Closing Terminology Gaps to Drive Precision Oncology
Jacqueline M. Clark, MS¹, Andrew K. Stewart, MA¹, Abel N. Kho, MD, MS¹²¹ Tempus Labs, Chicago, IL;² Northwestern University, Chicago, IL

Abstract

Precision oncology requires large standardized datasets to accurately characterize patient, tumor, treatments, and outcomes. Standardized terminologies provide unambiguous concept codes, but may not cover specialized domains. We characterized gaps in terminology coverage identified while mapping over 300,000 cancer concepts. Target specific treatments and condition specific variants accounted for the majority of gaps in terminology coverage. Coordination between standards development organizations and domain specific experts will be needed to close existing and future concept gaps.

Introduction

Advances in precision oncology promise to deliver tailored treatments to cancer patients for improved outcomes. However, access to standardized clinical data to accurately characterize patients, their disease, treatments, and outcomes remains a limiting factor. Several factors limit the ability to bring these data together for large scale analyses, including variably complete EHR data which requires significant manual data extraction, limited informatics resources, concerns of data privacy and governance, and inconsistent mapping of clinical data to widely recognized terminology standards such as ICD, ICD-O, LOINC, SNOMED or RxNorm.

Standardizing clinical data to consistent terminologies (concept mapping) is both art and science, sometimes with multiple choices available to represent a single clinical concept. Since 2017, the informatics team at Tempus has focused on transforming clinical data for oncology patients into standardized datasets to inform more precise cancer treatments. In this presentation we characterize and classify gaps in terminology coverage identified from 2 years of data to inform future terminology development efforts.

Methods

Both structured and unstructured data extracted from Electronic Health Records for patients using Tempus laboratory testing is abstracted through a standardized data pipeline with a focus on oncology relevant domains (e.g. cancer stage, treatment, outcomes, etc). Our team of terminology experts map concepts to standardized clinical terminologies including SNOMED, ICD 9/10/O, LOINC, and RxNorm, and also cross reference concepts with the National Cancer Institute Thesaurus (NCIt) when a code is not available in a primary terminology. Concepts without clear terminology coverage at the time of abstraction and concept mapping are flagged for creation of custom codes.

Results

Out of over 300,000 clinical concepts extracted from Electronic Health Record data, we identified 82 concepts without coverage from standardized clinical terminologies. We grouped uncovered concepts into 6 categories (Table 1). Beyond this subset of concepts deemed as immediately essential to cancer treatment, an abundance of granular data is not captured regularly due to the lack of detailed standardized codes.

<table>
<thead>
<tr>
<th>Type</th>
<th>Number</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target specific treatment</td>
<td>32</td>
<td>Anti-HER3 Monoclonal Antibody instead of Antineoplastic antibody (C129822)</td>
</tr>
<tr>
<td>Condition specific variant</td>
<td>24</td>
<td>Specific gene variant KIT p.K642E</td>
</tr>
</tbody>
</table>
Sequence of Line of Therapy | 10 | Ninth Line of Therapy (NCIt only has codes up to 8th line therapy)
---|---|---
Pathogenic effect of a specific mutation | 7 | Splice site deletion
Missing codes to represent specific exon numbers | 5 | Certain numbers and ranges are missing: 11-13, 25-26, 25-27, 34, 53
Generic tissue type descriptors | 3 | Tumor tissue vs. normal tissue
Specific Lab Test | 1 | EndoPredict

These 82 unique concepts represent a majority of over 141,000 custom codes by volume; e.g. a genetic test result of “overexpressed” with closest NCIT Code C18211 “Protein Overexpression” occurred over 800 times in reviewed patient records.

**Discussion**

Standardized terminologies continue to grow and evolve with the scientific communities they support. While many organizations maintaining standard terminologies allow the informatics community to request new content when insufficiencies are identified, release cadence typically lags behind the current state of the scientific knowledge. The interim use of an organized internal custom code system has become essential for this reason.

Our experience suggest several focused domains to guide terminology development efforts within the precision oncology community. In particular, specific codes to represent targeted cancer treatments (e.g. interventional drugs targeting specific receptors) as well as standard concept codes for pathologic molecular variants will likely be a significant area for growth and thoughtful concept modelling as new testing modalities emerge and more normal and pathologic variants are inevitably discovered. Because both the tests and test results which correlate with molecular and genetic variants are likely to increase at a rapid pace, adopting a strategy to represent cancer genetic testing and results modeled on prior terminology mapping efforts as espoused in the LOINC community may be an ideal approach.

Recently, cancer researchers have developed and evaluated ontologies to provide coverage of key concepts for solid tumor outcomes or adverse events of immunotherapies. Building on these efforts, we anticipate that domain gaps in terminologies will be best addressed through coordinated community efforts given the likely rapid future growth of precision oncology discoveries.

**References**