Abstract

Retrospective cancer research requires identification of patients matching both categorical and temporal inclusion criteria, often based on factors exclusively available in clinical notes. Although natural language processing approaches for inferring higher-level concepts have shown promise for bringing structure to clinical texts, interpreting results is often challenging, involving the need to move between abstracted representations and constituent text elements. We discuss qualitative inquiry into user tasks and goals, data elements and models resulting in an innovative natural language processing pipeline and a visual analytics tool designed to facilitate interpretation of patient summaries and identification of cohorts for retrospective research.

Introduction

The complexities of cancer care create significant challenges for the extraction of information for retrospective research. As patients progress through diagnosis to treatment and subsequent monitoring, multiple encounters with varying specialists generate a rich set of clinical notes. For patients undergoing lengthy or multimodal (e.g., a combination of surgery, chemotherapy, and radiotherapy) treatment, hundreds or thousands of notes can be generated along the cancer journey. Review of these notes can be a laborious interpretive challenge, often involving many hours of time for medical professionals who must read through collections of notes to prepare summarized abstractions in spreadsheets or databases. This process is also brittle, as reviews conducted for one study may miss items of potential interest to subsequent studies. Although ad hoc solutions such as the oncologic history have spontaneously developed as information collection devices, they are not necessarily universal, accurate, or complete.

The Cancer Deep Phenotype Extraction (DeepPhe) project is developing informatics solutions to overcome these inefficiencies. Unlike prior work applying Natural Language Processing (NLP) techniques to individual cancer documents, DeepPhe combines details from multiple documents to form longitudinal summaries. Classic and state-of-the-art NLP techniques for extracting individual concepts are used alongside cross-document co-resolution techniques and inference rules and a rich information model, to summarize diagnoses, treatments, responses and temporal relationships as needed to support retrospective research.

Support for manual review of these results is critical. Visual and interactive tools help researchers interpret the complexities of relationships between cancers, tumors, treatments, responses, biomarkers, and other key attributes. Tools must support interpretation at multiple levels, moving from higher-level summaries to specific textual items informing those summaries. Tackling these challenges requires linking summary views with original text sources.

We draw upon a substantial body of prior work on visual cohort extraction tools, many of which have used temporal or flow metaphors to characterize temporal trends or transitions across patient populations. DeepPhe visualization tools will extend these efforts with facilities for addressing challenges associated with the ambiguities of interpreting natural language.

We have developed a preliminary tool for visualization of patient summaries and identification of cohorts grouped by high-level concepts, e.g., cancer stage. The design of this tool was motivated by insights from qualitative inquiries with potential users and informed by a multi-level information model designed to bridge the gap between individual text spans and concepts relevant to domain experts.
Methods

Qualitative Inquiry

Interviews with cancer clinical researchers informed design of DeepPhe artifacts including an information model, an NLP pipeline, and visual analytics tools. Discussions focused on challenges in cancer retrospective research, including goals, information needs, representations, bottlenecks, and challenges. Information artifacts currently in use for managing abstracted retrospective were discussed. All interviews were audio-recorded. Interview notes and recordings were analyzed to extract information needs, problems, design suggestions, and other relevant information. Results from these analyses were used to develop user personas describing potential DeepPhe users, user stories describing specific tasks, and competency questions detailing specific information requirements. Subsequent to preliminary interviews, selected informants participated in card-sorting exercises focusing on identifying information items needed by researchers in specific domains. The University of Pittsburgh Human Research Protection Office classified the qualitative inquiries as exempt (PRO13120154).

Information Model and Natural Language Processing Tool Development

Qualitative inquiry results were used to develop an information model capable of representing relevant items and attributes at multiple granularities, ranging from individual text mentions to patient summaries. Existing cTAKES pipelines were extended to extract cancer information, to use rules to infer higher-level summaries, and to store results in a Neo4j graph database (www.neo4j.com). The initial DeepPhe architecture is described in detail by Savova, et al.

Visualization

Insights from qualitative inquiries informed the development and iterative revision of a set of requirements and a corresponding series of low-fidelity prototypes for the visual analytics tools. As implementation of the prototype visualization proceeded in parallel with development of the NLP pipeline, speculative additions to the data model were constructed as needed to support visualization development for data items and attributes that were not fully supported by the NLP tool.

The resulting visualization tool was developed as a web application, using the Node.JS (www.nodejs.org) web platform to provide a middle-ware layer capable of retrieving data through the Neo4j REST interface. The visualization interface was implemented in HTML, CSS, Javascript, and the D3 visualization toolkit.

Results

Qualitative Inquiry and Visual Analytics Requirements

User challenges identified during interviews involved difficulties with information availability, access, quality, and interpretation. Although some issues were specific to the types of cancer or the context of care, most were more broadly applicable (Table 1).

Together with informant descriptions of information needs and goals, these challenges informed the creation of more than 20 user stories detailing specific tasks to be conducted for individual patients and/or at the cohort level. These user stories were broadly grouped into 12 requirement categories (Table 2).

Information Model and Natural Language Processing Tools

Qualitative interview responses led to the development of a four-level information model for detailed cancer phenotypes, described previously and summarized here:

Mentions: Text spans in source documents discussing concepts and relations of interest, including tumors, body locations, treatments, stage indicators, biomarkers, and other key elements.
<table>
<thead>
<tr>
<th>Category</th>
<th>Specific challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability</td>
<td>Missing data for out-of-network patients; Missing dates; Progression/response to therapy missing; missing data regarding key questions or procedures (smoking, use of contraceptives, etc.); Limited follow-up information; Unreported information (endometriosis in pathology reports); Difficulty in finding healthy controls</td>
</tr>
<tr>
<td>Access</td>
<td>Information spread across multiple systems</td>
</tr>
<tr>
<td>Quality</td>
<td>Inconsistencies in drug name presentation challenge interpretation; Shorthands for conjunction or alternative drugs</td>
</tr>
<tr>
<td>Interpretation</td>
<td>Reconciliation across report types; Information about medications found in both physician and nursing reports; Tumor registry data might not match records; Contradictory information; Differences in interpretation across groups (cancer registry uses TNM staging while physicians use FIGO); Verification; Extracting multiple values for trend detection; Interpreting change in patients in the context of changes in practice (advent of neo-adjuvant therapy); Interpreting outcomes of procedures.</td>
</tr>
</tbody>
</table>

**Table 1:** User information challenges identified during contextual interviews.

*Compositions:* Aggregation of all details from a given clinical note.

*Episodes:* Collection of documents into key event intervals, initially including work-up, diagnosis, medical decision making, treatment, and follow-up.

*Phenotypes:* Descriptions of cancers, tumors, treatments, and genomics, abstracted across the entire span of the patient history.

NLP tools for extraction of basic elements were implemented within the cTAKES environment\(^\text{12}\), with appropriate extensions for rules processing to support inference of higher-level summaries from text mentions. Distinct episode models were trained for breast, ovarian, and melanoma corpora. Additional details on the system architecture have been published previously\(^\text{7}\).

As development of the DeepPhe NLP tools is an ongoing effort, prototype implementation of the visual analytics tools has been facilitated by the construction of synthetic details to complete fields that cannot yet be extracted by DeepPhe. The current prototype displays extracted results for cancer stage, diagnosis, treatments, tumor size, histologic type, tumor extent, cancer cell line, body site, and biomarkers. Synthesized results for date of birth and menopausal status are also displayed.

**Visual Analytics Environment**

Initial development of the information exploration tool has focused on a subset of requirements (R1:Overview, R2: Temporal, R3: Text, R7: Stage, and R8: Genomic) at the patient level:

*Patient details* (Figure 1A) are shown in a series of panels, beginning with patient demographics and summary cancer attributes, treatments, cancer stage, cell line, and TNM values\(^\text{14}\).

*Tumor details* (1B) are shown as expandable lists of attributes colored to indicate classes of information (body site, diagnosis, biomarker values, etc.). As many patients have multiple tumors and information availability might vary across tumors, this approach provides a compact summary. A toggle above the tumor summary pane supports switching to tabular views when desired. Tumor and cancer details can be selected to reveal individual text spans contributing to the summary element.

The *Clinical note timeline* (1C) arranges notes on a timeline with one lane for each type of note (progress, radiology, and surgical pathology). Notes are color-coded according by episode. A double-thumb scroll-bar below the timeline allows zoom and panning across the extent, which spans from the interval between the
**Table 2:** System requirements and associated user stories

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Granularity</th>
<th>User Stories</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1: Overview</td>
<td>Cohort</td>
<td>Explore distribution of available data</td>
</tr>
<tr>
<td>R2: Temporal</td>
<td>Cohort</td>
<td>Note time points or intervals corresponding to changes in clinical practice or other contextual factors that might influence results; Examine trends in lab values; Review sequences of care episodes; Ask temporal queries; Determine length of overall and progression-free survival</td>
</tr>
<tr>
<td>R3: Text</td>
<td>Patient</td>
<td>Refer to original clinical text</td>
</tr>
<tr>
<td>R4: Provenance</td>
<td>Cohort</td>
<td>Identify original text used to extract/infer observations</td>
</tr>
<tr>
<td>R5: Multirecords</td>
<td>Cohort</td>
<td>Use multiple types of records to bolster interpretations</td>
</tr>
<tr>
<td>R6: QA</td>
<td>Cohort</td>
<td>Verify accuracy and relevance of extracted data; Check for consistency within documents set, and between documents and unstructured or registry data</td>
</tr>
<tr>
<td>R7: Stage</td>
<td>Cohort</td>
<td>Identify cancer stage</td>
</tr>
<tr>
<td>R8: Genomics</td>
<td>Cohort</td>
<td>Determine biomarker status and method used to make determination</td>
</tr>
<tr>
<td>R9: Treatments</td>
<td>Cohort</td>
<td>Identify treatments provided, number of cycles, doses, treatments offered but not taken, reasons for discontinuing treatments; Relate start and end dates of treatments to outcomes; Understand response to treatment/progression, including failure to respond, recurrence/no benefit, metastasis, developed resistance</td>
</tr>
<tr>
<td>R10: Filters</td>
<td>Patient</td>
<td>Filter cohorts based on categorical or temporal values; Focus patient exploration based on relevant constraints</td>
</tr>
<tr>
<td>R11: Uncertainty</td>
<td>Patient</td>
<td>Communicate both inherent ambiguity in notes and confidence in extracted information</td>
</tr>
<tr>
<td>R12: Search</td>
<td>Patient</td>
<td>Conduct searches based on clinical criteria and temporal relationships</td>
</tr>
<tr>
<td>R13: Comparison</td>
<td>Patient</td>
<td>Compare outcomes across selected cohorts</td>
</tr>
<tr>
<td>R14: Criteria</td>
<td>Patient</td>
<td>Report criteria used to identify or compare cohorts</td>
</tr>
</tbody>
</table>

first and last available documents. Episode labels above the timeline can be clicked to zoom the timeline to documents contained in the specified episode.

Below the timeline, the explanation panel (1D) bridges the gap between the inferred attributes of the cancer and tumor summaries (1A and 1B) and the text of the clinical note (1F). Selection of a summary items from the cancer or tumor summary lists leads to a display in the explanation panel describing the selected fact, along with information about its derivation from the given document.

The mention pane (1E) provides a summary mentions extracted from the selected document. Each mention can be clicked to highlight the appropriate scan in the note view (1F).

Navigation through multiple levels of abstraction is illustrated in Figure 1: The selection of tumor summary item “ER+” (1B) led to the display of the “Estrogen Receptor Status” in the explanation pane (1D), and the display of relevant mentions from Report 22 (E). Clicking on the “positive” mention leads to text confirming the ER-positive observation (1F).

The DeepPhe cohort viewer is in early stages of development. Preliminary features include selectable histograms displaying frequencies of cancer stages and biomarker status, and box plots for distribution of age at first encounter across stages, as well as distributions of stage and diagnosis information. (Figure 2).

**Discussion**

The substantial amounts of clinical text associate with cancer patient histories present significant challenges for retrospective research. With histories involving dozens of relevant notes (one dataset used in DeepPhe
averaged 30 notes/patient, after clearly irrelevant notes were removed), manual expert review will not be sufficient for the large-scale analyses needed to drive innovation. Although advances in cross-document co-reference and other techniques currently being explored by the DeepPhe project show great promise in increasing the utility of clinical text, NLP is only a first step, providing an intermediate representation not directly consumable by end-users.

Two-key challenges must be addressed to convert concepts and relationships extracted by NLP systems into artifacts useful for clinicians and researchers. Aggregation of individual observations into higher-level clinically meaningful constructs will be necessary to easily answer key research questions such as “which patients were treated with neo-adjuvant therapy?” Once such aggregations are available, visual analytics tools will be needed to help domain experts interpret the results.

The DeepPhe visualization tools represent a first step toward these goals, providing preliminary patient- and cohort- views of cancer patient data at multiple granularities. Although limited to a subset of desired data types, the current prototype provides illustration of basic functionality needed to address key requirements (Table 2) and outstanding challenges that have been identified during the evolution of the tools.

Unlike many previous text analytics tools that focus on classification or more exploratory analysis of large text corpora, the DeepPhe tools attempt to combine NLP results with an analytics interface, thus forming a complete analytics platform. DeepPhe is perhaps most similar to HARVEST, which presents observation extracted from NLP in a timeline view. However, DeepPhe’s information model and inference rules provide support for higher-level abstractions not found in HARVEST.

Expanding the utility of the clinical text for both identifying cohorts and individual patients may aid in the interpretive process. Improved displays for both rendering and interpreting inference rules linking higher-order abstractions to individual text mentions may be helpful for complex inferences, particularly when cross-document inference is involved. Techniques for linking observations across documents will also prove useful for identifying recurring concepts identified through cross-document co-reference resolution. At the cohort level, visualization of text patterns, perhaps enhanced through a Word Tree or similar visualization,
Figure 2: Preliminary cohort view: (A) Distributions of cancer stage and ages. (B) Diagnoses by patient, with mouse-over highlight of three diagnoses associated with the selected patient. (C) Biomarkers. (D) Bubble graph, sized by patient age at date of first note. (E) Patient list by age.

might help users interpret key phrases indicative of observations of interest.

DeepPhe visualization functionality will evolve alongside NLP capabilities. Although extraction and classification of individual mentions has led to promising results in many of the attributes currently shown in the prototype visualizations, much work remains to be done in the inference of higher-level aggregations, and, subsequently, the inclusion of these representations in the visualization. Two key examples involve tumors and treatments. Linking multiple tumor references across temporal extents, and including these intervals in the timeline view, will provide valuable perspective on cancer progression and response.

Enhanced temporal aggregation will also drive extensions of the DeepPhe cohort view. Incorporation of per-document episode enhancement techniques alongside orderings of treatments and time spans of specific tumors, will support temporally-aligned cohort analysis using techniques similar to those used in Outflow9, Frequence10, EventFlow9 and related systems11. Temporal9 and logical20 search facilities are also planned, with pattern search11, 21, a possibility for future work. Similar to previous tools focused on specific domains22 or care pathways and treatment plans23–25, we will use episode annotations and the semantics of the DeepPhe information model to focus designs on the specific challenges of interpreting cancer data.

Inclusion of treatment information, particularly for chemotherapeutic regimens, may provide investigators with insights into treatment histories and possible impacts. Effectively displaying treatments will require inferences not only of specific start and stop times of various drugs, but ideally of identification or inference of multi-drug protocols. Extension of DeepPhe NLP tools to identify medication regimens based on the HemOnc ontology26, 27 is a high priority.

As DeepPhe visual analytics tools evolve to include these new data elements, appropriate handling of uncertainty and missing information will become increasingly critical. NLP temporal modeling techniques28, might be used in combination with structured EHR data to eliminate some ambiguity, but many details will likely remain unspecified. Cohort and patient tools will need both appropriate display of these under-specified constraints and appropriate semantics for any related queries or filters. Temporal ambiguities also underscore the importance of tools for explicitly describing search criteria explicit and for facilitating comparisons between cohorts as techniques that might reduce the risk of misinterpretation.
Conclusion

Facilitating the extraction of understanding from complex, longitudinal patient histories is an important challenge for understanding cancer treatment and outcomes. The DeepPhe project uses a multi-faceted approach, combining both NLP, inference, information modeling, and visual analytics, to provide researchers with detailed descriptions that span the gap between key phenomena of interest and specific documentary evidence. Extension of proposed prototype designs to handle richer data, particularly involving temporal spans, will set the stage for deployment with clinical researchers and subsequent evaluation studies.

Acknowledgments

Rebecca Jacobson, Eugene Tseytlin, Girish Chavan, and Melissa Castine made important contributions to early phases of this work. DeepPhe is supported by NCI grant 1U24CA184407.

References


