Building Cancer Diagnosis Text to OncoTree Mapping Pipelines for Clinical Sequencing Data Integration and Curation

Adhithya Narayanan¹ ², Umit Topaloglu, Ph.D. ³, Javier A. Laurini, M.D.² ³, Franck Diaz-Garelli, Ph.D. ² ⁴

¹University of North Carolina at Chapel Hill, Chapel Hill, NC
²Wake Forest Baptist Medical Center, Winston Salem, NC
³Montefiore Health System, Bronx, NY
⁴University of North Carolina at Charlotte, Charlotte, NC

Abstract

Precision oncology research seeks to derive knowledge from existing data. Current work seeks to integrate clinical and genomic data across cancer centers to enable impactful secondary use. However, integrated data reliability depends on the data curation method used and its systematicity. In practice, data integration and mapping are often done manually even though crucial data such as oncological diagnoses (DX) show varying accuracy and specificity levels. We hypothesized that mapping of text-form cancer DX to a standardized terminology (OncoTree) could be automated using existing methods (e.g., natural language processing (NLP) modules and application programming interfaces [APIs]). We found that our best-performing pipeline prototype was effective but limited by API development limitations (accurately mapped 96.2% of textual DX dataset to NCI Thesaurus (NCIt), 44.2% through NCIt to OncoTree). These results suggest the pipeline model could be viable to automate data curation. Such techniques may become increasingly more reliable with further development.

Introduction

Genomics testing has opened the door to new possibilities in oncology care¹ but also the advancement of clinical research², population health³ and building the learning healthcare system⁴. Genomic testing from providers such as Foundation Medicine, Caris and Guardant enable oncologists to differentiate mutations within tumors and go beyond histological information gathered from pathology reports⁵. This increased level of information allows the selection of specific treatment regimens tailored to each patient’s condition, mutations and tumor type, leading to better patient outcomes². Still, the ideal of the learning healthcare system⁴ requires clinical practices to go beyond care and learn from past experiences¹. Thus, data from cancer centers must be reused to improve future treatments through research via data reuse and aggregation⁶.

The American Association for Cancer Research’s (AACR) Project Genomics Evidence Neoplasia Information Exchange (GENIE) is an international research program involving data sharing among cancer centers. Its goal is to enable precision oncology research via clinical and genomic data integration from tens of thousands of patients for reuse and secondary analysis⁷. Cancer genomic data sharing has the potential to unlock data-driven insights that may expand existing knowledge about cancers and aid the development of therapies and treatments. This work also aims to support hypothesis generation to pave the way towards the future of cancer research⁸. However, data integration is a non-trivial and tedious task⁹ ¹⁰ ¹¹. For example, clinical and genomic data from different clinical sites with different workflows and different information technology infrastructure to record such information are often represented in different¹² and sometimes incompatible ways. This often requires the hand-curation of datasets for aggregation¹³.

The GENIE project uses the OncoTree standardized terminology⁷ to mediate these issues and standardize cancer DX data provide in clinical data and genomic data. OncoTree is a hierarchical structure of cancer diagnoses (DX), which classify DX descriptions by site and neoplasm type. Still, both clinicians and genomic reports often encode diagnostic data via free-text fields within clinical databases because of their complexity and ambiguity¹⁴ ¹⁵. This requires manual mapping of this free-text information to the OncoTree standard, which is both time-consuming and labor-intensive. The reviewer-based nature of the task is also likely to lead to inconsistencies and encoding errors. Given that the GENIE collaborative is composed of a growing number of cancer centers contributing data on a regular basis, the development of an automated method to standardize such mapping for integration and curation is crucial. Existing literature reports the development of biomedical named entity recognition (NER) tools capable of extracting facts about entities described in the text but also medical concepts qualifying them¹⁶ ¹⁷ ¹⁸. However, most available research work is focused on providing tools rather than building automated pipelines to achieve a specific goal such as mapping DX cancer data for integration and curation.
In this paper, we report our findings evaluating automated data pipelines capable of mapping cancer DX from clinical genomic testing reports to the OncoTree terminology based on concept extraction and NER tools. This work aims to systematize this repetitive, yet crucial task of the integration and curation of genomic data in the context of the GENIE project. We hypothesized that the mapping of textual cancer DX to OncoTree concepts could be automated using existing NLP packages and APIs, building off existing research work. We built two pipeline prototypes using NOBLE Coder\textsuperscript{20}, the Unified Medical Language System (UMLS) API\textsuperscript{21} and the OncoTree API\textsuperscript{22}. We tested the performance of our pipeline using free-text DX from cancer registry genomic testing reports from two genomic testing providers (i.e., Caris and Foundation Medicine [FM]). The outputs of this pipeline and the mappings were validated by a pathologist to provide and measure accuracy for our mappings. Our work contributes to standardizing and automating oncological data integration for research in within the GENIE project and beyond. These preliminary results aim to pave the way towards reliable automatic data curation\textsuperscript{13} as a means of supporting reproducibility in data reuse\textsuperscript{6,23}. Our automated mapping pipeline and its results will be shared across contributing GENIE sites to improve data reporting.

Methods

We extracted structured oncology DX data from the Wake Forest Baptist Medical Center’s cancer registry. We used distinct sample textual DX data from Caris and FM to develop a mapping algorithm to facilitate DX reporting to the GENIE collaborative. We developed this process by mapping textual DX using NOBLE Coder, an NLP tool\textsuperscript{20}, and the UMLS search API\textsuperscript{21} to extract concepts from free-text DX. We then mapped the resulting concepts to OncoTree by using the OncoTree API. The mapped output was assessed by a pathologist to determine clinical accuracy and confirm concordance with original text diagnosis. Our study was approved by Wake Forest University School of Medicine’s Institutional Review Board (IRB) before any data extraction or analysis.

Our pipelines consist of an NLP-based concept extraction module paired to the OncoTree concept mapping API (Figure 1). Our prototype development was based on simplicity and the ability to reuse existing tools to automatically map textual DX data to OncoTree concepts. Two the two simplest viable prototypes were considered and compared in this paper. For our first prototype, we used NOBLE Coder\textsuperscript{20} to pre-process our raw DX text and extract clinical concepts. We used the National Cancer Institute Thesaurus (NCIt)\textsuperscript{24} as the intermediary terminology between raw text and the OncoTree terminology\textsuperscript{25}. We used the default best match algorithm to minimize irrelevant results. To eliminate remaining irrelevant results, we filtered out irrelevant NCIt concepts by semantic type. For example, concepts such as “grade 1,” a “Classification” semantic type and “typical,” a “Qualitative Concept” semantic type were found to be irrelevant. Concepts types such as “Cell” and “Tissue” were also filtered out if there was a more specific semantic type within the NOBLE Coder output for the same text DX. We privileged the “Neoplastic Process” semantic type as this was the most relevant semantic type to the original DX based on preliminary review in most cases. On the other hand, we used the UMLS search API\textsuperscript{21} to map the raw textual DX to UMLS Metathesaurus concepts\textsuperscript{21} as an intermediate terminology. The pipeline begins with a string normalization process that includes removing any punctuation, extra spaces, and insignificant parts, such as NOS, in order to produce the most relevant results. We then queried the normalized strings using the exact match algorithm for the search API to return concepts in UMLS. In both pipelines, we used the OncoTree API to map the resulting concepts from the first mapping to relevant OncoTree concepts. The OncoTree Tumor Types API was used with the parameters type, which were “nci” or “umls”, and query, which corresponds to the NCIt or UMLS concept code. This segment of the pipeline produced corresponding OncoTree concepts in JavaScript Object Notation (JSON) form. Each JSON string contained consisted of the OncoTree codes, OncoTree diagnosis names, levels in the OncoTree hierarchy, and associated UMLS and NCI codes.

Figure 1. DX Mapping Pipeline Prototypes. Prototype 1 is shown at the top with the NOBLE Coder encoding module followed by filtering of semantic types and mapping to OncoTree concepts using the OncoTree API. Prototype 2 appears at the bottom with the string normalization process followed by UMLS Search API calls to extract UMLS concepts that are then mapped to OncoTree concepts via API calls.
We calculated concept coverage rates (i.e., how many raw DX descriptions could be covered by each DX mapping pipeline) as percentages that represented the ratio of the number of matches in each mapping to the number of initial diagnoses from their corresponding source. Specifically, we counted the number of raw DX descriptions returning concepts in NCIt after removal of irrelevant semantic type. We then calculated the ratio of this number to the total number of raw DX text samples. We used the same method to calculate coverage of concepts for the UMLS search API and for mappings completed through the OncoTree API in both pipelines.

We were only able to validate the clinical accuracy of one pipeline due to the preliminary stage of our work. We compared the performance of each pipeline in two ways. First, we compared the number of raw DX items that could not be mapped to each intermediary terminology (i.e., NCIt and UMLS). Comparing these raw numbers gave us an idea of the potential performance of each method. Then, we proceeded to compare the performance of both methods statistically. Specifically, we built count regressions\textsuperscript{26} to predict the number of concepts mapped to a raw DX entry from based on each pipeline. We tested for zero-inflation using Van den Broek’s zero-inflation test\textsuperscript{27} and for overdispersion by comparing means and standard deviations in our main outcome variables (i.e., number of concepts mapped) and using R’s AER package\textsuperscript{28}. The data comparing mappings to intermediary terminologies was neither overdispersed nor zero-inflated, for which we selected a traditional Poisson regression using R’s generalized linear model (GLM)\textsuperscript{29}. Data describing mapping results from each intermediary terminology was both overdispersed and zero-inflated, which led us to select a zero-inflated negative binomial regression. We used the pscl R package\textsuperscript{29} to build this model. To validate our best-performing pipeline’s output, a practicing clinical pathologist compared the raw text DX with their corresponding NCIt to the final OncoTree concepts to assess their accuracy. For the raw text DX, the mapping was considered a match when there was at least one concept in NCIt returned by NOBLE Coder that matched the diagnosis. The same process was used for OncoTree mappings. The rates of accuracy were calculated based on the number of tests and the number of accurate mappings.

Multiple software tools were used to carry out this analysis. Data extraction and preprocessing was done using a DataGrip software client (version 2019.1.3, JetBrains s.r.o., Prague, Czech Republic). Visual exploration and analyses were done using Tableau (version 10.2.4, Tableau Software, Inc., Seattle, WA). Our pipeline relied on the use of NOBLE Coder\textsuperscript{20} for the extraction of concepts from free text as well as the use of API querying from both the UMLS Search API (version, U.S. National Library of Medicine, Bethesda, MD) and OncoTree API (version 1.0.0, Memorial Sloan Kettering Cancer Center, New York, NY). We handled our datasets and API querying using Python (version 3.7.4, Python Software Foundation, Beaverton, OR) along with the Pandas library\textsuperscript{30} for data frame handling and JSON library to parse out JSON API outputs. All statistical analyses and data manipulation such as data scrubbing and reshaping were done in R version 3.6.1 and RStudio (version 1.3.1335, RStudio, Inc., Boston, MA). Statistical significance was set at p=0.05 for all models.

**Results**

Our initial dataset contained raw DX descriptions extracted from genomic testing reports. Each was a unique textual description of cancer DX from two genomic testing providers (i.e., Caris and FM) extracted from a cancer registry containing over 3500 individual patient records. The input data for our pipeline consisted of 193 distinct DX from Caris and 205 distinct DX from Foundation Medicine. These diagnoses were not in a standardized format, but free-text entries written by clinicians in order to avoid the potential ambiguity generated by DX codes in medical terminologies.

<table>
<thead>
<tr>
<th>Provider</th>
<th>Number of Initial DX</th>
<th>Number of DX by Concepts Returned in NCIt</th>
<th>Number of DX by Concepts Returned in UMLS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Caris</td>
<td>193</td>
<td>2</td>
<td>120</td>
</tr>
<tr>
<td>Foundation Medicine</td>
<td>205</td>
<td>0</td>
<td>39</td>
</tr>
</tbody>
</table>

Table 1. Caris and FM DX and extracted concepts in NCIt/UMLS. Zero, one, or multiple concepts were extracted in UMLS and NCIt from each DX provided.
We found slightly different mapping results while mapping our raw DX to our intermediate terminologies (i.e., NCIt and UMLS). On one hand, the NCIt mapping for the Caris dataset (193 DX records total) failed to map 2 DX (1.04%) to any concept in NCIt (Table 1). The remaining 191 DX were mapped to one or multiple concepts (Figure 2a). The NCIt mapping for the FM dataset (205 DX total) however, was able to map all raw DX to concepts to at least one concept. On the other hand, the UMLS mapping for the Caris dataset failed to map 40 DX (20.7%) to a concept in UMLS (Figure 2b). The remaining 153 DX were mapped to at least one concept. For the FM dataset of 205 DX the UMLS API failed to map the majority (Figure 2d), 148 DX (72.2%) to any concept in UMLS. 57 DX were mapped to at least one concept.

![Caris/FM to NCIt Mapping](image)

![Caris/FM to UMLS Mapping](image)

Figure 2. Distribution of concepts extracted through mapping of textual DX from Caris and FM to NCIt and UMLS. Mapping Caris and FM DX through NOBLE Coder or the UMLS search API yielded multiple concepts in some cases, due to splitting of the initial DX data or multiple search results.

Table 2. Caris and FM DX and extracted concepts in OncoTree from NCIt/UMLS terminologies. NCIt/UMLS concept unique identifiers for DX that returned at least once concept in NCIt and UMLS were queried using the OncoTree API.

<table>
<thead>
<tr>
<th>Terminology</th>
<th>Provider</th>
<th>DX in Termination</th>
<th>DX by Concepts Returned in OncoTree from Termination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>NCIt</td>
<td>Caris</td>
<td>191</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>Foundation Medicine</td>
<td>205</td>
<td>89</td>
</tr>
<tr>
<td>UMLS</td>
<td>Caris</td>
<td>153</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>Foundation Medicine</td>
<td>57</td>
<td>31</td>
</tr>
</tbody>
</table>
The number of raw concepts mapping from our intermediate terminologies to the OncoTree terminology seemed to be higher for our NCIt pipeline (Table 2). 191 of the 193 initial Caris DX (99.0%) were mapped to at least one concept in NCIt. Of these 191 DX, 123 (64.4%) were not mapped to any concept in OncoTree (Figure 3a). Of the initial 193 DX, 68 DX (35.2%) were mapped all the way to OncoTree. All 205 initial FM DX were mapped to a concept in NCIt, but 89 DX (43.4%) were not mapped to a concept in OncoTree. 116 DX (56.6%) of the 205 initial DX were mapped all the way to OncoTree. Moreover, there were much fewer UMLS concept mappings to OncoTree than NCIt concept mappings. For Caris, 153 DX (79.3%) of 193 initial DX were mapped to UMLS concepts (Figure 3b). Of these 153 DX, 92 DX (60.1%) were not mapped to an OncoTree concept. 61 DX (31.6%) of the initial 193 Caris DX were mapped all the way to OncoTree. For the 205 FM DX, only 57 DX (27.8%) were mapped to UMLS concepts. Of these 57 DX, 31 DX (54.4%) were not mapped to OncoTree concepts. Overall, only 26 DX (12.7%) of the 205 initial FM DX were mapped all the way to OncoTree.

The NCIt pipeline seemed to perform better than the UMLS pipeline at generating concepts for raw DX entries. On one hand, the NCIt pipeline had fewer raw DX entries returning no concepts than the UMLS pipeline. Specifically, our pipelines returned 2 unmatched concepts for NCIt versus 188 unmatched concepts for UMLS for the first segment of the pipeline and 214 versus 311 for the second stage of each pipeline. This larger number of unmapped concepts was likely due to missing mappings in the OncoTree API. On the other hand, we found statistically significant differences between each pipeline’s performance data (Table 3) hinting at the NCIt pipeline being capable of identifying over twice more concepts on average than the UMLS pipeline for both pipeline segments. NCIt identified 2.3 times more concepts per raw DX entry ($p<.0001$) compared to UMLS on the intermediate mapping segment of each pipeline. The NCIt pipeline identified 2.7 more concepts than UMLS ($p<.0001$) for the OncoTree mapping segment of each pipeline.
Table 3. Statistical Comparison of the Number of Concepts Mapped to Raw DX Entries. NCIt mapping returned significantly more concepts per raw DX entry. UMLS was the reference in both models.

<table>
<thead>
<tr>
<th>Pipeline Segment</th>
<th>Model Type</th>
<th>Rates Ratio (Exp(β)) (Ref=UMLS)</th>
<th>Estimate (β)</th>
<th>Confidence Interval (95%)</th>
<th>Std Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate Mapping</td>
<td>Poisson</td>
<td>2.3</td>
<td>0.85</td>
<td>0.72</td>
<td>0.99</td>
<td>0.068</td>
</tr>
<tr>
<td>OncoTree Mapping</td>
<td>Zero-Inflated Negative Binomial</td>
<td>2.7</td>
<td>0.98</td>
<td>0.74</td>
<td>1.22</td>
<td>0.12</td>
</tr>
</tbody>
</table>

The pathologist-driven validation of our mapping revealed that the overwhelming majority of our mappings were accurate for our selected NCIt pipeline (Table 4). Overall, 191 of 193 Caris DX were mapped to at least one concept in NCIt, and 181 DX were accurately mapped to a concept in NCIt. From NCIt, 68 of the Caris DX were mapped to at least one concept in OncoTree, and 63 of the mappings were accurate. All of the 205 FM DX were mapped to at least one concept in NCIt, and 202 DX were accurately mapped. 113 of 116 DX mapped to OncoTree from the FM NCIt concepts were accurate. Among the DX that were not accurately mapped to a concept in NCIt, a portion were mapped by NOBLE Coder to non-human concepts. For example, the DX “Nasopharynx and paranasal sinuses squamous cell carcinoma” was mapped to the NCIt concept “Squamous Cell Carcinoma of the Rat Nasopharynx,” and was not mapped to any other relevant concepts. From NCIt concepts to OncoTree, incorrect mappings were sometimes caused by an inaccurate map from the initial DX to NCIt. Of the NCIt concepts to which the DX “Pancreas neuroendocrine tumor (pNET)” was mapped, one was “Primitive Neuroectodermal Tumor,” likely due to NOBLE Coder’s interpretation of the abbreviation “pNET” alone. This concept mapped to the corresponding OncoTree concept, which did not match the initial DX. Other incorrect mappings to OncoTree were likely caused by the association of the incorrect NCIt concept code to an OncoTree concept.

Table 4. Accuracy of mapping to NCIt and OncoTree. The majority of initial DX were accurately mapped to NCIt, but there was a significant decrease in overall DX mapped to OncoTree.

<table>
<thead>
<tr>
<th>Source</th>
<th>Number of Initial DX</th>
<th>Source to NCI-T Mapping</th>
<th>Source to OncoTree Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mapped DX</td>
<td>Exact Matches</td>
</tr>
<tr>
<td>Caris</td>
<td>193</td>
<td>191 (98.96%)</td>
<td>181 (93.78%)</td>
</tr>
<tr>
<td>FM</td>
<td>205</td>
<td>205 (100.0%)</td>
<td>202 (98.54%)</td>
</tr>
</tbody>
</table>

Discussion

We were able to map free-text cancer DX from extracted from genomic report data using a data processing pipeline composed of an NLP module and a concept mapping API. NOBLE Coder outperformed UMLS search in mapping concepts to raw DX and proved to be a simple and effective method for the initial mapping to the NCIt terminology. However, many semantically irrelevant NCIt concepts were present in the output of the mappings and had to be filtered with a custom algorithm. Also, the mapping to OncoTree revealed limitations in the current state of mapping definitions, reducing the usefulness of our pipeline. Still, the produced mappings from raw DX to NCIt and OncoTree were found to be clinically accurate via clinician review.

Our study extends the existing literature by testing a method to improve data quality and reliability of textual DX data manually entered by clinicians in order to avoid ambiguity in existing DX coding systems. Because DX data reporting patterns may vary, the lack of standardization in DX records reduces their viability for secondary use. Previous approaches to this problem include deep learning and neural network models, which would require large, potentially expensive training datasets. Our pipeline uses existing clinical knowledge, which may be more cost-efficient and less time consuming than developing machine learning models. We incorporate NER and NLP methods to interpret the textual DX data and map the DX to a structured form. Our results show that it is possible to automatically map raw textual DX descriptions to support the integration and curation of clinical and genomic diagnosis data in oncology. We have successfully created a pipeline that automates DX reporting and returns standardized results for textual cancer DX. Through this pipeline, clinicians’ manually entered cancer DX data can be standardized into a format that is reliable and viable for secondary use. Further work could be done to maximize the accuracy of mappings of text form DX to the intermediate terminology, and to complete the mappings between the intermediate and OncoTree.
Further pipeline development requires testing with more sites and pathologists. Additionally, the mapping outputs of the pipeline can be used to improve mappings from the intermediate terminologies to OncoTree, which would in turn result in greater coverage and accuracy for the pipeline. Supplemental methods to maximize accuracy of mappings from textual DX data to the intermediate terminologies could also be analyzed. For example, the word “Pancreas” in the initial DX was unable to map to a concept in NCIt containing the word “Pancreatic” instead, despite the words’ semantic similarity. This could be remedied by studying semantically similar words to those which do not map accurately. Machine learning methods may be integrated into the pipeline to analyze semantic similarity between DX and concepts in a terminology\textsuperscript{35}. Additionally, methods such as fuzzy string matching\textsuperscript{36} may be used to identify patterns in the initial DX and match them with concepts in the intermediate terminology. This pipeline can also allow us to improve and further develop mappings by showing inconsistencies among the initial DX, the NCIt concept, and the final OncoTree result. The pipeline can also provide a testbed for future mappings as a framework to understand the limitations and shortcomings of existing mappings to existing DX data. The pipeline could also be adapted to serve purposes other than cancer DX mapping alone. The scope of this solution can be further broadened in order to expand automated DX mapping from other data sources, using other terminologies relevant to other clinical contexts.

Our analysis presents four core limitations related to its preliminary nature. First, the initial DX data came only from a single GENIE site and two genomic testing providers. This limits the potential generalizability of our results and the breadth of raw textual DX data used to develop and validate the pipeline. However, we successfully mapped close to 400 raw DX entries to with a very high degree of clinical accuracy. Still, we will further validate our pipeline using raw DX entries from other sites and genomic testing providers. Second, we were able to map a total of 398 DX, but we are uncertain of what proportion of all possible DX reported by genomic testing sites. We did map all available DX available at our site for the two main genomic testing providers using two different methods, which is a first step towards defining a DX mapping pipeline for data integration and reporting. Future work will include new mappings gathered from these and other genomic testing providers. Third, we had to develop an algorithm to filter irrelevant NCIt concept semantic types that was optimized for our specific case but has not been tested for generalizability. We were only able to filter out irrelevant semantic types appearing in our dataset, although there may be more potentially inaccurate semantic types when the pipeline is expanded to other cases. This will be addressed in future work. Finally, due to time and clinician resource availability constraints we were only able to validate the clinical accuracy and relevance of our textual DX to NCIt and NCIt to OncoTree mappings. These mappings were also validated by a single pathologist. However, we did find encouraging validation results for this preliminary phase of our study. We will expand our validation to multiple pathologist reviewers and multiple versions of our optimized pipelines.

Future work will be divided into two segments: further development of our mapping pipeline and further validation of the existing modules. On one hand, we will expand our development work to process data from sites other than Wake Forest Baptist Health’s Comprehensive Cancer Center registries. This will expand the number of sites, the number of genomic testing provider reports supported by our pipeline and the number of raw DX data used for development and validation. This will enable us to further refine and validate our irrelevant concept filtering algorithm to improve the performance and expand the scope of our pipeline. We will also explore other methods of entity recognition to maximize accuracy and specificity of mappings. On the other hand, we will use additional data to validate the clinical accuracy of our pipeline’s mappings but also their clinical relevance, their specificity and the repeatability of running data from multiple sites through our pipeline. We will also validate with multiple pathologists to confirm the generalizability of our accuracy results, taking into account the specificity of the DX as a measure of mapping quality. We will also assess whether there is a high level of agreement across our validations to confirm that our pipeline is approximating expert reasoning consensus rather than reflecting a single pathologist’s thought process.

**Conclusion**

Our cancer DX mapping pipeline through NCIt was shown to be highly effective and accurate when mapping from initial text-form cancer diagnoses to NCIt concepts (96.2% accuracy for our total of 398 text form DX), and adequate when mapping from those concepts to OncoTree concepts. The mapping from NCIt to OncoTree was limited in that not all OncoTree concepts were associated with a NCIt concept code, rendering our mappings incomplete. However, 176 of our 398 initial DX (44.2%) were accurately mapped all the way to OncoTree despite the NCIt to OncoTree mapping being incomplete. Having shown preliminary evidence of higher accuracy mapping effectiveness compared to the UMLS API pipeline, the NCIt pipeline would be preferable for further development and use for semi-automated oncological data integration and curation. Future work will involve increasing our textual DX data set by including new DX from the WFBMC database and updating our mappings.
Acknowledgements

This work was partially supported by the Cancer Center Support Grant from the National Cancer Institute to the Comprehensive Cancer Center of Wake Forest Baptist Medical Center (P30 CA012197), by the National Institute of General Medical Sciences’ Institutional Research and Academic Career Development Award (IRACDA) program (K12-GM102773) and by WakeHealth’s Center for Biomedical Informatics’ first Pilot Award. The authors acknowledge use of the services and facilities, funded by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (UL1TR001420) and support provided by the National Science Foundation Research Experience for Undergraduates IMPACT award (1559700).

References