically states that an abnormal cell is a human cell. This prevents the NCIT from consistently modeling data about abnormal cells from non-human species despite the fact that the NCIT does contain concepts for mouse diseases, such as *Mouse Carcinoma* (C24010). Given the importance of mouse models in cancer research, not being able to represent data from mouse studies correctly is a severe limitation.

The definitions for *Neoplastic Cell* and *Malignant Cell* do not provide much clarity about how these cells relate to neoplasms. Since a neoplasm may also contain normal cell types, more details are needed about what it means to be a neoplastic cell other than being derived from a tumor. Furthermore, while a metastasis may be said in some sense to derive from a tumor, this cannot be said of the originating neoplastic cells that first started proliferating during the tumor formation process. Lastly, it needs to be pointed out that these cell types form a hierarchy. A *Malignant Cell* is a type of *Neoplastic Cell*, and *Neoplastic Cell* is a type of *Abnormal Cell*. This information is not contained in the textual definitions in an Aristotelian fashion, although it is represented in NCIT's taxonomic relations.

3.3 Systematized Nomenclature of Medicine Clinical Terms

Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a comprehensive health terminology that provides a standardized way to represent clinical information in an electronic health record.[5] Although SNOMED CT has a large number of terms for clinical findings and disorders, it does not have worked out terms for other terms

[1] For brevity, we use the general term 'ICD' to refer to the number of different versions of ICD, such as ICD-10 and ICD-O.

[2] http://www.who.int/classifications/icd/en, accessed 2017-06-20.

[3] https://training.seer.cancer.gov/icdo3, accessed 2017-06-20.

[4] https://ncit.nci.nih.gov/ncitbrowser, accessed 2017-06-21.

[5] http://www.snomed.org/snomed-ct/what-is-snomed-ct, accessed 2017-06-21.

related to neoplasms. For example, the concept *Tumor cell* (SCTID 252987004) is defined as subtype of the concept *Abnormal cell* (SCTID 39266006), but this does not specify if the concept *Tumor cell* represents malignant cells.[6] Furthermore, the concept *Malignant tumor cells* (SCTID: 88400008) is defined as being a subtype of the concept *Malignant neoplasm, primary* (SCTID: 86049000).[7] This classification is incorrect for at least two reasons. First, although malignant tumor cells are often part of a malignant neoplasm, they are not a kind of malignant neoplasm. A malignant neoplasm (as stated above) will also include a number of non-cancer cells as part of its makeup. Second, even if we accept that a malignant tumor cell is a kind of malignant neoplasm, this definition is incorrect because a malignant tumor cell is also found in a metastasis (metastatic offshoot of a primary tumor). Finally, SNOMED CT classifies *Malignant tumor cells* as a kind of *Morphologic abnormality* (SCTID 4147007), and not a *Disorder* (SCTID 64572001). In SNOMED CT, the distinction between a morphologic abnormality and a disorder is that some underlying pathological process supports a disorder.[8] However, the reason that cell becomes malignant is because of underlying pathological processes (resulting from dysregulation) occurring with it.

3.4 Disease Ontology

The Disease Ontology (DO) is an OBO Foundry ontology built for the purposes of providing the biomedical community with consistent, standardized, and reusable definitions to represent the range of human diseases (Schriml et al. 2015). Although we found the DO to have decent coverage for cancer types, there are two difficulties with it that made the DO not suitable for our purposes.

First, the DO is not consistent in its use of the terms 'cancer' and 'neoplasm'. In DO, cancer is defined as a kind of disposition. This means that cancer is not a material thing (i.e., does not have mass), but rather is a kind of latent potential that is actualized when cells start proliferating out of control. Malignant neoplasms, are material objects that come into being due to uncontrolled cell proliferation. In the DO, however, there are number of terms in the cancer branch that reference neoplasms as material things and not the disposition of cancer. For example, *ovary neuroendocrine neoplasm* is defined as a subtype of *ovarian cancer*. Because of DO's inconsistent use of terms 'cancer' and 'neoplasm' and our remaining true to the OBO Foundry principles, we decided it would be beneficial to the development of our ontology to use the term 'malignant neoplasm' and avoid using the term 'cancer' when possible.

Second, the DO is missing needed formal axioms that relate entities having the disposition of cancer to the anatomical structures in which these entities are located. For instance, the DO term *ovary epithelial cancer* does not have axioms that formally relate the disposition to the epithelial cells that are part of the ovary. The lack of these axioms can make it difficult to query data modeled using the DO. For example, it is not possible to query for the most common anatomical structures in which malignant neoplasms are found.

3.5 On carcinomas and other pathological entities

In Smith et al. (2005b), the Ontology for Biomedical Reality (Rosse et al. 2005) is modified to account for material anatomical entities, material pathological entities, and pathological formations. Material anatomical entities are anatomical structures (e.g., organs, cells) or bodily substances (e.g., blood) that are found in a healthy organism. Anatomical structures are defined as being material anatomical entities that have an inherent 3D structure generated by the coordinated expression organism's own structural genes (Smith et al. 2005b). They include both canonical and variant anatomical structures. Canonical anatomical structures belong to 'idealized' healthy human beings. Variant anatomical structures are entities that deviate from the norm (e.g., having extra fingers), but are not pathological in the sense discussed below.

An anatomical entity is defined as being a material pathological entity when (Smith et al. 2005b):

- It has come into being as a result of changes in some pre-existing canonical anatomical structure through processes other than the expression of the normal complement of genes of an organism of the given type.
- It is predisposed to have health-related consequences for the organism in question manifested by symptoms and signs.

Material pathological entities include pathological structures and pathological bodily substances. These are anatomical structures and body substances, respectively, that host some kind of pathological formation, a formation being pathological when it affects an organism's physiological processes to the degree that they give rise to signs and symptoms. For instance, a carcinoma is a pathological formation that arises within an anatomical structure, such as an ovary.

A high-level summary of the hierarchy for material anatomical, material pathological entities is depicted below:

- material anatomical entity
    - anatomical structure
        - canonical anatomical structure
        - variant anatomical structure

[6] http://browser.ihtsdotools.org, accessed 2017-06-22.

[7] Ibid.

[8] https://confluence.ihtsdotools.org/display/DOCEG/6.1.1+Clinical++-+definition, accessed 2017-06-22.

  - portion of canonical body substance (e.g., portion of blood)
- material pathological entity
  - pathological structure (e.g., neoplasm)
  - Portion of pathological substance (e.g., portion of pus)

Pathological formations are then related to their hosts and the entities out of they originate using the following relations from the Open Biomedical Ontology (Smith et al. 2005a):[9]

- **instance of**: A primitive relation that holds between a particular individual and the universal (type or kind) that the particular individual instantiates at particular time. For example, particular patient is an **instance of** a human being at a particular time.
- **part of**: A primitive relation between instances of parts and wholes at a particular time. For example, a particular mass of malignant epithelia tissue is **part of** a particular ovary at a particular time.
- **is a**: *A* **is a** *B* means that *A* and *B* are universals and for all times *t* every particular individual *i*, if *i* **instance of** *A* at *t*, then *i* **instance of** *B* at *t*. For example, a human being is a mammal.
- **derived from**:[10] A primitive relation between two distinct instances *i, j* and times *t, t'* and is such that changes in *i* at *t* results in a new second entity *j* at *t'*. For example, a particular blastocyst **derived from** a particular zygote.
- **transformation of**: A **transformation of** B means that are universals and for all times *t* if *i* **instance of** *A* at *t*, then there is an earlier time *t'* at which *i* was an **instance of** *B*.

As an example, suppose a patient (*patient1*) has a carcinoma (*carinoma1*) that originated within her ovary (*ovary1*). We represent this using the axioms:

ovary1 at *t* **part of** *patient1* at *t*
carcinoma1 at *t* **part of** *ovary1* at *t*
carcinoma1 at *t* **instance of** *pathological structure* at *t*

Because carcinomas arise from the epithelial tissue lining of organs, we can assert the following about the patient's tumor:

carcinoma1 at *t* **derived from** some *epithelial cell* at some *t'* prior to *t*

And, since the **part of** relation is transitive, we infer that:

carcincoma1 at *t* **part of** *patient1* at *t*

Although this inference is trivial, the advantage of representing the patient's tumor in this manner is that we are not required to explicitly state this within an information system using the ontology. Rather, we let the computer system handle this through automated inferencing.

The benefit of doing this becomes apparent when we consider the multiple ways we classify malignant neoplasms. A malignant neoplasm may be classified according to:

- The cell type from which the neoplasm is originates, e.g., carcinomas arise from epithelial cells, and sarcomas arise from non-epithelial cells.
- The organ in which the neoplasm develops, e.g., an ovarian carcinoma originates in the ovary.
- The organ system to which the organ of origin belongs, e.g., an ovarian carcinoma is a kind of reproductive system cancer
- The anatomical site or region in which the organ of origin is found, e.g., a tongue carcinoma is a kind of head and neck cancer.

When such classification information is axiomatized, we can then query the information system along these multiple axes without have to maintain complex data structures that explicitly assert this information. For instance, we can now query an information system for all carcinomas (i.e., malignant neoplasms that are derived from epithelial cells) that belong to patients' reproductive systems without having to explicitly link each kind of carcinoma (e.g., ovarian, uterine, testicular) to the organ and associated organ system.

4 OUR PROPOSED ONTOLOGY

While we consider the work of Smith et al. to be a significant improvement over the aforementioned classification systems and terminologies, a number of ontologies have been developed after this work was published. We take advantage of these more recent ontologies as follows. First, we make use of the terms and relations from the Cell Ontology (CL) (Diehl et al. 2016) to represent the types of cells from which a malignant neoplasm arises. Moreover, as a result of our work, the CL added the terms *abnormal cell*, *neoplastic cell*, and *malignant cell* in order to better represent cell types that play integral roles in tumor formation:

- *abnormal cell*: A cell found in an organism or derived from an organism exhibiting a phenotype that

[9] Hereafter, relations are represented in **bold.**

[10] In the referenced Open Biomedical Ontology relations, the name the relation is named 'derives from'. However, to avoid confusion, we use the term as presented in the paper.

deviates from the expected phenotype of any native cell type of that organism. Abnormal cells are typically found in disease states or disease models.

- *neoplastic cell*: An abnormal cell exhibiting dysregulation of cell proliferation or programmed cell death and capable of forming a neoplasm, an aggregate of cells in the form of a tumor mass or an excess number of abnormal cells (liquid tumor) within an organism.
- *malignant cell*: A neoplastic cell that is capable of entering a surrounding tissue.

Second, an important criterion in Smith et al.'s definition of an entity being pathological is that it is predisposed to have health related consequences (Smith et al. 2005b). To more precisely account for predispositions of this sort, we adopt Ontology for General Medical Sciences' (OGMS) model of disease (Scheuermann et al. 2009). In OGMS, a disease is type of disposition that is manifested (or realized) during those processes that compromise an organism's physiological health. This permits us to represent that an organism may have a disease even though the disease is not currently being realized. A malignant neoplasm, for instance, may shed malignant cells that remain dormant in the patient until at some later time they begin to proliferate. During this dormant period, these malignant cells possess the disposition for undergoing uncontrolled cell proliferation, although the disposition is not being realized. Similarly, the genome within a native cell may have mutations in its BRCA1 or BRCA2 genes, but the cell may behave normally until certain cellular processes uncover the pathological effects of these mutations. Using the dispositional account of disease, we then incorporate the Disease Ontology's (DO) representation of cancer as follows:

- *disease*: A disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism.[11]
- *disease of cellular proliferation*: A disease that is characterized by abnormally rapid cell division.
- *cancer*: A *disease of cellular proliferation* that is malignant and primary, characterized by uncontrolled cellular proliferation, local cell invasion and metastasis.

Recall that above we criticized the DO for its inconsistent usage of the term 'neoplasm'. However, given the need to represent the dispositional aspect of cancer, we find DO's hierarchy appropriate for characterizing cancer as we are clear and consistent about which sense of 'cancer' we are using.

[11] DO uses the OGMS term *disease.*

Third, in order to account for Smith et al.'s distinction between material pathological entities and material anatomical entities, we adopt OGMS' account of a *disease* (as a disposition) being based on a *disorder*:

> *disorder*: A material entity which is clinically abnormal and part of an extended organism. Disorders are the physical basis of disease.

Since OGMS uses the Basic Formal Ontology (BFO) as its upper-level framework and a *disease* in OGMS is type of BFO *disposition*, it cannot (like all dispositions) exist on its own. Rather, a *disease* must be borne by a *disorder* whose structural abnormalities serve as a *disease's* basis. For example, a sprained ankle is a *disorder* in the sense that the physical structures are clinically abnormal, and these physiological abnormalities are the reason that a sprained ankle is disposed to swell.

Fourth, to relate a *disease* to the *disorder* upon which it is based, we define the **has material basis in** relation as follows:

> **has material basis in**: A primitive relation between an instance of a disease *i* and an instance of a disorder *j* at particular time *t* in which *i* exists because of the physical makeup of some part of *j* at time *t*.

In addition to relating a *disease* to its basis, we must also account for the processes that realize (or make manifest) an instance of a *disease*. For this we use OGMS' term *pathological bodily process*:

> *pathological bodily process*: A bodily process that is clinically abnormal.

As observed in the definition, a *pathological bodily process* is a type of bodily process. However, the term *bodily process* is not defined in OGMS.

Fifth, in order to account for the temporal development of malignant neoplasms, we make use of the Relations Ontology's **derives from** and **develops from** relations (Smith et al. 2005a). The **derives from** relation is similar to the aforementioned **derived from** relation, but adds the criteria that the originating entity ceases to exist when the new entity is created and the newly created entity inherits a significant portion of its matter from the originating entity. For example, the assertion:

> *abnormal cell* **derives from** *native cell*

entails that a particular *native cell* no longer exists once the *abnormal cell* derived from it comes into existence.

The **develops from** relation also represents new entities that arise from previously existing entities, but does not require that the originating entity cease to exist. This allows us to represent that an instance of a secondary neoplasm

develops from an instance of a primary neoplasm without having to commit the primary neoplasm's ceasing to exist.

Sixth, given the importance of representing the anatomical structures in which malignant neoplasm from, we incorporate the Uberon's *anatomical structure* and OGMS' *pathological anatomical structure* terms, and define them as follows (Mungall et al. 2012):

- *material anatomical entity*: An anatomical entity that has mass.
- *anatomical structure*: A material anatomical entity that is a single connected structure with inherent 3D shape generated by coordinated expression of the organism's own genome.
- *pathological anatomical structure*: A material entity that comes into being as a result of changes in some pre-existing anatomical structure through processes other than the expression of the normal complement of genes of an organism of the given type, and is predisposed to have health-related consequences for the organism in question manifested by symptoms and signs.

We note here that although intuitively a *pathological anatomical structure* is a type of *anatomical structure*, for reasons that will be discussed below, we classify them in separate hierarchies. Moreover, we assert that a particular *pathological anatomical structure* (1) **develops from** an instance of a previously existing *anatomical structure*, and (2) **has part** an instance of a *disorder*. These two assertions define both necessary and sufficient conditions for an entity to be a *pathological anatomical structure*.

Lastly, with the above modifications in place, we define the following terms to necessary for an ontology of malignant neoplasms:

- *dysregulation of cell proliferation*: A *pathological bodily process* during which cell proliferation occurs at a level not normal for that cell type in its native context.
- *neoplasm*: A *disorder* that results from dysregulation of cell proliferation (uncontrolled cell proliferation).
- *malignant neoplasm*: A *neoplasm* that has acquired the disposition to invade surrounding tissues and spread to remote anatomical sites.
- *primary neoplasm*: A *malignant neoplasm* that is found in the site where the malignant cells first began proliferating.
- *secondary neoplasm*: A *malignant neoplasm* that develops from a *primary neoplasm*.

A summary of proposed ontology of malignant neoplasms is depicted in **Figure 1**.

5 DISCUSSION

We began our work in order to build an application ontology to assist us in analyzing data in an ovarian cancer patient registry (work in progress). Because of our commitment to OBO Foundry principles and ontological realism, we began our ontology development by considering existing ontologies, including OGMS and DO, and related ontologies such as Uberon and CL. Our aim has been to reuse ontology classes where possible and create new classes and hierarchies where existing ontologies either are missing classes or providing faulty modeling of the domain.

We have found the NCIT to be a very useful source of information about cancer related entities, their definitions, and their relationships to each other. Although the NCIT is very large and has been developed over many years, it really remains a terminology rather than an ontology. For example, the NCIT includes the term *Disease or Disorder* defined as*:*

> Any abnormal condition of the body or mind that causes discomfort, dysfunction, or distress to the person affected or those in contact with the person. The term is often used broadly to include injuries, disabilities, syndromes, symptoms, deviant behaviors, and atypical variations of structure and function.

This definition does not adequately distinguish between the processes and material entities that result in abnormal conditions. This distinction is important for precisely representing the nature of a malady. If a cancer patient has difficulty breathing due to metastatic tumors spreading throughout the lungs, both the difficulty in breathing and the tumors are abnormal conditions, and hence, are would be classified using the term *Disease or Disorder* (C2991). But, in reality, the process of breathing is a distinct kind of entity than a tumor, which is a material entity. There are past and current efforts to redevelop NCIT or at least sections of it into a proper ontology. Our hope is these efforts will make the NCIT more aligned with OBO Foundry principles. One important result of our work was the addition the abnormal cell, neoplastic cell, and malignant cell types to CL. These CL classes parallel the naming and relationships of the NCIT concepts, but as discussed above, we chose to write new definitions that better define these cell types and do not limit their applicability unnecessarily.

In considering the Disease Ontology, we found it to be a useful catalog of cancer types, but as discussed above, we find that there is confusion as to whether neoplasms are dispositions or disorders. Because of our need to represent pathological findings, we need to reflect that these findings

are about disorders (which are material entities) that are observed by pathologists, and not about dispositions, which are not directly observable.

An important finding of our work is that we found that OBO Foundry ontologies have difficulty representing abnormal or pathological entities. Two prominent examples are pathological anatomical structures and pathological processes. Intuitively, a pathological anatomical structure is a kind of anatomical structure. For instance, an ovary containing a carcinoma is still an instance of an ovary. However, the standard definition (with some variations) for *anatomical structure* found in Uberon, the Common Anatomy Reference Ontology, the Foundational Model of Anatomy, and the Anatomical Entity Ontology does not allow for this:[12]

> Material anatomical entity that has inherent 3D shape and is generated by coordinated expression of the organism's own genome.

This issue is that disorders (such as neoplasms and fractures) that arise with anatomical structures are not necessarily generated by the organism's genome. Thus, the definition is too strong. Smith et al. are aware of this propose an anatomical hierarchy consisting of top-level *anatomical structure* term with subtypes of *canonical anatomical structure*, *variant anatomical structure*, and *pathological anatomical structure* (Smith et al. 2005a):

- Anatomical structure
    - Canonical anatomical structure
    - Variant anatomical structure
    - Pathological anatomical structure

While we think this a reasonable proposal, the lack of a definition for *anatomical structure* makes is unclear as to what canonical, variant, and pathological structures have in common.

A similar problem exists for abnormal processes (such as *dysregulation of cell proliferation*). The OGMS, for its part, does provide the term *pathological bodily process*. But, this term is orphaned from other biological processes found in other OBO ontologies. For example, the Gene Ontology (GO) includes the term *biological process*:[13]

> Any process specifically pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms. A process is a collection of molecular events with a defined beginning and end.

Again, intuitively it makes senses that a *pathological bodily process* should be a subtype of *biological process*. However, the definition of *biological process* does not permit this.

Although we do not have any concrete solutions, at this point, for how to align pathological structures and processes with these other OBO terms, we look forward to collaborating with the OBO Foundry community on creating a coherent structure for these upper level classes that is shared among all OBO Foundry ontologies. Thus, we simply leave *pathological anatomical structure* as a subtype of *material anatomical entity*, and *pathological bodily process* in its current OGMS hierarchy.

Our goal is to contribute to the oncology domain by creating a strong and consistent ontological foundation for providing metadata and data analysis of patient cancer data for both research and clinical applications including clinical decision support. The ontological framework described herein attempts to solve some continuing issues in the representation of cancer as a disease and the disorders (neoplasms) in which it presents. Our framework is intended to be useful for the description and classification of data used in cancer diagnosis and treatment. In future work, we will be adding classes to represent additional entities associated with cancer such as laboratory methods and results, treatments, and outcomes. We hope our ontology will support other oncology researchers in exploiting the full potential of patient data registries and other cancer-related datasets.

ACKNOWLEDGEMENTS

We gratefully acknowledge support as follows. William Duncan and Carmelo Gaudioso received support from the Clinical Data Network, a Roswell Park Cancer Institute Cancer Center Support Grant shared resource funded by NCI P30CA16056. Alexander Diehl received support from NCATS 5UL1TR001412. All three authors received support from NCI P50CA159981.

REFERENCES

Arp, R., Smith, B., & Spear, A.D. (2015). *Building Ontologies With Basic Formal Ontology*. The MIT Press. doi:10.7551/mitpress/9780262527-811.001.0001.

Diehl, A. D., Meehan, T. F., Bradford, Y. M., Brush, M. H., Dahdul, W. M., Dougall, D. S., … Mungall, C. J. (2016). The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, *7*(44). doi: 10.1186/s13326-016-0088-7. PMCID: PMC4932724. https://github.com/obophenotype/cell-ontology.

Hanna, J., Joseph, E., Brochhausen, M., & Hogan, W. R. (2013). Building a drug ontology based on RxNorm and other sources. *Journal of Biomedical Semantics*, 4(44). doi:10.1186/2041-1480-4-44. PMCID: PMC3931349.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., & Steinbeck, C. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research*, 41(Database issue), D456–D463. doi:10.1093/nar/gks1146.

Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., & Haendel, M. A. (2012). Uberon, an integrative multi-species anatomy ontolo-

[12] Definitions retrieved from www.ontobee.org, accessed 2017-07-06.

[13] Definition retrieved from www.ontobee.org, accessed 2017-07-06.

gy. *Genome Biology*, *13*(1), R5. doi: 10.1186/gb-2012-13-1-r5. PMCID: PMC3334586. http://uberon.github.io.

Rosse, C., Kumar, A., Mejino, J. L., Cook, D. L., Detwiler, L. T., & Smith, B. (2005). A Strategy for Improving and Integrating Biomedical Ontologies. *AMIA Annual Symposium Proceedings*, 2005, 639–643. PMCID: PMC1560467

Scheuermann, R.H., Ceusters, W., & Smith ,B. Toward an ontological treatment of disease and diagnosis. San Francisco: *Proceedings of the 2009 AMIA Summit on Translational Bioinformatics*, 2009, 116–120. PMCID: PMC3041577. https://github.com/OGMS/ogms.

Schriml, L. M., & Mitraka, E. (2015). The Disease Ontology: fostering interoperability between biological and clinical human disease-related data. *Mammalian Genome*, 26(9-10): 584–589. doi: 10.1007/s00335-015-9576-9. PMCID: PMC4602048. http://disease-ontology.org.

Sioutos, N., de Coronado S., Haber M.W., Hartel F.W., Shaiu WL, & Wright LW. (2007). NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1): 30-43. doi: 10.1016/j.jbi.2006.02.013. PMID: 16697710.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., & Rosse, C. (2005a). Relations in biomedical ontologies. *Genome Biology*, 6(5): R46. doi: 10.1186/gb-2005-6-5-r46. PMCID: PMC1175958.

Smith B., Kumar A., Ceusters W., and Rosse C. (2005b). On Carcinomas and Other Pathological Entities. *Comparative and Functional Genomics*, 6(7-8): 379-387. doi:10.1002/cfg.497. PMCID: PMC2447494.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., The OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., & Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnology*, 25(11): 1251–1255. PMCID: PMC2814.

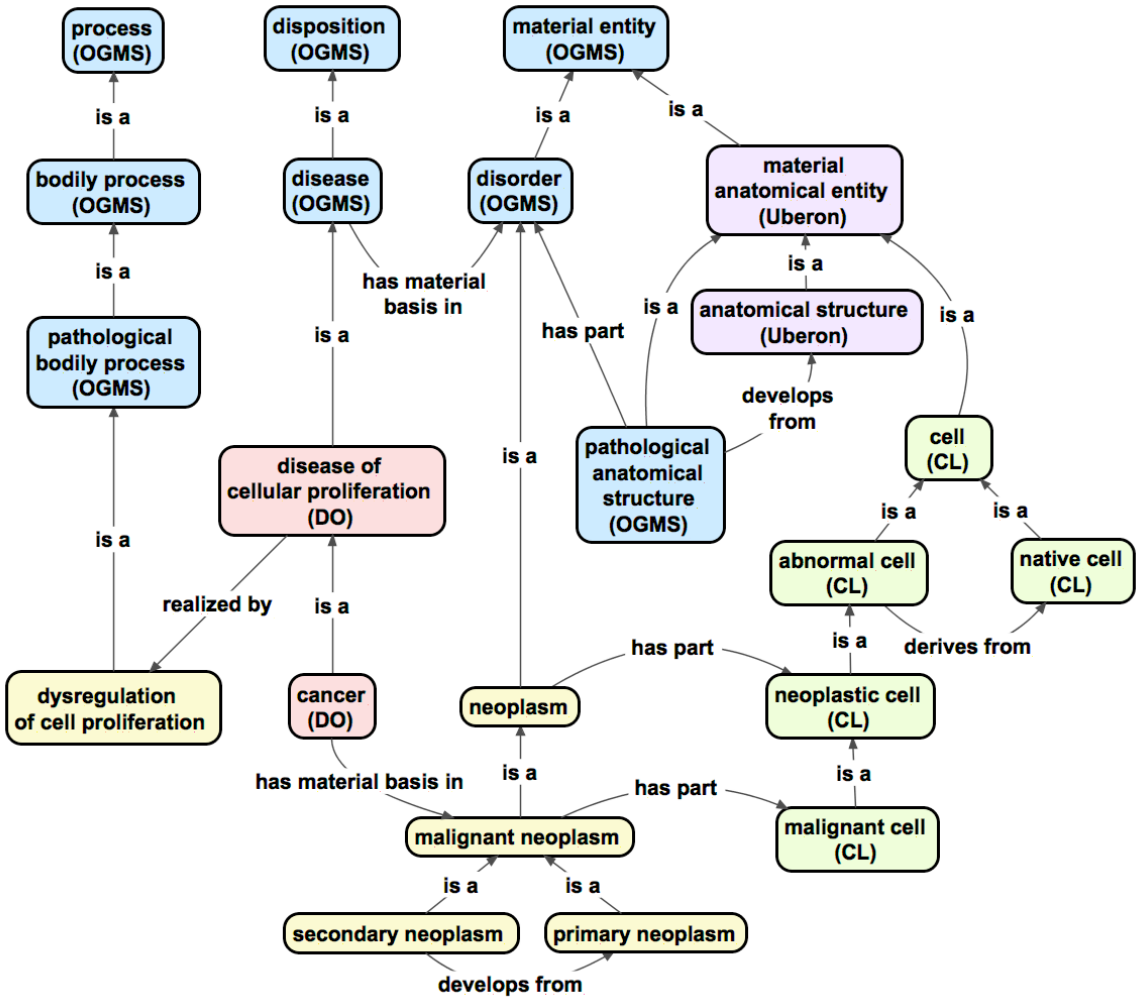

Figure 1. This figure illustrates our proposed ontology for representing malignant neoplasms. Our upper-level framework consists of classes imported from the Ontology for General Medical Sciences (OGMS), Cell Ontology (CL), Disease Ontology (DO), and Uberon. We extend the upper-level framework by adding the classes *dysregulation of cell proliferation*, *neoplasm*, *malignant neoplasm*, *primary neoplasm*, and *secondary neoplasm*.