An Ontology for TNM Clinical Stage Inference

Felipe Massicano[2], **Ariane Sasso**[1], **Henrique Amaral-Silva**[1], **Michel Oleynik**[3], **Calebe Nobrega**[1], **Diogo F. C. Patrão**[1]

[1]CIPE - A. C. Camargo Cancer Center

[2]IPEN - USP

{djogo,ariane.sasso,henrique.silva,cnobrega,michel}@cipe.accamargo.org.br

massicano@gmail.com

Abstract. *TNM is a classification system for assessment of progression stage of malignant tumors. The physician, upon patient examination, classifies a tumor using three variables: T, N and M. Definitions of values for T, N and M depend on the tumor topography (or body part), specified as ICD-O codes. These values are then used to infer the Clinical Stage (CS) and reflect the disease progression, which can be 0 (no malignant tumor), IS (in situ), I, II, III, or IV. The rules for inference are different for each topography and may depend on other factors such as age. With the objective of evaluating missing CS information on A. C. Camargo Cancer Center databases, we developed an open ontology to represent TNM concepts and rules for CS inference. It was designed to be easily expansible and fast to compute.*

1. Introduction

Originally developed in 1958 and since then maintained by the Union for International Cancer Control (UICC), the TNM staging system is a cancer classification scheme used mainly to predict survival rates given the disease severity. Based on the fact that patients with localized tumors present higher survival rates when compared to patients with distant metastasis, the TNM staging system aims to help doctors with treatment planning, disease prognosis, interpretation of treatment results and also to facilitate information sharing and improve cancer research [Sobin and Wittekind C 2002].

The classification is based on three main discrete variables: T (0-4), for the evaluation of the primary tumor extension; N (0-3), for the appraisal of the presence and the extension of metastasis in regional lymph nodes; and M (0-1), to annotate the absence or presence of distant metastasis. Some topographies include an additional character in the range $a - d$ for specifying subcategories. Additional characters can also be included to define the information source (clinical exam or pathology biopsy); the diagnosis stage (before/after treatment, after recurrence or through autopsy); and the existence of multiples tumors in the same site. Moreover, other symbols describe optional lymphatic and venous invasion, the histological grade, the metastasis site, presence of isolated tumor cells, sentinel lymph node invasion status, the degree of certainty and the presence of residual tumor after the treatment [Sobin and Wittekind C 2002].

Additionally, each topography has rules for mapping the TNM staging into one variable called clinical stage. The clinical stage ranges from 0 to IV, with an additional character for some sites. Although rules differ for each topography, higher clinical stages correlates with worse prognosis. Therefore, its determination is a central point in the cancer diagnostic process.

The rules for clinical staging inference, standardized by the TNM staging system, should be used by the physicians during the medical appointment; however, many factors contribute to this not being largely adopted, such as: resistance by physicians to extra paperwork, physicians uncertainty concerning the current staging system and lack of regulatory processes to enforce compliance with the standard [Schmoll 2003]. Many efforts have been made lately to reach that, including its recommendation by specialized medical societies and its use as a mandatory prerequisite for quality accreditation on oncology care [Neuss et al. 2005].

Moreover, the TNM staging information is also crucial for cancer research. As the different clinical stages indicates better or worse response to certain treatments and better or worse prognosis, cancer studies usually focus on diseases of a specific tissue, and a specific clinical stage. If the clinical database does not contain this information for a relevant fraction of the patients, the researchers may have to resort to manually assessing the patient records to find out the sample size.

Since the rules for clinical stage coding are explicitly defined in the TNM publication, it is possible to create a computer program to automatically evaluate them. Such a program would validate existing values, or even provide this information when it is missing. However, representing all rules directly on a computer programming language is an extenuating and repetitive task, and may lead to code maintenance issues. In addition, it would be difficult to a oncology expert, untrained in computer programming, to validate the algorithms.

In order to overcome these difficulties, a proposal to model the concepts, descriptions and rules in TNM clinical stages is to use ontologies. In summary, the term ontology means a specification of a conceptualization and it has been applied to create standardized dictionaries in several fields. [Gruber 1993].

Standardized ontologies have been developed in many areas in such a way that domain experts can share and annotate information in their respective fields. In medicine, well-known standardized and structured vocabularies such as Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT) [1], RadLex [Langlotz 2006], Unified Medical Language System (UMLS) [Lindberg et al. 1993], Medical Subject Headings (MeSH) [Nelson et al. 2001] and others have been used for clinical and research purposes. Although new general and specialized ontologies are emerging fast, there is no published ontology yet that approaches the TNM clinical stage coding problem. Yet, some ontologies may represent some of the TNM concepts.

The National Cancer Institute Thesaurus (NCIt) is a reference terminology that covers the clinical care, basic and translational research, public data and also the administrative domain regarding the National Cancer Institute (NCI). It was built upon the NCI Metathesaurus from the UMLS and it is based on description logic with relationships between semantically rich concepts [Smith et al. 2005]. It is coded on OWL Lite, a subset of OWL-DL with enough complexity to represent the ontology data [Bechhofer et al. 2004]. It provides some of the TNM concepts for 6th and 7th edition and each topography has its

[1] http://www.ihtsdo.org/snomed-ct

own T, N, M and CS classes with annotations in English. When a concept has the same definition in the 6th and 7th edition, it is defined as a single class, or else specific classes for each version are defined. There is no definition of axioms for inference of Clinical Stage based on values of T, N and M.

The SNOMED CT is a vocabulary comprising more than 310.000 concepts hierarchically organized. There are concepts to represent all TNM (including individual definitions for T, N, M and CS for each topography), however, there are no compositional rules connecting the T, N, M and the topography to the CS. Moreover, its license is not open and there is no official or non-official translation to Portuguese.

Dameron et al. propose the creation of an ontology for automatic grading of lung tumours using OWL-DL description logic language, inspired by the controlled vocabulary for cancer, the NCIt and also by the Foundational Model of Anatomy (FMA) for its anatomical decomposition [Dameron et al. 2006]. Marquet et al. also developed an ontology based on the NCIt for automatic classification of glioma tumors using the WHO grading system. Their ontology contained 243 classes (234 of them corresponding to NCIt classes) which correctly classified simulated tests and graded correctly ten clinical reports out of eleven used on the test for clinical data [Marquet and Dameron 2007]. The links mentioned on both manuscripts for downloading the ontologies were not active at the time of this writing.

The TNM ontology [Boeker et al. 2014] is a thorough representation of the TNM concepts for breast cancer using OWL-DL with SRI expressivity. The focus there was representation of the clinical meaning of each concept: T, N and M, with links to the Foundational Model of Medicine [Rosse and Jr. 2003]. They depict how to represent the tumor, the lymph node, distant metastasis, the organ locations specified and the tumor invasion pattern. Complete as it is, there is no rules for inference of clinical stage, nor the concepts related to the latter.

In this work we present an ontology for allowing inference of the TNM clinical stage of tumors, based on given values of T, N, M, the ICD-O topographic code and other information. This ontology should provide annotations with the original descriptions from the reference, and links to the NCIt ontology wherever applicable.

2. Materials and Methods

The first step was to identify the most common topographies on A. C. Camargo Cancer Center patients. Upon interview with an oncologist expert, we created a list of the ten most relevant topographies for research on this institution. We used the TNM 6th edition, because most of the relevant databases in the institution used this version of the coding system.

To achieve the goal of a fast-computing ontology, we kept its expressivity at the bare minimum while preserving the intended meaning of concepts. We used only subclass, intersection, equivalence, disjunction between classes, and object properties. As seen on Figure 1, the ontology is divided in four files (Figure 1): the main ontology, with the general TNM concepts and the imports of all others; the ICD-O topography, with the topographic classes referred by the TNM; a file with the annotations and finally a file with the clinical stage inference axioms.

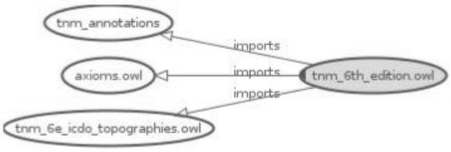

Figure 1. TNM Ontology components and imports diagram.

The concepts for representing T, N, M and CS were created as an hierarchy of classes; the root concept TNM_6th_edition, and its direct subclasses T, N, M and EC (the portuguese acronym for CS). There are subclasses that describes the general classification for all tumors, according to the introduction of the TNM reference. There may be an additional level of subclasses for representing concepts such as T1b or CS IIIa (as defined in some topographies such as breast cancer). We called all those the general staging classes. See Figure 2.

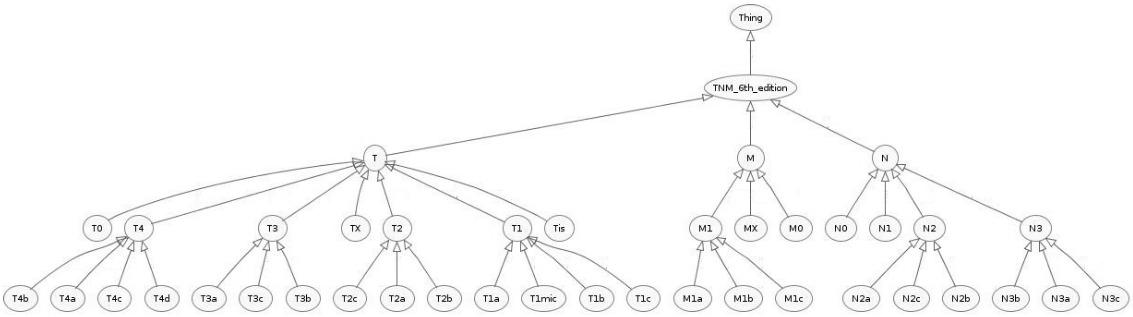

Figure 2. Class hierarchy for TNM concepts.

As the clinical stage rules depends on the tumor topography, the axioms for inference would need reference to ICD-O topography concepts. We could not find any ICD-O ontology available, and it was beyond the scope of our work to create one. However, as ICD-O topographic codes were based on ICD-10 cancer codes, we reused an ICD-10 ontology, available on the BioPortal[2]. We kept only the C00-C80 range of codes, removed some undefined codes within this range (such as C43, C78 and C79) and added C42 (as described in the ICD-O introduction). We also changed the ontology namespace and changed the label annotation property to skos:label. Reference to the prior ontology was kept. In Figure 3 there is a depiction of the ICD-O ontology.

To represent actual patient data, there should be an instance of class Patient, related to one or more instances of class tumor. In order to use the ontology to represent data, an instance representing the tumor should be created and related to subclasses of T, N, M, CS and ICD-O Topography classes. Following the TNM guidelines for staging, a patient with two primary tumors should be represented as one instance of a patient linked to two tumor instances; however, a patient with one tumor that metastasised should have only one tumor instance. The patient instance should be linked to the tumor instances by an object property.

A tumor should not belong to more than one topography class. First, it does not make clinical sense: a tumor should be located on a specific location or organ. It may

[2] `http://bioportal.bioontology.org/ontologies/ICD10`

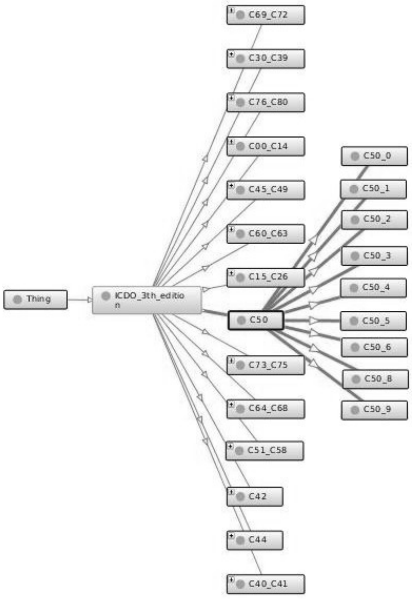

Figure 3. Class hierarchy for ICD-O concepts.

happen to spread itself to neighbour tissues or the precise location maybe be dubious (such as the gastroesophageal junction). In these cases the most probable tumor location should be selected and linked to the instance. The ICD-O Topography ontology states disjunction axioms for all their classes, preventing a tumor instance to belong to two topographic locations at once.

As each topography has different definitions for individual values of the general staging classes T, N and M, we created a script to parse a text file and create a RDF/XML file defining specific staging classes and inference axioms for a pair of T, N or M values and one topography, plus annotations using rdf:Description annotation property. We manually created text files based on the TNM definitions. The axioms are subclasses relating the specific staging classes to the intersection of one general staging class and one topography class.

Whenever a corresponding NCIt concept was available, it was linked to the specific staging class by the property owl:equivalentTo (see Figure 4). Not all concepts defined on TNM were present on NCIt, for instance, the T4 for Breast Cancer.

$$C50 \sqcap M1 \sqsubseteq C50\_M1 \equiv NCIt : C49009$$

Figure 4. Relation between an annotation from the current ontology and a NCIt class.

The standard procedure at the A.C. Camargo Cancer Center is to encode the TNM staging and the ICD-O topography during clinical attendance. As a result, structured information about the clinical stage is not promptly available in its databases. Based on this, we use the previously constructed inference axioms that considered the values of T, N, M and ICD-O to infer the clinical stage (CS) values.

The format starts with a first line containing the name of the determined clinical stage class. The second line contains one or more topography classes, which are linked to that clinical stage class and separated by a space character. The other lines have a relation of conjunction between the group T, N and M with each specified ICD-O topography. See Figure 5 for an excerpt of these axioms.

$$C50 \sqcap Tis \sqcap N0 \sqcap M0 \sqsubseteq BreastCancer\_CS\_0$$

$$C50 \sqcap T1 \sqcap N0 \sqcap M0 \sqsubseteq BreastCancer\_CS\_I$$

$$C50 \sqcap T2 \sqcap N1 \sqcap M0 \sqsubseteq BreastCancer\_CS\_IIB$$

$$C50 \sqcap N3 \sqcap M0 \sqsubseteq BreastCancer\_CS\_IIIC$$

$$C50 \sqcap M1 \sqsubseteq BreastCancer\_CS\_IV$$

Figure 5. Axioms for inference of clinical stage (CS) based on ICD-O topography and T, N and M classes.

For testing purposes we created another ontology with subjects and patients and assignments to specific classes of this ontology. For each subject we included a topographic class which includes the TNM for each test according to the example below.

$$patientTest00100 : Patient \sqcap hasTumor\ value\ patientTest00100\_Tumor1$$

$$patientTest00100\_Tumor1 : C50 \sqcap Tis \sqcap N0 \sqcap M0$$

After the inference, we can check the TNM annotation classes and also the respective NCIt code class. Thus we reach the ontology objective informing the inferred class to their respective clinical staging. We created a script to generate 566 tests based on the text mappings, as instances of Patient class with exactly one Tumour instance related to it. There were one test for each possible combination of T, N, and other variables for which could be inferred a clinical stage. We created then two queries, one for assessing test instances without any clinical stage inferred (it should have none) and other listing the inferred plus the expected clinical stage for each test.

The software we used to create the ontologies was Protégé [3]. The scripts for the creation of OWL files based on text files were developed in Python. The inferences were computed using Pellet[4].

3. Results

The resulting TNM ontology is divided in four files: main TNM concepts, ICD-O topography, annotations and clinical stage axioms. The main TNM ontology contains the

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

[4] `https://github.com/complexible/pellet`

general staging classes and includes the other ontologies. The ICD-O topography ontology contains the topographic codes and superclasses (such as *C00-C14 - Head and Neck*), with English descriptions. The annotation ontology define the specific TNM classes (such as C50_T1 and C61_M0) and their corresponding description in Portuguese and English. Finally, the clinical stage axioms ontology define the logical axioms that allows the inference of clinical stage based on ICD-O topography and TNM values.

The consolidated ontologies have ALC (Attributive Concept Language with Complements) expressivity. It consisted of 4.382 axioms, 2.954 logical axioms, and 772 classes. It defines 1.690 subClassOf axions, 16 EquivalentTo axioms, 1.248 disjointClasses axioms and 643 AnnotationAssertion axioms. The ontology, the scripts and the text files used to generate it were released under the APACHE-2.0 [5] open source license and are available online at

`https://github.com/djogopatrao/tnm_ontology/tree/master/ontologies`

All 566 test instances were assigned a clinical stage, and only one was assigned two clinical stages. $PatientTest\_51$ was supposed to be assigned Prostate Cancer Clinical Stage I, however an additional concept, Clinical Stage II, was present. This is because the definition of those clinical stages, as stated on the original reference, is ambiguous; Clinical Stage I is defined as T1a, N0, M0 and G1 (Gleason 2-4, discreet anaplasia), while Clinical Stage II, among other definitions, can be T1, N0, M0 and any G. T1, for prostate cancer in the 6th edition of TNM, means ”Clinically inapparent tumor neither palpable nor visible by imaging, while T1a (a subconcept for the former) is defined as ”Tumor incidental histologic finding in 5% or less of tissue resected. Therefore, as T1a is also T1, so Clinical Stage II is also applicable, and the definition of Clinical Stages in the prostate section of TNM 6th edition contained an ambiguity, detected by means of the ontology.

4. Discussion

We successfully represented the desired TNM rules using an ontology with a simple expressivity profile. That will allow the classification of tumors to remain computable.

The NCIt and SNOMED CT ontologies provide the general concepts involved with tumor staging: the values and description for T, N, M and CS for each topography. However, NCIt does not contains all codes for all topographies. SNOMED CT, in the other hand, does not define which TNM edition their concepts refer to. Neither defined axioms for inferring the clinical stage.

The work by Dameron et al. focus at the anatomical decomposition of a single topography, whereas the present work approaches several topographies, focusing on inference of clinical stage. Besides that, there is no description of the final ontology in the mentioned paper and the links provided are not available [Dameron et al. 2006].

In the paper by Boeker et al, a very detailed description of breast cancer TNM definitions is formalized in a very expressive ontology. The main objective of their work seems to be the formal representation of clinical examination findings for each value of T, N and M, with links to the anatomical and tumoral invasion patterns concepts. That

[5] `http://www.apache.org/licenses/LICENSE-2.0`

allowed the analysis of inconsistencies and inaccuracies in the definitions of TNM itself [Boeker et al. 2014]. However, the ontology at the time of this writing does not include the clinical stage classes, and thus does not provide axioms for their inference. Moreover, this ontology high level of expressivity (SRI) would arguably be less efficient than ALC for a given A-Box.

The tests showed that the inference worked as expected, except in one case, in which the definition provided by the original reference is ambiguous. A related work [Boeker et al. 2014] also found similar ambiguities; this shows how ontologies can be used to prevent classification definition errors.

The presented ontology may be applied to perform validation of existing databases or classify tumors based on TNM values. The usage of relational database to ontology mapping software [Calvanese et al. 2011] [Bizer 2004] [Cullot et al. 2007] allows the usage of the present ontology and inference tools on relational databases, the *de facto* industry standard. As it provides annotations for the meaning of individual T, N and M values for each topography, it may also serve as a reference for physicians and cancer registry workers.

As future work, the presented ontology may be completed to include all topographies and alignment with the NCIt ontology. Alignments with the TNM Ontology [Boeker et al. 2014] may also be of interest. Currently, there are annotations in both Portuguese and English, and other languages may be added. The ontology may be updated to represent the TNM 7th edition, possibly representing an alignment between it and the 6th edition, which may help database migration efforts. Finally, the pathological stage and other modifiers (such as stage post treatment) may also be implemented.

5. Conclusion

We showed that the presented ontology accurately represents the descriptions and inference rules from the selected topographies, fulfilling the main objective of this work. It may be useful in a number of tasks involving tumor staging. It is open source, allowing scrutiny and contributions from the scientific community. It has means to be linked to other TNM ontology efforts and well-established vocabularies, increasing its interoperability. Finally, it is lightweight to compute, being a valuable tool to validate or complete TNM databases.

References

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., and Stein, L. A. (2004). OWL Web Ontology Language Reference. Technical report, W3C, http://www.w3.org/TR/owl-ref/.

Bizer, C. (2004). D2rq - treating non-rdf databases as virtual rdf graphs. In *In Proceedings of the 3rd International Semantic Web Conference (ISWC2004*.

Boeker, M., Faria, R., and Schulz, S. (2014). A Proposal for an Ontology for the Tumor-Node-Metastasis Classification of Malignant Tumors: a Study on Breast Tumors. In Jansen, L., Boeker, M., Herre, H., and Loebe, F., editors, *Ontologies and Data in Life Sciences*, number 1, pages B1–B5, Freiburg.

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., Rosati, R., Ruzzi, M., and Savo, D. F. (2011). The mastro system for ontology-based data access. *Semantic Web Journal*, 2(1):43–53. Listed among the 5 most cited papers in the first five years of the Semantic Web Journal.

Cullot, N., Ghawi, R., and Yétongnon, K. (2007). Db2owl : A tool for automatic database-to-ontology mapping. In Ceci, M., Malerba, D., and Tanca, L., editors, *SEBD*, pages 491–494.

Dameron, O., Roques, E., Rubin, D., Marquet, G., and Burgun, A. (2006). Grading lung tumors using OWL-DL based reasoning. In *9th Intl. Protégé Conference*, pages 1–4, Stanford, California.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199 – 220.

Langlotz, C. P. (2006). Radlex: A new method for indexing online educational materials. *RadioGraphics*, 26(6):1595–1597. PMID: 17102038.

Lindberg, D. A., Humphreys, B. L., and McCray, A. T. (1993). The unified medical language system. *Methods Archive*, 32(4):281–291.

Marquet, G. and Dameron, O. (2007). Grading glioma tumors using OWL-DL and NCI thesaurus. *AMIA Annual . . .* , pages 508–512.

Nelson, S., Johnston, W. D., and Humphreys, B. (2001). volume 2 of *Information Science and Knowledge Management*, chapter Relationships in Medical Subject Headings (MeSH), pages 171–184. Springer Netherlands.

Neuss, M. N., Desch, C. E., McNiff, K. K., Eisenberg, P. D., Gesme, D. H., Jacobson, J. O., Jahanzeb, M., Padberg, J. J., Rainey, J. M., Guo, J. J., and Simone, J. V. (2005). A process for measuring the quality of cancer care: The quality oncology practice initiative. *Journal of Clinical Oncology*, 23(25):6233–6239.

Rosse, C. and Jr., J. L. M. (2003). A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36(6):478 – 500. Unified Medical Language System.

Schmoll, H.-J. (2003). F.l. greene, d.l. page, i.d. fleming et al. (eds). ajcc cancer staging manual, 6th edition. *Annals of Oncology*, 14(2):345–346.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A. L., and Rosse, C. (2005). Relations in biomedical ontologies. *Genome Biol*, 6(5):R46–R46. gb-2005-6-5-r46[PII].

Sobin, L. and Wittekind C (2002). *Classificação de Tumores Malignos*. Wiley and Sons, New York, 6th edition.