Journal of the American Medical Informatics Association, 0(0), 2017, 1–8
doi: 10.1093/jamia/ocx097
Research and Applications

Research and Applications

A computable pathology report for precision medicine: extending an observables ontology unifying SNOMED CT and LOINC

Walter S Campbell,[1] Daniel Karlsson,[2] Daniel J Vreeman,[3] Audrey J Lazenby,[1] Geoffrey A Talmon,[1] and James R Campbell[4]

[1]Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, NE, USA, [2]Department of Biomedical Engineering, Linköping University, Linköping, Sweden, [3]Regenstrief Institute, Indiana University School of Medicine, Indianapolis, IN, USA and [4]Department of Internal Medicine, University of Nebraska Medical Center, Omaha, NE, USA

Corresponding Author: Walter S Campbell, Department of Pathology and Microbiology, University of Nebraska Medical Center, 985900 Nebraska Medical Center, DRC2 8064, Omaha, NE 68198-5900, USA. E-mail: wcampbel@unmc.edu. Phone: 402-559-9593

Received 11 April 2017; Revised 21 July 2017; Editorial Decision 9 August 2017; Accepted 28 August 2017

ABSTRACT

Background: The College of American Pathologists (CAP) introduced the first cancer synoptic reporting protocols in 1998. However, the objective of a fully computable and machine-readable cancer synoptic report remains elusive due to insufficient definitional content in Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) and Logical Observation Identifiers Names and Codes (LOINC). To address this terminology gap, investigators at the University of Nebraska Medical Center (UNMC) are developing, authoring, and testing a SNOMED CT observable ontology to represent the data elements identified by the synoptic worksheets of CAP.

Methods: Investigators along with collaborators from the US National Library of Medicine, CAP, the International Health Terminology Standards Development Organization, and the UK Health and Social Care Information Centre analyzed and assessed required data elements for colorectal cancer and invasive breast cancer synoptic reporting. SNOMED CT concept expressions were developed at UNMC in the Nebraska Lexicon© SNOMED CT namespace. LOINC codes for each SNOMED CT expression were issued by the Regenstrief Institute. SNOMED CT concepts represented observation answer value sets.

Results: UNMC investigators created a total of 194 SNOMED CT observable entity concept definitions to represent required data elements for CAP colorectal and breast cancer synoptic worksheets, including biomarkers. Concepts were bound to colorectal and invasive breast cancer reports in the UNMC pathology system and successfully used to populate a UNMC biobank.

Discussion: The absence of a robust observables ontology represents a barrier to data capture and reuse in clinical areas founded upon observational information. Terminology developed in this project establishes the model to characterize pathology data for information exchange, public health, and research analytics.

Key words: SNOMED CT, LOINC, Ontology, cancer synoptic reports, interoperability

BACKGROUND

The surgical pathology report is the summative assessment written by a pathologist to provide the basis for diagnosis and treatment of cancer. Concerted efforts by professional societies to move the format of the pathology report from a narrative to a structured format to ensure consistent and complete reporting of pathology data are ongoing.[1,2] The College of American Pathologists (CAP) produced its first cancer-reporting protocols as a synoptic reporting tool in 1998.[3] Shortly thereafter, CAP began publishing electronic versions of the cancer protocols as electronic Cancer Checklists for incorporation into information systems. Other international pathology societies produce similar protocols.[4–6] While the adoption of synoptic reporting has increased and is often mandated by regional or national authorities,[7–9] synoptic reporting will not achieve its maximum potential for patient care, clinical decision support, and secondary reuse by researchers and public health agencies until the data elements are reported in computable form.[10] A fundamental barrier to computable pathology data is a gap in the structure and content of standardized, controlled terminologies[11] essential for efficient and effective computation of pathology observations.[12–14] The purpose of this investigation is to develop and enhance the necessary computable terminology elements within the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) and Logical Observation Identifiers Names and Codes (LOINC) standardized terminologies to precisely represent the pathology data specified by the synoptic report. Successful terminology development in this domain addresses a gap in synoptic data representation and prepares synoptic data for clinical decision support in precision medicine, analytics, research, and public health use cases.

About clinical EHR terminology

Research into terminologies for an electronic health record (EHR) began 5 decades ago with the assumption that billing classifications such as the International Classification of Diseases, Ninth Revision, Clinical Modification would serve all needs. This hypothesis was studied and discarded as clinically incomplete and nonexpressive. A variety of competing controlled terminology resources were developed in the years following, each addressing the needs of a niche in the EHR.[15] A series of studies in the 1990s[16,17] culminated with an evaluation by the National Committee for Vital and Health Statistics commissioned by the secretary of Health and Human Services as directed by the Health Insurance Portability and Accountability Act. This evaluation[18] concluded that features of reference terminologies for recording of medical record data were required to assure scalability and interoperability of EHR deployment in the United States. A reference terminology is a conceptually based coding system that meets requirements of uniqueness, nonambiguity, and historicity, supplemented by a knowledge base of conceptual relationships that define the concepts and provide pragmatic information of use for querying and decision-making. In the years since that study, the mathematics of formal ontologies have matured, and we have seen an evolution of the domain ontology[19,20] as an architecture that employs all the features of the reference terminology but is further supported by a well-defined and semantically consistent concept model and maintained with logical consistency and rigor by a description logic classifier. A domain ontology supports complex data queries of EHR data, has robust relationships to support inference, and provides best interoperation due to its sound mathematical underpinnings. In this paper, references to "computable terminology" are meant to specify a terminology resource with the features of a domain ontology.

About the synoptic pathology report

A pathology synoptic report format consists of a series of observations posed as question-answer pairs, where each question characterizes a particular aspect or feature of the malignancy to be used for staging, treatment, and an estimation of prognosis of the cancer. Anatomic pathology (AP) observations assess physical and morphologic alterations of tissue, and molecular pathology (MP) observations address subcellular changes in the genetic or protein structure of the cells.[21,22] Synoptic questions semantically conform to the "Observable" entity hierarchy of the SNOMED CT concept model, as well as to the intended semantics of LOINC. LOINC[23] is specified for representing laboratory test order questions, and SNOMED CT[24] is mandated for populating categorical answer value sets within Meaningful Use guidelines as issued by the Office of the National Coordinator.[25,26]

Despite the depth of each terminology in many domains, it is impossible with current releases of either terminology to precisely and reproducibly report structured pathology data for clinical care, quality improvement, public health, or research. This means that basic research questions are difficult and laborious to answer. For example, consider a clinician attempting to identify all female breast cancer patients with high-grade tumors negative for estrogen receptors, progesterone receptors, and human epidermal growth factor receptor 2 for the purpose of introducing new therapies. Questions that require detailed genomic information are even more difficult to research. Consider a researcher seeking to analyze the survival data for patients who have a BRAF or KRAS gene mutation and whose invasive tumor originated in the colon, breast, or pancreas and involved regional lymph nodes but did not directly extend outside of the primary organ. Lacking efficient, precise reporting tools, these queries become manual, resource-intensive tasks. If the data were reported with computable terminologies, clinicians, researchers, and epidemiologists could make more effective and efficient use of the rich trove of information contained in synoptic pathology reports rendered in the course of daily health care delivery.

The terminology gaps of SNOMED CT and LOINC for observables were initially demonstrated by the Reporting Pathology Protocols (RPP) studies sponsored by the US Centers for Disease Control and Prevention.[12,13] The RPP projects investigated pathology reporting based on the CAP synoptic cancer worksheet for colorectal cancers in 2005 and for breast, prostate, and melanoma malignancies in 2009. These studies recommended that standard computable terminology be bound to each data element in the cancer report and specifically recommended the use of LOINC and SNOMED CT.[14] However, due to limitations of the concept models for LOINC and SNOMED CT observables, the RPP projects finally concluded that a link to the original data-collection instrument, specifically the CAP cancer protocol version and question-answer pair, should be maintained to provide the necessary context to interpret the data.

The LOINC definitional model is based upon 5 parts that define aspects of the LOINC term, but the relationships between LOINC parts are not defined. Therefore, queries of aggregation and subsumption are not possible using the LOINC coding system. The SNOMED CT concept model is a polyhierarchy that adheres to the principles of concept orientation, formal definitions, and multiple granularities[11] that support queries by attribute and subsumption.

However, until recently no concept model was agreed for the Observable entity semantic axis, so there was an insufficiently detailed definition in SNOMED CT to precisely represent Observables in general, and pathology observables in particular. Hence, the RPP study concluded that deficiencies of content and expression in the available reference terminologies would lead to ambiguous representation of pathology data.

In 2013, the Regenstrief Institute and the International Health Terminology Standards Development Organization (IHTSDO), curators of LOINC and SNOMED CT, respectively, reached a long-term cooperative agreement.[27] The collaborative work initiated the SNOMED CT Observable and Investigation Model Project, which serves as a working group to develop, test, and deploy an ontology-based definitional structure of all observables. The Observables project developed an extension to the SNOMED CT concept model for observable entities that increased the expressivity and specificity sufficiently to support full definitions of Observable concepts across the diverse subject matter of the SNOMED CT and LOINC terminologies.[28]

Extending these widely adopted terminologies is of great value and use to the international medical and research communities. This paper describes a project to author Observables content capturing the data elements contained in AP and MP synoptic reports. The technical approach to modeling this content is described, as well as the extensive international collaboration of advisors that guided the work, the deployment of the coded terminology within the research databases and EHR at the University of Nebraska Medical Center (UNMC), and some of the lessons learned from this project.

MATERIALS AND METHODS

The SNOMED CT concept model for observables

In the SNOMED CT concept model, observable concepts are found as subtypes in the 363787002|Observable entity| hierarchy. The SNOMED CT harmonized Observables concept model forms the basis for defining meaning and linking LOINC terms with the ontologic structure of SNOMED CT.[29] Within the harmonized concept model, the SNOMED CT attributes populate relationships in the form of attribute-value pairs that define features of the Observable concept. The March 2017 version of the Observables concept model[28] serves as the basis for all concept modeling in this project.

Selection of data elements for modeling

To capture and represent AP and MP data in computable form, investigators at UNMC in collaboration with IHTSDO, the National Library of Medicine (NLM), CAP, and the UK National Health Services (NHS) Health and Social Care Information Centre worked to develop, test, and deploy harmonized LOINC–SNOMED CT content found in AP and MP synoptic reports. In September 2015, a meeting of the International Pathology and Laboratory Medicine Special Interest Group of the IHTSDO convened in London to solicit a broad base of input to ensure reproducible and valid concept authoring for the broader AP community. Attendees included anatomic pathologists and molecular pathologists from the UK and the United States, representatives from CAP, NLM, NHS, the Health and Social Care Information Centre, terminologists, and IHTSDO editorial leadership. Participants analyzed synoptic data elements contained within the CAP colorectal worksheet version 3.3.0.0[30] and the CAP breast cancer worksheet version 3.0.0.0.[31] Analysis of MP concepts entailed review of the CAP colorectal biomarker worksheet version 1.2.0.0[32] and the CAP breast cancer biomarker worksheet version 1.0.0.0.[33] The UK Royal College of Pathology tissue pathway protocol data elements supplemented the CAP colorectal worksheet analysis.

Synoptic worksheet review and development of observable modeling recommendations

UNMC investigators inventoried "questions" contained in the CAP worksheets and presented each distinct element to the meeting attendees. Participants then attempted to describe the clinical meaning and intent of each data element. The exercise included developing a fully specified name (FSN) for each data element.[34] Within the SNOMED CT concept model, an FSN is a context-free description of the concept that states its precise meaning, including an assertion of its semantic domain or SNOMED hierarchy. Formation of an FSN is one required element for development of SNOMED CT concepts and definitional expressions, as it informs the terminologist of the precise clinical meaning of the concept being modeled.

Participating consultant terminologists trained by the IHTSDO and other team members analyzed the meaning of the concepts and developed SNOMED CT definitions employing the Observables concept model. As the data elements from the synoptic reports were modeled, terminologists and pathologists reviewed each element in order to ensure fidelity between the concept definition and the clinical intent of the authored term.

A design consensus was reached in London for the modeling of more than 100 histopathology concepts for breast and colorectal malignancies. The design model was confirmed by the Observables working group at the IHTSDO business meeting in October 2015. Consensus for MP concept representation and design was reached at the IHTSDO business meeting in April 2016. The design templates developed from these sessions formed the basis for subsequent terminology authoring of SNOMED CT Observables.

Terminology modeling and mapping approach

UNMC investigators modeled, authored, and tested concepts required for a comprehensive representation of AP and MP synoptic reports using a terminology-authoring environment that supports IHTSDO protocols. The Nebraska Lexicon© (IHTSDO extension reference 1000004) extension namespace[34,35] is copyrighted by UNMC and maintained by the Snow Owl® ontology development platform (B2i Healthcare, Singapore) employing the FaCT++[36] description logic classifier. We deployed the Observables concept model[28] in the machine-readable concept model integrated within Snow Owl and modeled Observables concepts that were required, accompanied by mappings to LOINC codes for all AP and MP concepts.

Each observation (question) included in the CAP structured cancer reports that has nominal or ordinal scale type is linked to a list of "answer" codes, from which the pathologist chooses a value. The project team reviewed the sets of answers to each selected question and attempted to find appropriate concept matches from the appropriate hierarchy of SNOMED CT. The SNOMED CT International edition of January 2016 and the US extension published by NLM in March 2016 were used for all mapping and concept modeling and updated with each SNOMED CT version release. Questions defined with quantitative or narrative scale types do not require value sets.

Table 1. New concepts developed for anatomic pathology and molecular pathology of colorectal and breast cancer by SNOMED CT hierarchy

SNOMED CT hierarchy	No. of new AP concepts	No. of new MP concepts	Total new concepts
Observable entities	61	32	93
Body structures	10	29	39
Clinical findings	6	7	13
Techniques	4	7	11
Property types	8	2	10
Scale types	0	9	9
Situations	1	0	1
Substances	0	11	11
Attributes	2	3	5
Qualifiers	2	0	2
Total	94	100	194

Pilot implementation at UNMC

The Cerner Copath® AP laboratory information system (Cerner Corp., Kansas City, MO, USA) is used by the pathology department at UNMC to develop and report AP and MP tests and observational data. To test the capability of the production systems at UNMC to handle the terminology-rich synoptic structure, investigators developed Copath® data entry templates for synoptic worksheets, including all CAP-required synoptic data elements, and bound them to LOINC and SNOMED CT codes. Health Level-7 (HL7) v2.3.1–formatted messages transmitted encoded worksheets to downstream clinical systems, including a clinical data warehouse and tissue biobank repository. HL7 interface with the EHR (Epic®, Verona, WI, USA) is currently being deployed and tested.

RESULTS

Terminology mapping and authoring

Existing LOINC and SNOMED CT content was examined for published codes that might capture the synoptic data elements. Review of existing LOINC content that related to these "questions" revealed that published meaning often implied a concept that was a more general supertype of the observable represented by the synoptic data elements. When published LOINC or SNOMED CT observables exactly captured the meaning of the use case, investigators employed them and modeled their meaning using the harmonized concept model. A total of 41 LOINC terms were required for colorectal cancer, of which 25 did not exist in LOINC. Breast cancer reports required 53 new LOINC terms of the total 57 terms included on the worksheet. As expected, all but 3 preexisting SNOMED CT observable entity concepts required modeling for definition. Only observable entity concepts pertaining to American Joint Commission on Cancer tumor staging remain without any defining SNOMED CT attribute value pairs, due to licensing restrictions specific to the joint commission.

A total of 243 existing SNOMED CT concepts were employed, of which 80% (194) were in the clinical finding and body structures hierarchies. Where there were gaps in the existing LOINC and SNOMED CT content, new concepts were authored as necessary to accurately represent the data elements in the synoptic worksheets. A total of 61 new Observables concepts were authored for histopathologic assessment of colorectal and breast cancers, and 32 were authored for biomarkers. The numbers of concepts authored across all SNOMED CT semantic types are shown in Table 1. All new observable concepts have been submitted to the LOINC committee for assignment of LOINC codes. A comprehensive listing of concepts and associated value sets for each CAP synoptic worksheet are available for review, including annotated CAP worksheets, at www.unmc.edu/pathology/informatics/tdc. SNOMED CT content can be downloaded with a valid Unified Medical Language System user account from this site.

Table 2. New property concepts developed for pathology observables. Property concepts created to represent anatomic pathology data. All concepts are children of the concept 118598001|Property of measurement (qualifier value)|.

New AP Concepts for Property of Measurement
160161921000004107|Morphology (qualifier value)|
644113361000004102|Histologic feature (qualifier value)|
169429731000004101|Histologic grade (qualifier value)|
257717701000004105|Histologic invasiveness (qualifier value)|
353715521000004107|Entitic integrity (qualifier value)|
372886811000004101|Location property (qualifier value)|
582585561000004109|Anatomic location property (qualifier value)|
733834701000004104|Radial direction property (qualifier value)|

Modeling and authoring in AP

In AP, investigators found that SNOMED CT content lacked sufficient expressivity to completely define new Observable entities. In particular, SNOMED CT did not have concepts that could accurately populate the 704318007|Property type (attribute)|. Therefore, one of the more significant additions to SNOMED CT identified during the AP modeling was the need for new Property type qualifiers, as enumerated in Table 2.

Microscopic tumor invasion or extension into adjacent tissues required for cancer staging provides a representative example of a concept authored for AP content in this project. SNOMED CT content for tumor extension by direct growth consisted of 370052007|Status of invasion by tumor (observable entity)| and its 17 descendant concepts. However, all of the concepts were primitive and did not capture accurate meaning of the use case concept. Therefore, researchers authored a local extension concept, 89000100004107|Status of microscopic invasion of excised colon malignant neoplasm (observable entity)|, with a stated definition as shown in Figure 1. The CAP worksheet element is reproduced in Figure 2A, and the encoded elements and value sets are shown in Figure 2B.

Modeling and authoring in MP

CAP biomarker (MP) synoptic worksheets contained fewer clinical questions, thus fewer observable entity concepts were developed for MP. The majority of new SNOMED CT concepts authored for MP reflected a set of shortcomings of the SNOMED CT content required for genomic observables. SNOMED CT concepts for gene loci, nucleotide sequences, sequence variants, and proteins required development. Investigators modeled gene loci as subcellular body structures and defined them by the chromosome location. An example of the model for the BRAF gene locus[37] as currently deployed is shown in Figure 3.

The Human Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC)[38] is the authoritative body for naming human genes. LOINC uses HGNC's terminology to name genes and the Human Genome Variation Society's (HGVS)[39,40] syntax to code the sequence variants of interest.[41] HGNC editorially reviews

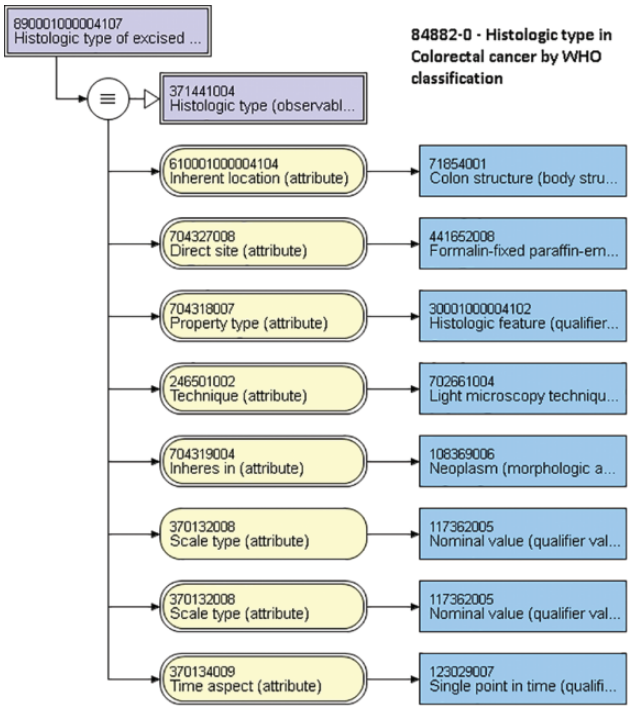
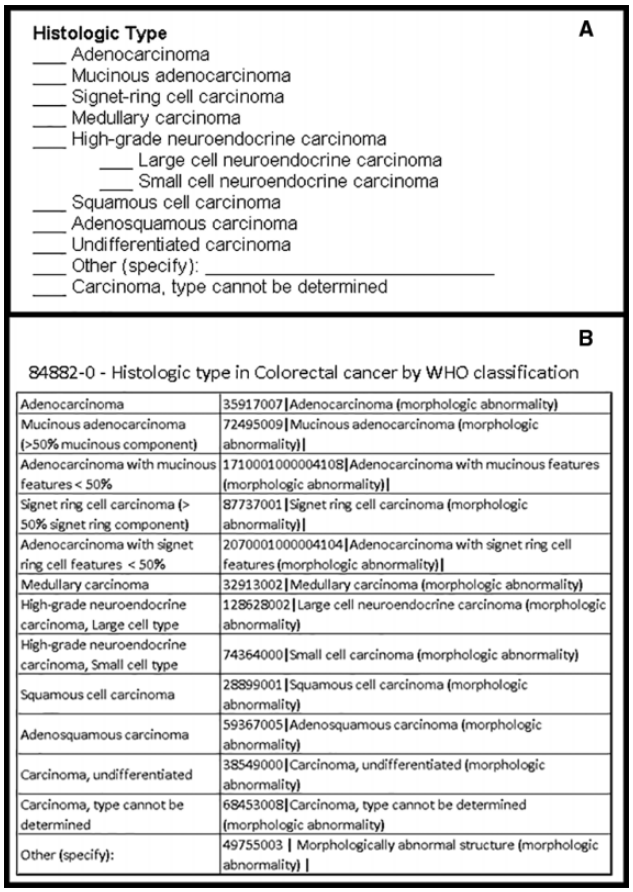


Figure 1. Stated definition of "Histologic type of excised colon neoplasm (observable entity)" using the SNOMED CT diagram specifications and the corresponding LOINC code for data exchange.

SNOMED CT: Rectangle with single line border = primitive concept; rectangle with double line border = fully defined concept; rectangle with rounded edges and double line border = attribute concept; open-headed arrow = IS_A relationship. Arrow points to parent concept; circle with 3 parallel lines = concept is equivalent to; arrow with solid head indicates directional link between concepts (https://confluence.ihtsdotools.org/display/DOCDIAG/Diagramming+Guideline).



Adenocarcinoma	35917007 | Adenocarcinoma (morphologic abnormality)
Mucinous adenocarcinoma (>50% mucinous component)	72495009 | Mucinous adenocarcinoma (morphologic abnormality) |
Adenocarcinoma with mucinous features < 50%	171000100000 4108 | Adenocarcinoma with mucinous features (morphologic abnormality) |
Signet ring cell carcinoma (> 50% signet ring component)	87737001 | Signet ring cell carcinoma (morphologic abnormality) |
Adenocarcinoma with signet ring cell features < 50%	2070001000004104 | Adenocarcinoma with signet ring cell features (morphologic abnormality) |
Medullary carcinoma	32913002 | Medullary carcinoma (morphologic abnormality)
High-grade neuroendocrine carcinoma, Large cell type	128628002 | Large cell neuroendocrine carcinoma (morphologic abnormality)
High-grade neuroendocrine carcinoma, Small cell type	74364000 | Small cell carcinoma (morphologic abnormality)
Squamous cell carcinoma	28899001 | Squamous cell carcinoma (morphologic abnormality)
Adenosquamous carcinoma	59367005 | Adenosquamous carcinoma (morphologic abnormality)
Carcinoma, undifferentiated	38549000 | Carcinoma, undifferentiated (morphologic abnormality)
Carcinoma, type cannot be determined	68453008 | Carcinoma, type cannot be determined (morphologic abnormality)
Other (specify):	49755003 | Morphologically abnormal structure (morphologic abnormality) |

Figure 2. Data elements and terminology binding example. (**A**) Data elements for histologic type of colorectal neoplasm from the College of American Pathologists' protocol for the examination of specimens from patients with carcinomas of the colon and rectum. (**B**) Encoded value set for CAP protocol data elements in Figure 2a. LOINC code used for observable entity (question). SNOMED CT concepts used for answer value set.

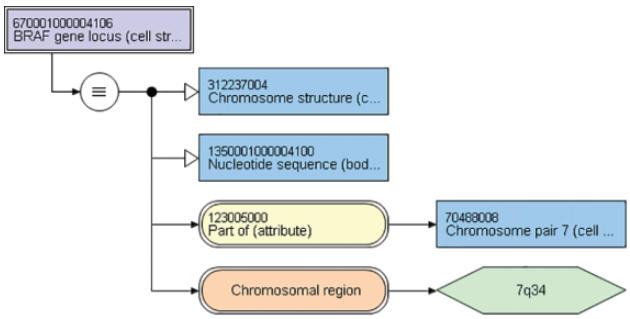


Figure 3. Stated definition of BRAF gene locus, primitive concept.

progress in the science of human genomic discovery, standardizes naming of gene loci, annotates and cross-references gene nucleotide addresses and genetic mutations, and provides tool sets to analyze gene sequence information. Consultant pathologists urged the investigators not to duplicate the genetic detail in the HGNC model as SNOMED CT artifacts. Instead, UNMC investigators created a map set linking SNOMED CT concepts for gene loci to HGNC IDs and the integrated HUGO reference data from the National Center for Biomedical Ontology.[42] The map set for gene names further provides users with a representational state transfer (REST) application program interface (API) Uniform Resource Locator to the Extensible Markup Language–formatted HGNC data for each named gene to facilitate cross-references between SNOMED CT and HGNC (Table 3).

The draft model does not attempt to completely represent the domain of human genetics. Instead, the intent was to create a construct that frames and interprets the complex data of molecular analysis and captures meaning defined as relevant by the clinical standard of care. Observable entities within MP were defined using the known high-level architecture of the genome; that is, the DNA molecule, chromosome, and gene. Only known, clinically validated, named genes, their protein products, and variants of documented clinical significance as presented in the CAP check sheets were modeled in SNOMED CT.

Using the building block concepts of genes, nucleotide sequences, proteins, and MP techniques, UNMC terminologists developed observables for biomarker reporting, including observations from sequencing, immunohistochemistry (IHC), or other molecular techniques. The consultant pathologists recommended that concepts be modeled such that the ontology would provide comprehensive retrieval results irrespective of the assessment technique employed. Some observables assess for sequence variation of a gene with protein tests such as IHC, while others use direct genetic sequencing procedures, such as pyrosequencing or Sanger sequencing, or they

Table 3. SNOMED CT to HGNC map example for BRAF gene locus. SNOMED CT concept for BRAF gene locus with map to Human Genome Naming Committee concept using REST API.

SNOMED CT Concept	Mapped HGNC Identifier and REST API to HGNC Metadata
100670521000004106|BRAF gene locus|	HGNC:1097^http://rest.genenames.org/fetch/symbol/BRAF

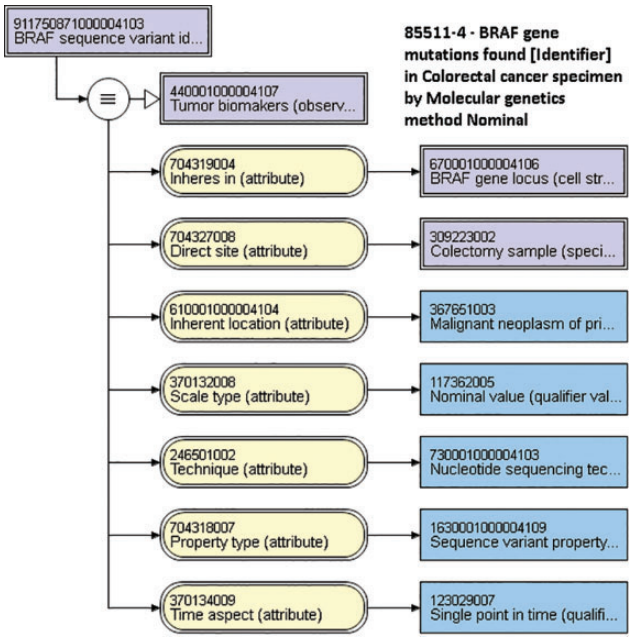


Figure 4. Concept model for a nucleotide sequence variant of the BRAF gene of a colorectal cancer specimen, including the corresponding LOINC code for data exchange.

```
OBX-1|OBX|
OBX-2|OBX sequence|
OBX-3|CWE|
OBX-4|Variant detected in BRAF locus^85511-4^LN|
OBX-5|NM_004333.4(BRAF):c.1799T>A(p.Val600Glu)|
```

Figure 5. Partial HL7 version 2.x message for a molecular pathology finding of a BRAF v600e mutation in a colorectal cancer specimen. OBX4 contains the LOINC observable code, and OBX5 contains the HGVS string for the specific mutation. The specific gene, accession number, functional coding sequence change, and mutant protein changes are indicated.

employ deoxyribonucleic acid (DNA, RNA, mRNA) probes, such as in fluorescent in situ hybridization. An understandable and useful domain ontology model should support querying of all genetic data specifically for one or across all of these techniques. For this reason, the investigators agreed in discussion with the Observables project team that molecular observables evaluating nucleotide sequences should consistently *Inhere in* the gene and have *Direct substance* of the related protein if IHC testing is being done as a proxy for the gene nucleotide structure. (As defined by SNOMED CT, 704319004|Inheres in (attribute)| "specifies the independent continuant which bears the quality, and on which the dependent quality (of this observable) depends" [browser.ihtsdotools.org] [ie, what is assessed or measured]).

Given the large number of possible gene variants that can be detected by MP techniques and the evolutionary nature of clinical knowledge as to the significance of these nucleotide polymorphisms, pathology consultants on the project advised that the model should support concise sequence observation data at varied levels of granularity. This challenged the investigators to consider commonly used information artifacts in the field of MP to meet this user requirement. UNMC terminologists chose to use HGVS nomenclature as syntax for discrete sequence variant observations and variant call file format for aggregate output from multichannel sequencers. Using a *Scale type* of Nominal, gene sequence data can be transmitted in HL7 2.x message format with the OBX-4 segment containing the appropriate observable concept and OBX-5 containing the variant specific data using HGVS terms. An observable 911750871000004103|BRAF sequence variant identified in excised malignant neoplasm of colon (observable entity)| is shown modeled in Figure 4, with a sample HL7 message reporting the V600E mutation in Figure 5.

Using the extended observables model, the UNMC investigators successfully produced fully encoded synoptic worksheet summative reports from the Copath® surgical pathology system. Since October 2016, 49 colorectal cancer resections and 89 invasive breast cancer resections have been fully encoded and characterized in the UNMC cancer registry database and pathology information system.

DISCUSSION

This development project began with the intent to solve the UNMC research community's need to query AP and MP data in order to identify research candidates. Basic researchers also required a method to identify available tissue specimens that exhibit specific characteristics for hypothesis testing and translational research. The pathology reporting methods in use at UNMC employed structured reports and synoptic reporting for many conditions, but extracting AP and MP information from the laboratory information system was cumbersome, at best. It often required manual chart review or unwieldy natural language searches with unpredictable results. Although the AP system provided a mechanism to bind structured terminology to synoptic data, the current versions of SNOMED CT and LOINC did not contain AP or MP content with a domain ontology serving precise querying of case data. Without such a domain ontology, data retrieval is unpredictable and method-dependent. The AP and MP content authored in this project provides the necessary underpinnings to perform advanced data queries using the full semantic strengths of the SNOMED CT concept model.

Rapid scientific advances in the understanding of neoplastic disease at the molecular level and the drive for precision medicine[43] create a burden on clinical terminologies to further serve these endeavors. There were several previous efforts to bind molecular and genetic information into the clinical record.[44–50] Most recently, the HL7 Fast Healthcare Interoperability Resources (FHIR) genomics implementation guide contained Standard for Trial Use[51] in release 3, which holds great promise. The content authored as part of this project will enhance the semantic representation of data with a FHIR construct as adoption of the standard increases. However, the terminology development of this study also conforms to the broadly

used HL7 version 2.x standard and can be readily adopted in current health information systems. The content authored in this project extends previous work, including the FHIR genomics implementation guide, by adding clinical interpretation data in a highly compressed, computable form and is consistent with the tenets of Masys et al.[52] for integrating genomic data into the EHR. They write that molecular and genomic data should be presented to clinicians in a format that is conducive to clinical use, decision support, and patient care while simultaneously retaining complete or source representation of genomic data for use in discovery.

The modeling approach developed in this work compresses the key MP findings of clinical importance into a compact format suitable for use in patient care while retaining detailed genetic information with additional linkages. The application resolves the terminological ambiguity and resolves the conundrum faced by the RPP project.[12,13] The content modeled as an ontology employing the LOINC–SNOMED CT harmonized concept model integrates with standard terminologies that are widely deployed in commercial EHR systems. The nucleotide sequence details are connected to the standard vocabularies using a reference to the HGNC standard for gene naming, which avoids overburdening the terminology or the EHR. By representing sequence data using HGVS, the approach effectively retains a complete representation of genetic sequence data for oncogenes, tumor-suppressor genes, and variants of undetermined significance in a manageable size. In this way, scientific reference libraries maintained by Gene Ontology and HGNC sources are linked explicitly to clinical and research datasets and can be exploited for use throughout the health care enterprise.

Limitations and future directions

Thus far, representation of MP content has been limited to the data elements represented on the colorectal and invasive breast cancer worksheets. Additional investigation that tests the model's ability to represent a broader array of MP data and the concomitant clinical concepts would be valuable. Comparatively, the AP terminology domain is more stable. The model developed in this study for AP can be employed to represent a large spectrum of content with few changes.

The level of effort and resources necessary to undertake the content development project for AP and MP are significant but tractable. The scope of the project is bound by the clinical content represented in the CAP and similar cancer datasets. Second, the content developed may be large in number of new concepts, but the concept definitions follow a limited number of definitional patterns. All MP observable concepts were developed using one of 2 distinct patterns. AP content required <20 patterns. These patterns will repeat when developing content for the scope of this effort.

While UNMC is uniquely poised to assist in this objective, input from outside experts in pathology and terminology is necessary to improve the end product for large-scale use. The initial IHTSDO International Pathology and Laboratory Medicine project has expanded to include many more expert pathologists and terminologists from the United States, the UK, Australia, and Sweden.

CONCLUSION

This study developed a novel application of convergent terminology in order to enhance the computability of structured AP and MP reports in a way that maximizes utility for multiple communities of use in alignment with a national vision for the Learning Healthcare System.[53] The modeling approach provides a concise data representation of synoptic cancer checklist observations and demonstrates that the new terminology artifacts were implementable in research and clinical care systems. The AP and MP content authored in this project provides the necessary underpinnings to perform advanced data queries using the full semantic strengths of the SNOMED CT concept model while complying with the Standards and Interoperability Framework of the Office of the National Coordinator for Health IT.[26] By extending the foundation of existing internationally adopted terminology standards, this work can contribute to improved interoperability and computability of pathology data and patient care globally.

FUNDING

This work was support by National Institutes of Health grant number 1U01HG009455-01, Patient Centered Outcomes Research Institute grant number CDRN-1306-04631, and the Departments of Pathology/Microbiology and Internal Medicine at the University of Nebraska Medical Center.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

Each author contributed substantially to the design of the work and/or the analysis and interpretation of the data, the drafting of the manuscript, and approval of final version, and agree to be accountable for the accuracy and integrity of the work.

ACKNOWLEDGMENTS

The authors thank collaborators from SNOMED International, Regenstrief Institute Inc., the College of American Pathologists, Rajesh Dash, MD, Alexis Carter, MD, and Mary Kennedy and National Health Service Digital Health.

REFERENCES

1. Srigley JR, McGowan T, Maclean A, *et al*. Standardized synoptic cancer pathology reporting: a population-based approach. *J Surg Oncol*. 2009;99(8):517–24.
2. Markel SF, Hirsch SD. Synoptic surgical pathology reporting. Hum Pathol. 1991;22(8):807–10.
3. *An Overview of the College of American Pathologists Cancer Checklist*. www.cap.org/apps/docs/committees/cancer/cancer_protocols/Overview_CAP_Cancer_Checlkists_090115.pdf. Updated 2009. Accessed July 10, 2017.
4. *Cancer Datasets and Tissue Pathways*. www.rcpath.org/resource-library-homepage/publications/cancer-datasets.html. Updated 2016. Accessed July 10, 2017.
5. *RCPA Cancer Protocols*. www.rcpa.edu.au/Library/Practising-Pathology/Structured-Pathology-Reporting-of-Cancer/Cancer-Protocols. Updated 2013. Accessed July 10, 2017.
6. *International Collaborative on Cancer Reporting/Datasets*. www.iccr-cancer.org/datasets. Updated 2016. Accessed July 10, 2017.
7. *ACS Commission on Cancer Releases Updated Standards Manual*. www.facs.org/publications/newsscope/121815/cocmanual1218. Updated 2015. Accessed July 10, 2017.
8. *The Partnership Launches Electronic Synoptic Pathology Reporting Initiative (ESPRI) to Advance Pan-Canadian Standardized Cancer*

Pathology Reporting. www.partnershipagainstcancer.ca/the-partnership-launches-electronic-synoptic-pathology-reporting-initiative-espri-to-advance-pan-canadian-standardized-cancer-pathology-reporting/. Updated 2012. Accessed July 10, 2017.
9. *Cancer Protocols Frequently Asked Questions*. www.cap.org/web/home/resources/cancer-reporting-tools/cancer-protocol-frequently-asked-questions?_afrLoop=262923013927738#!%40%40%3F_afrLoop3D262923013927738%26_adf.ctrl-state%3Dvq1atfacv_98. Updated 2016. Accessed July 10, 2017.
10. Williams CL, Bjugn R, Hassell AL. Current status of discrete data capture in synoptic surgical pathology and cancer reporting. PLMI. 2015;7:11–22.
11. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf Med*. 1998;37(4–5):394–403.
12. Centers for Disease Control and Prevention. *Report on the Reporting Pathology Protocols for Breast and Prostate Cancers, and Melanomas*, RPP2. 2009.
13. Centers for Disease Control and Prevention. *Report on the Reporting Pathology Protocols for Colon and Rectum Cancers Project*, RPP1. 2005.
14. Centers for Disease Control and Prevention. *Report on the Reporting Pathology Protocols Project for Breast and Prostate Cancers and Melanomas – Executive Summary*. 2009.
15. Cimino JJ. Coding systems in health care. *Yearb Med Inform*. 1995;1(1):71–85.
16. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. For the Computer-Based Patient Record Institute's Work Group on Codes & Structures. *J Am Med Inform Assoc*. 1996;3(3):224–33.
17. Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, Warren J. Phase II evaluation of clinical coding schemes: Completeness, taxonomy, mapping, definitions, and clarity. CPRI Work Group on Codes and Structures. *J Am Med Inform Assoc*. 1997;4(3):238–51.
18. National Center for Vitals and Health Statistics. *Scope and Criteria for Selection of PMRI Terminologies*. 2002:1–98.
19. Schulz S, Balkanyi L, Cornet R, Bodenreider O. From concept representations to ontologies: a paradigm shift in health informatics? *Healthc Inform Res*. 2013;19(4):235–42.
20. Rector AL, Rogers J, Taweel A. Models and inference methods for clinical systems: a principled approach. *Stud Health Technol Inform*. 2004;107(Pt 1):79–83.
21. *Template for Reporting Results of Biomarker Testing of Specimens from Patients with Carcinoma of the Colon and Rectum*. www.cap.org/ShowProperty?nodePath=/UCMCon/Contribution%20Folders/WebContent/pdf/cp-colorectalbiomarker-14.pdf. Updated 2014. Accessed July 10, 2017.
22. Simpson RW, Berman MA, Foulis PR, *et al*. Cancer biomarkers: the role of structured data reporting. *Arch Pathol Lab Med*. 2015;139(5):587–93.
23. Regenstrief Institute. LOINC. www.loinc.org. Updated 2017. Accessed February 3, 2017.
24. SNOMED International. SNOMED CT worldwide. www.snomed.org/snomed-ct/snomed-ct-worldwide. Updated 2017. Accessed February 3, 2017.
25. Office of the National Coordinator for Health Information Technology, Department of Health and Human Services. Health information technology: Initial set of standards, implementation specifications, and certification criteria for electronic health record technology. *Interim final rule*. Fed Regist. 2010;75(8):2013–47.
26. Office of the National Coordinator for Health Information Technology. 2016 interoperability standards advisory. www.healthit.gov/sites/default/files/2016-interoperability-standards-advisory-final-508.pdf. Updated 2016. Accessed April 10, 2017.
27. IHTSDO. Cooperation agreement between international health standards terminology organization and Regenstrief Institute, Inc. http://loinc.org/collaboration/ihtsdo/agreement.pdf. Updated 2013. Accessed February 7, 2017.
28. Case J, ed. *SNOMED CT Editorial Guide*. January 2017 ed. London: SNOMED International; 2017.
29. *IHTSDO Observable and Investigation Model Project*. https://confluence.ihtsdotools.org/display/OBSERVABLE. Updated 2016.
30. Washington MK, Berlin J, Branton P, *et al*. Protocol for the examination of specimens from patients with primary carcinoma of the colon and rectum. *Arch Pathol Lab Med*. 2009;133(10):1539–51.
31. Lester SC, Bose S, Chen YY, *et al*. Protocol for the examination of specimens from patients with invasive carcinoma of the breast. *Arch Pathol Lab Med*. 2009;133(10):1515–38.
32. Bartley AN, Hamilton SR, Alsabeh R, *et al*. Template for reporting results of biomarker testing of specimens from patients with carcinoma of the colon and rectum. *Arch Pathol Lab Med*. 2014;138(2):166–70.
33. Fitzgibbons PL, Dillon DA, Alsabeh R, *et al*. Template for reporting results of biomarker testing of specimens from patients with carcinoma of the breast. *Arch Pathol Lab Med*. 2014;138(5):595–601.
34. International Health Terminology Standards Development Organization. *SNOMED-CT Technical Implementation Guide*. July 2012 International Release (US English) ed. Copenhagen: International Health Terminology Standards Development Organization; 2012.
35. https://confluence.ihtsdotools.org/display/REFSET/Requirements?preview=%2F6160816%2F6160916%2FSNOMED_CT_Namespace_Registry++OFFICIAL+20141021.pdf. Updated 2014. Accessed July 10, 2017.
36. FaCT++. http://owl.man.ac.uk/factplusplus/. Updated 2007. Accessed July 10, 2017.
37. Sithanandam G, Druck T, Cannizzaro LA, Leuzzi G, Huebner K, Rapp UR. B-raf and a B-raf pseudogene are located on 7q in man. *Oncogene*. 1992;7(4):795–99.
38. HUGO Gene Nomenclature Committee. www.genenames.org. Updated 2015. Accessed July 10, 2017.
39. den Dunnen JT, Dalgleish R, Maglott DR, *et al*. HGVS recommendations for the description of sequence variants: 2016 update. *Hum Mutat*. 2016;37(6):564–69.
40. den Dunnen JT. Sequence variant nomenclature. www.hgvs.org/varnomen. Accessed February 3, 2017.
41. Deckard J, McDonald CJ, Vreeman DJ. Supporting interoperability of genetic data with LOINC. *J Am Med Inform Assoc*. 2015;22(3):621–27.
42. Campbell JR, Talmon G, Cushman-Vokoun A, Karlsson D, Scott Campbell W. An extended SNOMED CT concept model for observations in molecular genetics. *AMIA Annu Symp Proc*. 2017;2016:352–60.
43. Obama B. *The Precision Medicine Initiative*. www.whitehouse.gov/precision-medicine. Updated 2015. Accessed July 10, 2017.
44. Hoffman M, Arnoldi C, Chuang I. The clinical bioinformatics ontology: a curated semantic network utilizing RefSeq information. *Pac Symp Biocomput*. 2005:139–50.
45. Hoffman MA. The genome-enabled electronic medical record. *J Biomed Inform*. 2007;40(1):44–46.
46. Jing X, Kay S, Marley T, Hardiker NR, Cimino JJ. Incorporating personalized gene sequence variants, molecular genetics knowledge, and health knowledge into an EHR prototype based on the continuity of care record standard. *J Biomed Inform*. 2012;45(1):82–92.
47. Sax U, Schmidt S. Integration of genomic data in electronic health records: opportunities and dilemmas. *Methods Inf Med*. 2005;44(4):546–50.
48. Deckard J, McDonald CJ, Vreeman DJ. Supporting interoperability of genetic data with LOINC. *J Am Med Inform Assoc*. 2015;22(3):621–27.
49. HL7. *HL7 version 2 Implementation Guide: Clinical Genomics; Fully LOINC-qualified Cytogenetic Model, release 1 (US Realm)*. www.hl7.org/implement/standards/product_brief.cfm?product_id=364. Updated 2017. Accessed February 3, 2017.
50. HL7. *HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-qualified Genetic Variation Model (US realm)*. www.hl7.org/implement/standards/product_brief.cfm?product_id=23. Updated 2017. Accessed February 3, 2017.
51. HL7. *FHIR Release 3 (STU) – Genomics Implementation Guidance*. www.hl7.org/FHIR/genomics.html#diagnosticreport-genetics. Updated 2017. Accessed July 10, 2017.
52. Masys DR, Jarvik GP, Abernethy NF, *et al*. Technical desiderata for the integration of genomic data into electronic health records. J Biomed Inform. 2012;45(3):419–22.
53. Institute of Medicine. *The Learning Healthcare System*. Washington, DC: National Academies Press; 2007.