Adapting a Natural Language Processing Tool to Facilitate Clinical Trial Curation for Personalized Cancer Therapy

Jia Zeng, PhD[1], Yonghui Wu, PhD[2], Ann Bailey, PhD[1], Amber Johnson, PhD[1], Vijaykumar Holla, PhD[1], Elmer V. Bernstam, MD, MSE[2], Hua Xu, PhD[2], Funda Meric-Bernstam, MD[1]
[1]Institute for Personalized Cancer Therapy, The University of Texas MD Anderson Cancer Center, Houston, TX; [2]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX

Abstract

The design of personalized cancer therapy based upon patients' molecular profile requires an enormous amount of effort to review, analyze and integrate molecular, pharmacological, clinical and patient-specific information. The vast size, rapid expansion and non-standardized formats of the relevant information sources make it difficult for oncologists to gather pertinent information that can support routine personalized treatment. In this paper, we introduce informatics tools that assist the retrieval and curation of cancer-related clinical trials involving targeted therapies. Particularly, we adapted and extended an existing natural language processing tool, and explored its applicability in facilitating our annotation efforts. The system was evaluated using a gold standard of 539 curated clinical trials, demonstrating promising performance and good generalizability (81% accuracy in predicting genotype-selected trials and an average recall of 0.85 in predicting specific selection criteria).

Introduction

It is now affordable to sequence an individual patient's genome and design personalized cancer treatment plans that directly target the underlying molecular aberrations. Personalizing therapy requires identifying the molecular alterations that "drive" cancer development in an individual patient, as well as associations between specific genomic alterations and specific targeted therapies. This information is used to optimally match patients to approved drugs and ongoing clinical trials of investigational targeted therapies. The process involves review and analysis of biomedical literature and other resources that provide information about molecular biology, targeted therapies and clinical trials. This laborious, manual process does not scale.

At MD Anderson Cancer Center (MD Anderson), the Institute for Personalized Cancer Therapy (IPCT) is dedicated to providing personalized cancer therapy to our patients. As part of our daily operation, IPCT's decision support team is manually curating biomedical literature, databases of targeted therapies and clinical trials to assist our physicians with personalized therapy selection. To expedite this curation effort, our institute is developing an informatics infrastructure that applies automated (or semi-automated) tools to achieve the following goals: 1) to retrieve and analyze molecular, pharmacological, clinical and patient-specific information from biomedical literature, targeted therapy and clinical trial databases as well as electronic health records; 2) to represent them in a standardized format and integrate them into a knowledge repository that is easy for the curators to navigate; and 3) to offer interfaces that enable the physicians to easily retrieve and visualize high quality curated information. In this paper, we report our progress in constructing a curated knowledge base of cancer-related clinical trials that involve targeted therapies, specifically regarding the identification of genotype-selected clinical trials and the genes used as their selection criteria.

To provide informatics support to facilitate this, we must properly identify gene entities inside the trial documents. However, gene name ambiguity is prevalent and causes a serious problem for the computational programs that try to extract genomic information from text [1]. Several methods have been proposed to disambiguate gene mentions in the biomedical literature [2-5]. However, very little work has been done to address this issue for clinical trial documents. Wu et al. developed a system that used natural language processing (NLP) techniques to disambiguate the status of a genetic lesion mentioned in a clinical trial document [6]. Specifically Wu's system captured four features of a gene mention: 1) the contextual words and the associated information; 2) words that have dependency relationships to the gene symbol based on the dependency parse tree from Stanford Parser [7]; 3) words expressing negation status; and 4) section headers including "title", "summary" and "eligibility criteria". Using a training set of 4332 manually annotated sentences, Wu et al. constructed a support vector machine-based classifier to identify the status of a gene mention as belonging to one of the following nine categories: 1) Drug Class (e.g., **AKT** inhibitor MK-2206); 2) Gene Status Altered or Gene Status Not Altered (e.g., any **HER2** status); 3) Gene Status Altered (e.g., with documented **BRAF** mutation); 4) Gene Status Not Altered (e.g., wild type **MET** status); 5) Gene Status Unknown (e.g., **KRAS** mutation unknown); 6) Alteration Detected or Not Detected (e.g., selecting patients' whose **HER2** status is measured); 7) Gene Status (e.g., to compare **MET** status); 8) Gene (e.g., **PIK3CA** is a gene involved in the PI3K/AKT pathway); and 9) English Word (e.g., these requirements have to be **met**). The system achieved a highest accuracy of 89.8%, demonstrating its applicability in the real-world task of clinical trial annotation.

In this study, we adapted Wu's system [6] and assessed its performance using trials curated by the MD Anderson Cancer Center IPCT decision support team as the gold standard. Our evaluation demonstrated the merit of applying the system to facilitate manual curation and also validated the generalizability of the existing NLP tool.

Methods

Our system consists of a clinical trial retrieval and preprocessing component, a gene recognition component and a gene mention disambiguation component adapted from Wu's system.

Clinical Trial Retrieval and Pre-processing – To assist IPCT's daily operation, we have developed a program that automatically retrieved and pre-processed potential targeted therapy clinical trials from Clinicaltrials.gov [8] and the MD Anderson clinical trial database [9]. Given a set of applicable targeted therapies, the program automatically expanded the drug names by including known aliases (based on NCI's drug dictionary [10]) and retrieved matching clinical trials from Clinicaltrials.gov via its RESTful API (by constructing the search term for the "Interventions" field using the list of drug names concatenated with Boolean operator OR and formulating a query URL accordingly). The criteria for a match include: 1) the trial has to be an ongoing study (recruiting, not yet recruiting or available for expanded access); 2) it has to mention the drug name/alias in either the intervention or title sections; and 3) it needs to be applicable to at least one cancer type. The program then parsed the trial records (in XML format) returned by Clinicaltrials.gov, extracted and pre-processed pertinent information, and stored them in a tabular format. The fields included in the reformatted records were: Unique Trial Identifier (NCTID), Drugs, Applicable Conditions, Broadly Categorized Conditions, Detailed Recruitment Status, Phase, Title, Inclusion Criteria, Exclusion Criteria, General Criteria, Sponsors/Collaborators, Locations, and Hyperlinked URL to the Trial Document. If a trial was conducted at MD Anderson, then additional MD Anderson-specific information including the PI's name, clinic, and MD Anderson recruitment status would be provided.

It is worth noting that the criteria for selecting or excluding patients were typically provided under the eligibility section of the trial document and our program further divided the text into subsections based upon word boundaries (Inclusion or Exclusion) so the fields of Inclusion Criteria and Exclusion Criteria could be auto-populated. When no such boundary was found, all the content under eligibility would be used to populate a field called General Criteria.

Gene Recognition Component – In the existing version of Wu's system, the primary focus was to automatically disambiguate a gene mention. To identify the gene entities, Wu and colleagues first applied a string matching technique to identify potential occurrences of gene mentions and then used domain experts' feedback (e.g., confirmation, rejection or suggestion of change) to finalize the component of gene recognition.

To evaluate the feasibility of recognizing gene mentions without human intervention, in our adapted system, we constructed a gene recognizer by modifying a component from IPCT's existing information retrieval (IR) pipeline. The IR pipeline used Lucene (a text search engine library) [11] to index any textual document repository. At query time, it could take a human gene symbol as input, automatically expand it to include the official gene name and known aliases as indicated by NCBI's Entrez gene database [12], and retrieve the matching documents. The IR pipeline could also highlight the matched terms in text. To recognize genes for the purpose of identifying genotype-selected trials and their selection criteria, we modified the IR pipeline by first indexing the inclusion, exclusion and general criteria of all the trials in the gold standard, then using a set of predefined genes as the query and labeling the matched gene mentions in the trial documents by their corresponding gene symbols. To maintain consistency with IPCT's priorities, for the predefined gene list, we used a set of 543 genes whose molecular abnormality can be detected by at least one of the four sequencing panels offered at MD Anderson: CMS46 (46 gene Ampliseq platform, Ion Torrent, LifeTechnologies, Carlsbad CA), T200 and T300 (MD Anderson in-house targeted exome sequencing research platforms), and CMS400 (409 gene Ion Proton platform, Life Technologies, Carlsbad CA).

Adaptation of the Gene Mention Disambiguator – The disambiguator reported by Wu and colleagues [6] was trained and tested using a 9-class categorization system. To make the tool applicable to IPCT's curation tasks, we made the following adaptation: if the 9-class disambiguator labeled a gene mention to be class 2, 3 or 6 and the occurrence of the gene mention was not inside the trial's exclusion criteria, then predicted the trial as genotype-selected and labeled the official symbol of the gene mention as a selection criterion.

Results

Manual Curation and Generation of the Gold Standard – MD Anderson Cancer Center IPCT decision support team routinely performs manual review of clinical trials to answer the following questions: 1) whether the trial is genotype-selected, i.e., selecting for patients who have specific molecular abnormalities (e.g., PIK3CA mutation, MET amplification); 2) for a genotype-selected trial, which genes are the selection criteria; 3) whether the trial is genotype-relevant, i.e., does the trial involve a targeted therapy that is applicable to treating patients with matching molecular profiles (e.g. targeting downstream signaling activated by a molecular alteration); and 4) for a genotype-relevant trial, alterations in which genes may be relevant. For trials that have multiple cohorts and a specific molecular alteration (e.g., HER2 amplification) only applies to one cohort, we annotated them as genotype-selected.

To facilitate this study, i.e., to identify genotype-selected trials and their gene selection criteria, we constructed a gold standard of 571 clinical trials manually annotated by the IPCT team, where 153 trials were genotype-selected and the rest (418 trials) were non-genotype-selected. Notably there was no overlap between these trials and those used to train the 9-class disambiguator. Using our gold standard as a testing set, we assessed the performance of the gene recognition component and the adapted disambiguation component respectively.

Gene Recognition Component – Of the 153 genotype-selected trials in our gold standard, our gene recognizer was able to correctly identify all genes annotated as selection criteria in 121 trials. In the remaining 32 trials, at least one gene was not properly recognized.

Adapted Gene Mention Disambiguation Component – To assess the performance of our adapted disambiguator, we excluded the 32 trials that were not properly labeled by the gene recognition component and constructed a test set from the gold standard which included 121 genotype-selected trials and 418 non-genotype-selected trials. The binary classifier predicted 193 genotype-selected trials and 346 non-genotype-selected trials, yielding an accuracy of 81%. Precision and recall were 0.55 and 0.88 respectively, yielding an F score of 0.68. With the understanding that recall was not perfect (15 trials were erroneously labeled as non-genotype-selected), we entertained the following hypothetical analysis: if our curators did not have to curate the trials that were predicted to be non-genotype-selected, they would have saved 346 minutes worth of man power (approximately 1 minute used for concluding a trial that is non-genotype-selected), which made up 83% of all the time spent on annotating non-genotype-selected trials in the gold standard.

We also evaluated the performance of the disambiguator in identifying genes that serve as selection criteria. The averages of precision, recall and F score were 0.69, 0.85 and 0.74 respectively. Overall, recall was higher than precision, which is consistent with our expectation, since for our task, a false negative (missed trial) is much worse than a false positive. Table 1 shows the performance of our system on nine genes that were annotated as the selection criteria for at least 5 clinical trials in the gold standard.

Discussion

In this paper, we presented an informatics system that facilitates the retrieval, analysis and curation of genotype-selected clinical trials to guide personalized cancer treatment. We have adapted and extended an existing NLP tool trained at a different institution and investigated its applicability to our curation tasks. Using IPCT's in-house curated clinical trials as the gold standard, we evaluated the performance of the modified system and observed promising results with an average accuracy of 81% in predicting genotype-selected trials and an average recall of 0.85 in predicting the genes that serve as selection criteria. To understand the limitations of our current system and to shed light in our future direction, we performed an error analysis of the components of gene recognition and disambiguation. In the ensuing text, we elaborate on our analysis and identify opportunities for improvement which will be explored in our future studies.

Gene Recognition Component – Our error analysis has revealed the following five reasons why the gene recognizer failed to identify all the gene mentions in 32 trials.

1) Translocation/fusion genes (e.g., EML4-ALK): the individual components of a fusion gene were properly identified yet their co-occurrence in this context was not tagged as a fusion gene. There were 7 trials that were incompletely labeled due to this. An enhancement that recognizes the pattern of translocation/fusion genes and tags them appropriately will overcome this limitation.

2) Gene mentioned outside of the eligibility criteria: in 8 trials, the selected genes were not mentioned in the eligibility criteria section, instead they were mentioned only in the title, summary or outcome description. While we still think it is reasonable to expect the trial document to be structured so critical information such as genes used as selection criteria would occur in the eligibility criteria section, we can easily expand the sections to be analyzed to include title, summary and outcome.

3) Incomplete dictionary: to enable automatic query expansion given a gene symbol, we used NCBI's Entrez gene database as a dictionary to look up the genes' common aliases and official name. However from the 6 mislabeled trials we learned that such a dictionary would require some expansion. For instance, RAS is often used to refer to a family of genes encoding proteins in the RAS family, including NRAS, KRAS and HRAS, yet it is not included as an alias for any of these three genes. Similar observations have been made between the following alias/symbol pairs: RAF for **BRAF**, MEK for **MAP2K1** or **MAP2K2**, PDGFR for **PDGFRA** or **PDGFRB,** and CD79 for **CD79A** or **CD79B**. To overcome this problem, we will supplement the current dictionary by integrating additional resources that contain information about gene names (such as GeneCard [13]), and/or design rules that extrapolate an alias based upon the gene symbol. For instance, we can assume that a gene ending with a letter (e.g., CD79A) encodes a subunit of a protein (e.g., CD79) and automatically assign the root portion of the symbol (e,g., CD79) as an alias. In our current gene recognizer, we have already applied a similar rule for gene symbols ending with a digit (e.g., FGFR1/FGFR2/FGFR3/FGFR4) that belong to the same family (e.g., FGFR) and automatically included the family name as an alias.

4) Tokenizer: a commonly adopted convention for describing a point mutation is to place a delimiter (white space or dash) between a gene symbol and the point mutation (e.g., BRAF V600E or BRAF-V600E). However, our error analysis of 3 mislabeled trials revealed that in some trial documents, the delimiter was omitted (e.g., BRAFV600E). While this may not cause a problem for a gene recognizer that applies substring matching, it does introduce an issue to more sophisticated (and probably more efficient) information retrieval strategies that use a tokenizer which relies on such delimiters to identify word boundaries. To resolve this, we will customize the default tokenizer to recognize a pattern of point mutations (e.g., V600E) as a separate word.

5) Inferred by curators: there were 8 trials where the genes that were annotated as the selection criteria were not mentioned explicitly in the trial documents but were inferred by the curators based upon their domain knowledge. For example, the following text occurred in one trial that was mislabeled: "*with tumor mutations/amplifications in one of 3 genetic pathways (DNA repair, PI3K or RAS/RAF)*". While our tool was capable of recognizing the genes whose symbol/aliases are consistent with the pathway name, it was unable to infer what genes are associated with the DNA repair pathway. To achieve a perfect recall in this category, we would have to integrate pathway information into the gene recognition component.

We understand that the task of gene recognition and normalization is very sophisticated in its own right and the aforementioned analysis was not intended to be a comprehensive assessment of this component. Due to the scope of this study, we did not construct our gold standard in a manner that supports an in-depth evaluation of the gene recognizer. For future work, we plan to address this issue as well as exploring existing tools such as those evaluated at the BioCreative gene normalization competitions [14] (e.g. GenNorm [15], GeneTUKit [16] and IASL-IISR [17]).

Adapted Gene Mention Disambiguator Component – An examination of several trials that were erroneously classified by our disambiguation component revealed the following three common causes.

1) Negation boundary: to identify negation status, Wu et al. used a training set of a localized negation/assertion lexicon constructed by the domain experts based on their review of the clinical trial training set and applied a support vector machine based method to identify the negation cases and assertion cases in the stage-2 classification. The SVMs considered a rich set of features regarding negation status including: the direction of the negation words in relation to the gene symbols, the distance between the negation cues and the target gene symbol, and the punctuations. An error analysis is as follows. In situations like the following: "*Patients with histologically/cytologically confirmed advanced solid tumors with **FGFR1** or **FGFR2** amplification or **FGFR3** mutation, for which **no** further effective standard anticancer treatment exists*", the disambiguator successfully predicted the first and second gene mentions as "Gene Status Altered" but wrongly classified the third (FGFR3) to be "Gene Status Not Altered" because it is very close to a negation cue ("no"). In the future, we may explore the application of some existing negation analyzers to overcome this issue (e.g. NegEx by Chapman et al.[18] and Bejan et al's assertion analyzer [19]).

2) Drug class identification: in some cases, the disambiguation program failed to properly recognize that a gene is mentioned in the context of a drug. For example, the following cases were mislabeled as "Gene Status Altered": "*anti-EGFR antibody (cetuximab or panitumumab)*"; "*epidermal growth factor receptor (EGFR)* *inhibitor*". Utilizing existing resources such as the UMLS that can help identify drug names could improve performance.

3) Disease acronym identification: some cancer types have acronyms that can be confused with an alias of a gene. For instance, papillary thyroid carcinoma is often abbreviated as PTC, which also happens to be an alias of the gene RET. In our current system, such ambiguity has not been taken into consideration. We plan to address this issue in our future work.

Conclusion

The existing NLP tool was generalizable. Informatics tools may partially automate the process of information gathering for the delivery of personalized cancer therapy. By conducting an error analysis, we identified several ways of further improving the performance of our system, which will be explored in our future studies.

Table 1. Performance evaluation of the adapted disambiguator from gene perspective.

Gene Symbol	# of Associated Genotype-Selected Trials in Gold Standard	Precision	Recall	F
BRAF	63	0.94	0.79	0.86
ERBB2	14	0.59	0.93	0.72
ALK	14	0.93	1.00	0.97
KRAS	11	0.65	1.00	0.79
PIK3CA	11	0.82	0.82	0.82
NRAS	8	0.88	0.88	0.88
PTEN	7	1.00	0.86	0.92
EGFR	6	0.45	0.83	0.59
MET	5	0.45	1.00	0.62

Acknowledgement

This study was supported in part by the MD Anderson Cancer Center Sheikh Khalifa Bin Zayed Al Nahyan Institute for Personalized Cancer Therapy, NCI grant 1U01CA180964 and the Center for Clinical and Translational Science (NCATS UL1 TR000371). The authors would like to thank MyCancerGenome team at Vanderbilt University who provided annotation to the training set used by the original version of the gene mention disambiguator [6].

References

1. Chen L, Liu H, Friedman C. Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics. 2005; 21(2): 248-56.
2. Schijvenaars BJ, Mons B, Weeber M, Schuemie MJ, van Mulligen EM, Wain HM, Kors JA. Thesaurus-based Disambiguation of gene symbols. BMC Bioinformatics. 2005; 6:149.
3. Xu H, Fan JW, Hripcsak G, Mendonca EA, Markatou M, Friedman C. Gene symbol disambiguation using Knowledge-based profiles. Bioinformatics. 2007; 23(8):1015-1022.

4. Farkas R. The strength of co-authorship in gene name disambiguation. BMC Bioinformatics. 2008; 9:69.
5. Stevenson M, Guo Y. Disambiguation in the biomedical domain: the role of ambiguity type. J Biomed Inform. 2010; 43(6):972-981.
6. Wu Y, Levy MA, Micheel CM, Yeh P, Tang B, Cantrell MJ, Cooreman SM, Xu H. Identifying the status of genetic lesions in cancer clinical trial documents using machine learning. BMC Genomics. 2012; 13:S21.
7. Klein D, Manning CD. Accurate unlexicalized parsing. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. 2003:423-430.
8. Clinicaltrials.gov: an NIH funded registry and results database of publicly and privately supported clinical studies of human participants conducted around the world. [http://clinicaltrials.gov/]
9. MD Anderson cancer center clinical trial registry. [http://www.mdanderson.org/patient-and-cancer-information/cancer-information/clinical-trials/clinical-trials-at-md-anderson/index.html]
10. NCI funded drug dictionary provides technical definitions and synonyms for drugs/agents used to treat patients with cancer or conditions related to cancer. [http://www.cancer.gov/drugdictionary]
11. McCandles M, Hatcher E, Gospodnetic O. Lucene in Action (2nd edition). Mannings Publications Co. 2010.
12. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2005; 33:D514-9.
13. Stelzer G, Harel A, Dalah A, Rosen N, Shmoish M, Iny Stein T, Sirota A, Madi A, Safran M and Lancet D. GeneCards: one stop site for human gene research. FISEB (ILANIT). 2008.
14. Lu Z, Kao HY, Wei CH, Huang M, Liu J, Kuo CJ, Hsu CN, Tsai RTH, Dai HJ, Okazaki N, Cho HC, Gerner M, Solt I, Agarwal S, Liu F, Vishnyakova D, Ruch P, Romacker M, Rinaldi F, Bhattacharya S, Srinivasan P, Liu H, Torii M, Matos S, Campos D, Verspoor K, Livingston KM, Wilbur WJ. The gene normalization task in BioCreative III. BMC Bioinformatics. 2011; 12(Suppl 8):S2.
15. GenNorm [http://ikmbio.csie.ncku.edu.tw/GN/]
16. Huang M, Liu J, Zhu X. GeneTUKit: a software for document-level gene normalization. Bioinformatics. 2011; 1(27):1032-3.
17. IASL-IISR Gene Mention/Normalization Tool. [http://sites.google.com/site/potinglai/downloads].
18. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34:301-10.
19. Bejan CA, Vanderwende L, Xia F, Yetisgen-Yildiz M. Assertion modeling and its role in clinical phenotype Identification. J Biomed Inform. 2013;46:68-74.