An Extended SNOMED CT Concept Model for Observations in Molecular Genetics

James R. Campbell MD[1], Geoffrey Talmon MD[1], Allison Cushman-Vokoun MD PhD[1], Daniel Karlsson PhD[2], W. Scott Campbell PhD[1]
[1]University of Nebraska Medical Center, Omaha NE USA; [2]Department of Biomedical Engineering, Linköping University; Linköping, Sweden

Abstract

Molecular genetics laboratory reports are multiplying and increasingly of clinical importance in diagnosis and treatment of cancer, infectious disease and public health. Little of this data is structured and maintained in the EHR in format useful for decision support or research. Structured, computable reporting is primarily limited by nonavailability of a domain ontology for these data. The IHTSDO and Regenstrief Institute(RI) have been collaborating since 2008 to develop a unified concept model and ontology of observable entities – concepts which represent the results of observations. In this paper we report the progress we have made to apply that unified concept model to the structured recording of observations in clinical molecular genetic pathology including immunohistochemistry and sequence variant findings. The primary use case for deployment is the structured reporting of Cancer checklist and biomarker data as as developed by by the College of American Pathologists(CAP) and the Royal College of Pathology(RCP).

Introduction

Molecular genetic pathology is a new scientific frontier exploding on the practice of clinical medicine. President Obama pushed the issue to the fore when he announced the national agenda for research into personalized medicine[1]. Unfortunately the extensive work in developing and managing information in genetic research has not translated into ontologies of use in support and documentation of clinical practice. Of the reference terminologies cited by the Office of the National Coordinator(ONC) as required by the US healthcare information architecture[2], only LOINC[f]has significant content addressing observations in molecular genetics. Unfortunately the granularity of the 1331 LOINC observables in molecular genetics does not capture details of the laboratory methods or support needs of the domain ontology that would be of use in epidemiology, research and clinical decision making[4]. Laboring behind the scenes in worldwide terminology management, harmonization efforts by the IHTSDO and RI[5] have been quietly working to expand the expressivity and utility of observable entities and clinical findings for use in molecular genetic structured clinical data. An observable entity is a concept with semantic overlap between LOINC and SNOMED CT and can best be described as a conceptual model for the results of an observation – administrative, clinical, laboratory or otherwise. Although RI has served the informatics community for years providing LOINC codes for laboratory medicine and molecular genetic pathology, the lack of a domain ontology for these concepts means that the there is no terminology support for the query aggregation of these observations. The harmonization work underway has developed a candidate unified concept model for genetic observables but the application of that work is not intuitive. ` This paper reports one small part of that effort focused upon the challenging issues of genetic biomarker observations in anatomic pathology supporting diagnosis and management of cancer.

The volume and types of molecular genetic data appearing in clinical medicine is growing exponentially. CAP first published cancer report protocols (checklists)[6] in 1998 defining a minimum dataset for anatomic pathologists to report when they evaluate surgical specimens concluding a diagnosis of cancer. Currently CAP publishes 82 separate checklists for various tissue pathways. The checklists were expanded in 2013 to include tumor biomarkers which have become referent findings required for staging and planning treatment.

The number of genetic observations – regarding either the tumor or the patient – important to outcomes in cancer treatment has grown monthly. The science behind molecular genetic testing is also undergoing rapid development. Historically the clinical pathologist could only detect genetic sequence variations by analyzing extracted tumor DNA for specific mutations in single genes that code for the mutant proteins they formed, with limited coverage of these genes by such assays. With the successful mapping of the human genome and improved technology, high throughput sequencing of human and neoplastic genomes became possible with next generation sequencing(NGS) resulting in relatively rapid turn-around of higher volumes of genetic sequence data. Anatomic pathology today employs both protein and sequence data in pathologic tumor diagnosis, prognostication and to aid in selection of targeted therapeutic regimens. A primary use case for a domain ontology of observables is query support of findings aggregated by genotypic variant, tissue of origin or histologic appearance. Query use cases implied by this expectation might include: "Find all cases of malignancies that tested positive for the BRAF p.V600E (c.1799T>A) mutation"; "Find all patients who have tested positive for genetic predisposition to breast cancer"; "Find all cases of colon cancer in which no biomarker testing was performed"

The SNOMED CT concept model[7] consists of a constrained set of relationships and accompanying target value sets allowed for incorporation within a computable concept definition. Allowed relationships and value sets are specified for domains of SNOMED CT, usually individual hierarchies. A SNOMED CT concept must also have at least one supertype (IS_A) relationship linking the concept within the hierarchy and a fully specified (context free) name which is the universal term denoting the concept. Each linguistic implementation of SNOMED CT can have one primary term and as many synonyms as required. When the concept model applied to the modelled SNOMED CT content is insufficient to F*ully define* a concept, the concept is declared as *Primitive.* The SNOMED CT ontology is subjected to description logic classification prior to publication as an editorial quality check and to compute the inferred relationships implied by the application of the concept model. The concept model is not complete for all segments of SNOMED CT and for years the Observable entities hierarchy has been published as a hierarchy of *Primitives* employing only stated supertype relationships as asserted by the modeling team.

SNOMED expects that extensions of SNOMED CT will be authored, compliant with the concept model, that represent material necessary for parochial or research needs not appropriate for the International release of SNOMED CT. In the US, the National Library of Medicine(NLM) develops and maintains the US extension to SNOMED that is required for Meaningful Use compliance by EHR vendors. The University of Nebraska maintains the Nebraska Lexicon extension, dependent upon the US extension, which supports terminology needs of our Epic® implementation and terminology development we have been supporting for the community since 2004.

Concept model development

The Observables and Investigation Model Project was formed by the IHTSDO in 2008 to develop a computable concept model for the Observable entity domain and prepare a model for interoperation of content with LOINC data bases. In summer of 2015, the project convened a meeting of experts from the NLM, CAP and Health and Social Care Information Center(HSCIC) of UK to discuss details of application of the proposed concept model to the set of observables necessary to the structured reporting of cancer checklists and biomarkers. The deployment model for testing from that conference is pictured in Figure 1 for Observable entities. Each attribute for refining the meaning of a concept is shown along with the valuesets of target concepts that are supported as well as the associated cardinality of the relationship.

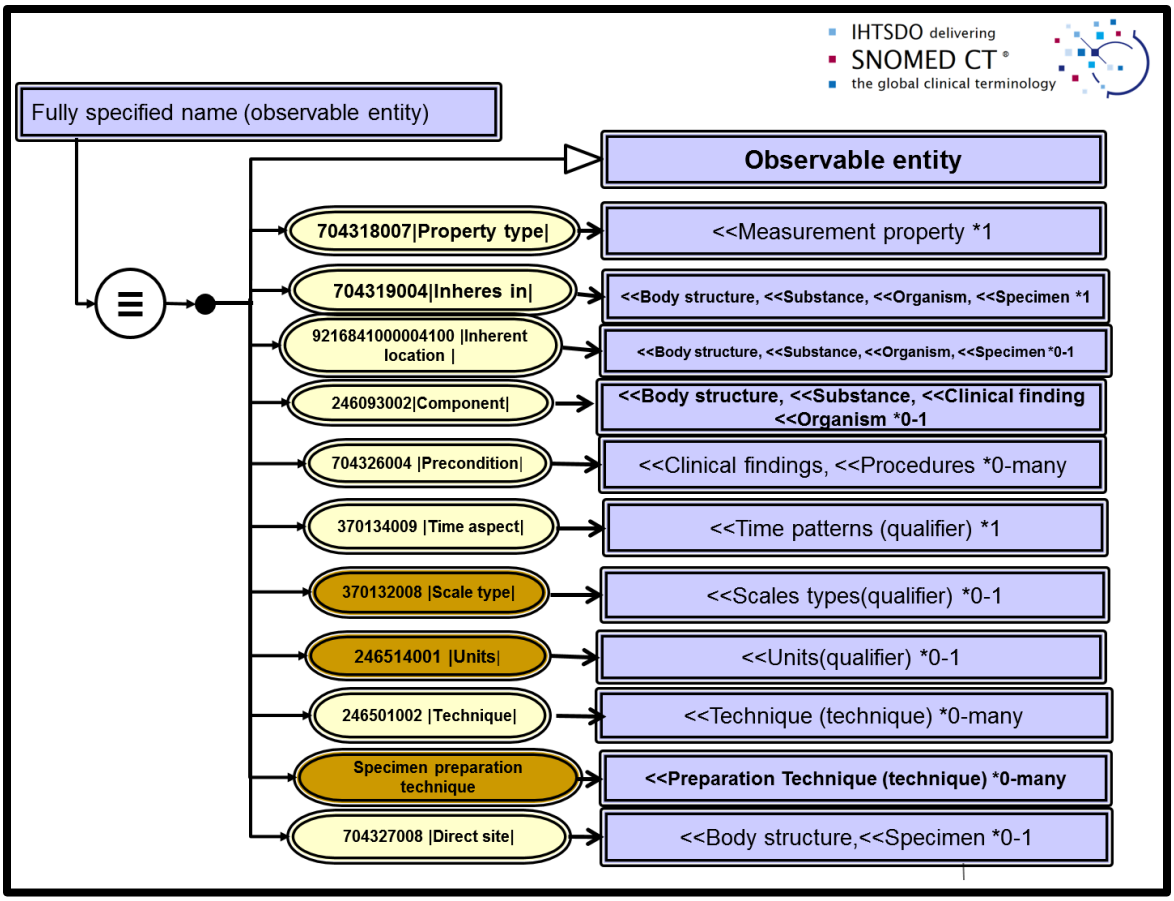

Figure 1. Harmonized observable concept model

Application of this model was reasonably straightforward for conventional observations in surgical pathology, requiring only model extensions of SNOMED to include 8 Properties and 4 Techniques. The Colon cancer checklist which we first modeled required 61 new observable entities for anatomic pathology. An exemplar from that development is in Figure 2.

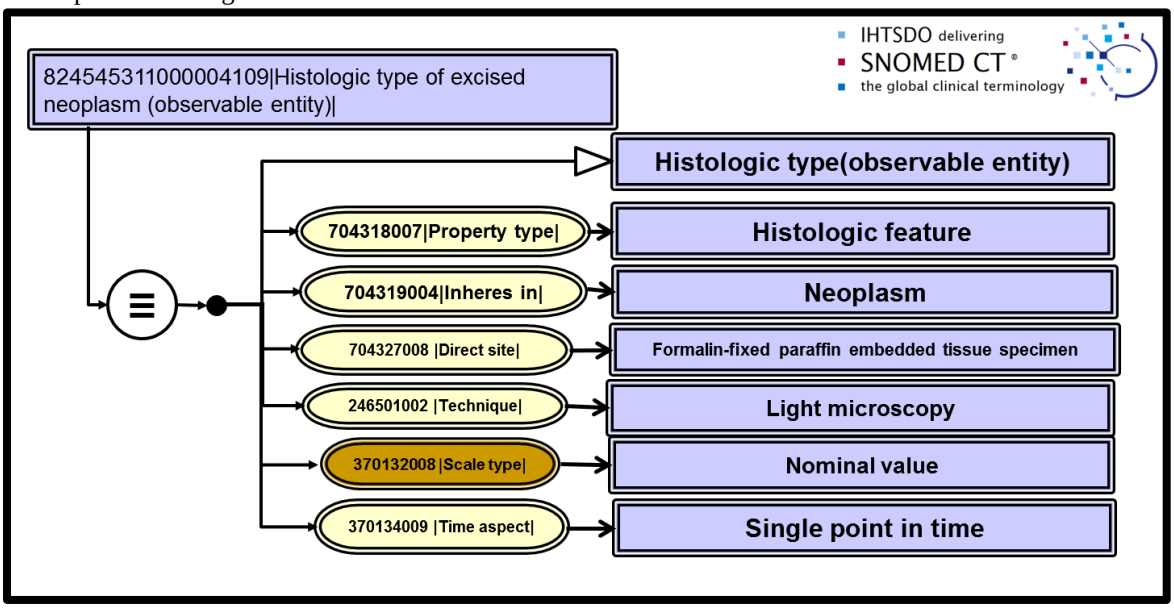

Figure 2. Anatomic pathology observable

Consensus on application of this model to observations in molecular genetic pathology could not be developed in our first several meetings and a deployment strategy was only achieved at the Montevideo meeting in fall 2015. Criticisms emerging from initial discussions included: a) excessive numbers of primitives, b) insufficient semantic granularity in genetic structures and c) failure to support both protein-based and sequence-based observations. In order to meet these challenges we proposed employing Concrete domains[8] in our model to uniquely define nucleotide sequence entities (genes, microsatellites and other nucleotide segments) and to support the complexity of sequence data expressed in *Variant Call Format*[9]. We proposed that a gene locus and other sequence based data could be fully defined in an expanded concept model specifying the genome reference, parent chromosome and nucleotide sequence addresses and modeled as cellular substructures, specifically nucleotide sequences . These data were defined using the reference naming of the Human Gene Nomenclature Committee[10] and the genetic datasets which it cross references such as Ensemble[11]. An example for the B-RAF proto-oncogene is included in figure 3. This particular definition employs sequence address data from the Genome Reference Consortium GRCh38 release in GenBank[12]. Human protein products of genetic transcription are currently implemented as primitive concepts and under discussion as to concept model definition.

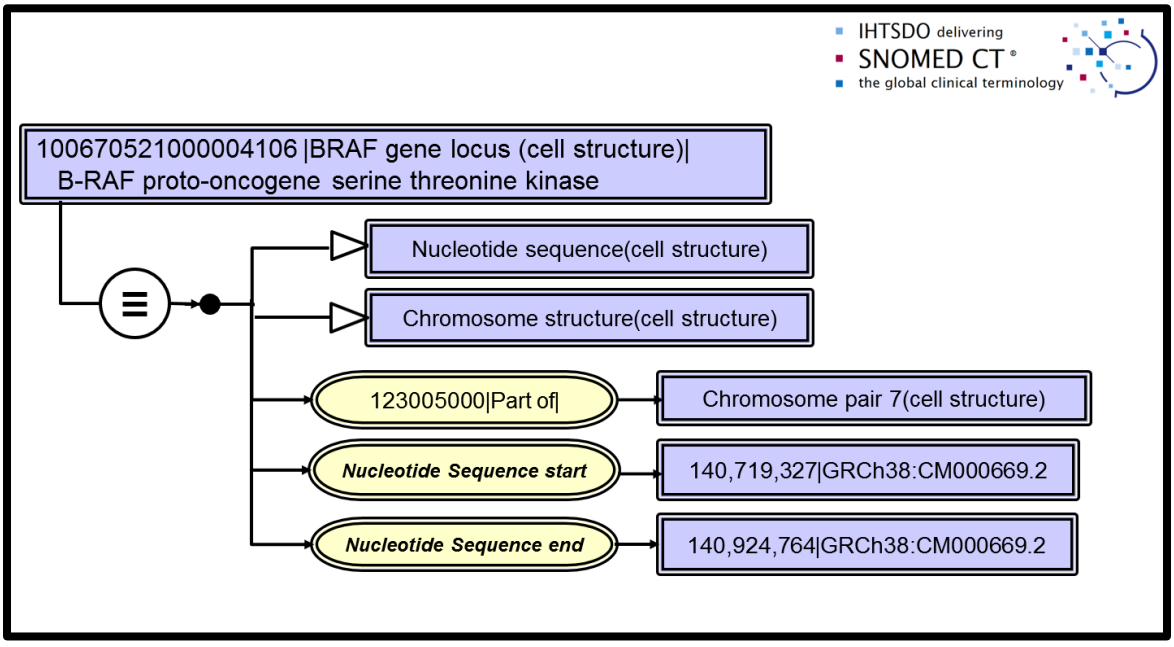

Figure 3. Cellular structure model for BRAF gene locus

Once the reference genetic material was fully defined in SNOMED, immunohistochemistry observations could be fully defined once we extended Techniques and Properties for molecular genetic pathology procedures. An example of an immunoperoxidase staining analysis for BRAF protein in a surgical tumor specimen is shown in figure 4. Immunoperoxidase and nuclear sequencing techniques are much more specific than the molecular procedures currently employed in LOINC 2.54 concept definitions. Therefore the observables concepts we have modeled for CAP checklists are generally subtypes (children) of molecular genetic observables now in LOINC.

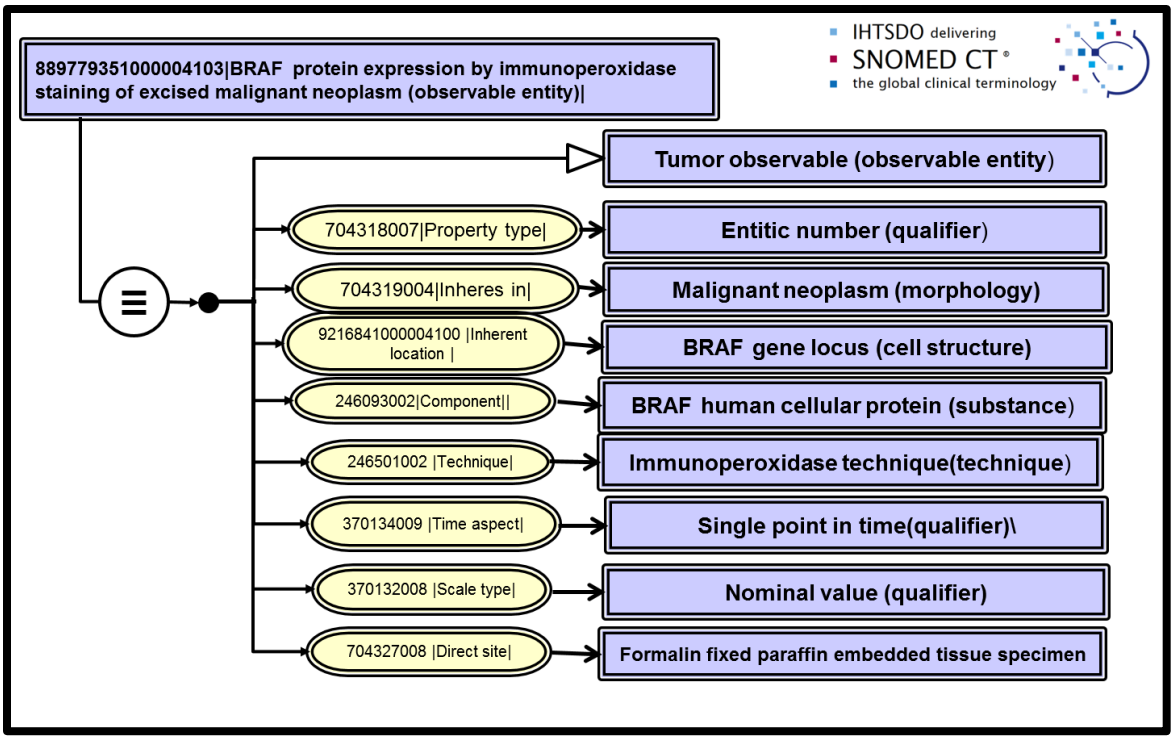

Figure 4. Immunohistochemistry observable: BRAF protein expression by immunoperoxidase staining

Nucleotide sequence observables were a particularly thorny issue we faced since input from our pathologist specialists required that we be able to store observations in our research databases with complete sequence data. An industry standard for identifying sequencing results has become the *Variant Call Format*[9]. These data files issue findings of sequence variants when compared to a reference standard genome and are typically ASCII files of a few kilobytes. Once again, employing Concrete domains allowed us to extend findings data and to deploy *Has value* attributes in place of *Has interpretation* in an extended concept model for Clinical findings. Figures 5 and 6 show the Observable entity model for sequence data of the BRAF gene locus along with a positive clinical finding for the BRAF p. V600E(c.1799T>A) mutation detected in the excised tumor.

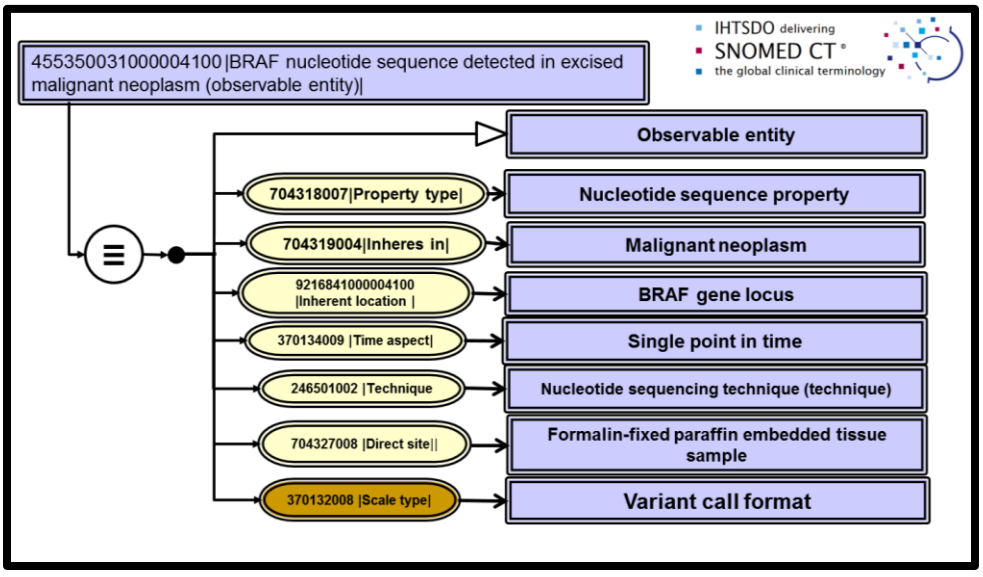

Figure 5. Nucleotide sequence observable for BRAF gene locus

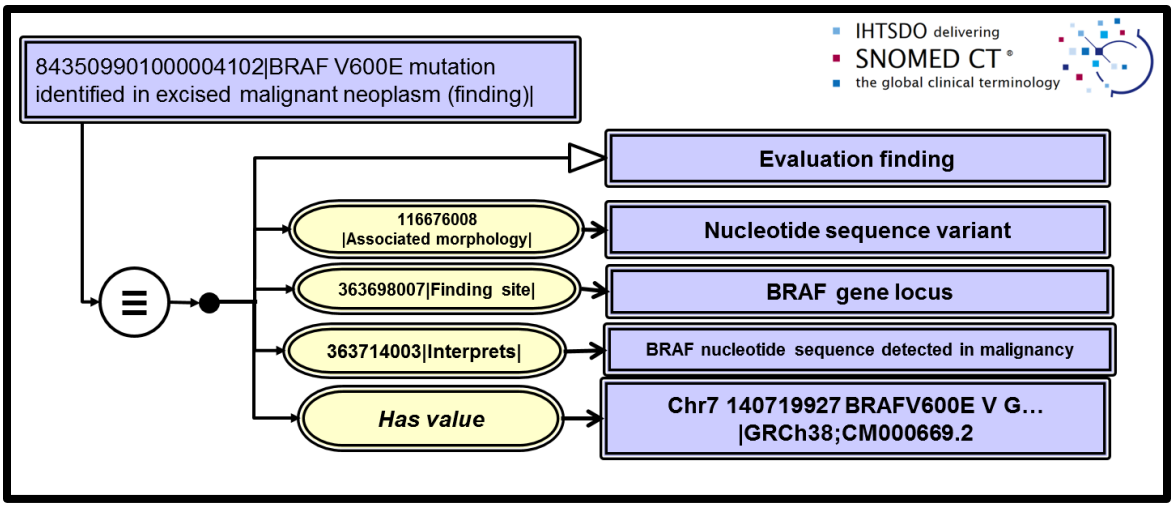

Figure 6. Observable for detection of BRAF V600E mutation in cancer

Results

Our deployment use cases are the 82 cancer and tumor biomarkers checklists published by CAP and supplemented by review with the RCP. There are many repeating observations across these checklists but our experience with deployment of our model for colon cancer required expansion of the Machine Readable Concept Model[13] for the SNOMED CT hierarchies of Observables, Body structure and Clinical Findings. The magnitude of the new conceptual content in our Nebraska SNOMED CT extension required for colon cancer is summarized in table 1. This lists the number of new concepts by hierarchy which we developed for anatomic pathology and molecular genetic pathology in colon cancer. The number of primitive concepts we required for each domain is also listed. This work has proceeded in collaboration with NLM, RI, IHTSDO, CAP and the RCP and was published on the NLM Knowledge Source Server in conjunction with the US extension to SNOMED CT in April 2016. It is available to all those with a UMLS license.

Table 1. Extension concept inventory for colon cancer checklist

SNOMED CT hierarchy	Anatomic Pathology Concepts/Primitives	Molecular Genetic Concepts/Primitives
Observable entities	61/1	32/3
Body Structures	10/9	29/3
Clinical findings	6/2	7/3
Procedures	2/1	0
Techniques	4/4	7/7
Property types	8/8	2/2
Scale types	0	9/9
Situations	1/0	0
Attributes	2/2	3/3
Qualifiers	2/2	0
TOTALS	88/29	87/30

Conclusion

We have presented in summary form the results of a two year collaboration with clinical and terminology standards developers which represents only a snapshot of harmonization work between the IHTSDO and RI in process since 2008. This report focuses just on Observable entities for molecular genetic observations which are a challenging but important component of a comprehensive terminology model for twenty-first century medicine. We have deployed this model for testing and evaluation by other informatics centers and published the work in collaboration with the NLM with the expectation and agreement that we will scale the work across the rest of molecular genetic pathology observations for cancer and expand the work into microbiology and human germline genetic disease testing.

The model we present interacts and draws upon NCBI resources and ontologies supported by the Genome Reference Consortium[14] expecting that the science in this field will continue to rapidly evolve and that the role of SNOMED CT is not to originate or control this data, but to develop and manage the ontology for representation and organization of genetic observational data as it demonstrates relevance for the practice of clinical medicine.

References

1. www.whitehouse.gov/precision-medicine
2. https://www.healthit.gov/policy-researchers-implementers/interoperability
3. {HYPERLINK "http://www.loinc.org"}
4. Simpson RW, Berman MA, Foulis PR et all. Cancer Biomarkers: the role of structured data reporting. Arch Pathol Lab Med. 2015;139:587-593; doi: 10.5858/arpa.2014-0082-RA.
5. LOINC-IHTSDO joint agreement text: http://www.ihtsdo.org/resource/resource/104
6. College of American Pathologists Cancer Protocols and Checklists [http://www.cap.org/apps/docs/cancer_protocols/protocols_index.html]
7. SNOMED CT Editorial Guide: January 2015 International Release. International Health Terminology Standards Development Organization. 2015. http://www.snomed.org/eg.pdf
8. Representation of nu8mber in SNOMED CT. International Health Terminology Standards Development Organization. 2010. https://confluence.ihtsdotools.org/download/attachments/15795743/
9. Variant Call Format Specifications. {HYPERLINK https://vcftools.github.io/specs.html}
10. HUGO Gene Nomenclature Committee. http://www.genename.org/
11. Ensembl Genome Browser 83. {HYPERLINK http://www.ensembl.org/}
12. Genbank repository for GRCh38. {HYPERLINK http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
13. SNOMED CT Technical Implementation Guide. IHTSDO. 2014. p463. {HYPERLINK http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_TechnicalImplementationGuide_Current-en-US_INT_20140813.pdf}
14. Genome Reference Consortium. {HYPERLINK http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/}

Learning objectives:

1) Appreciate the structure and harmonized concept model for deployment of a domain ontology for Observable entities
2) Understand the challenges facing the development of structured coded observations in molecular genetic pathology