Extracting genetic alteration information for personalized cancer therapy from ClinicalTrials.gov

RECEIVED 15 September 2015

REVISED 7 December 2015

ACCEPTED 13 January 2016

PUBLISHED ONLINE FIRST 24 March 2016

Jun Xu,¹ Hee-Jin Lee,¹ Jia Zeng,² Yonghui Wu,¹ Yaoyun Zhang,¹ Liang-Chin Huang,¹ Amber Johnson,² Vijaykumar Holla,² Ann M Bailey,² Trevor Cohen,¹ Funda Meric-Bernstam,^{2,3} Elmer V Bernstam,^{1,4} Hua Xu¹

ABSTRACT

Objective: Clinical trials investigating drugs that target specific genetic alterations in tumors are important for promoting personalized cancer therapy. The goal of this project is to create a knowledge base of cancer treatment trials with annotations about genetic alterations from ClinicalTrials.gov.

Methods: We developed a semi-automatic framework that combines advanced text-processing techniques with manual review to curate genetic alteration information in cancer trials. The framework consists of a document classification system to identify cancer treatment trials from ClinicalTrials.gov and an information extraction system to extract gene and alteration pairs from the Title and Eligibility Criteria sections of clinical trials. By applying the framework to trials at ClinicalTrials.gov, we created a knowledge base of cancer treatment trials with genetic alteration annotations. We then evaluated each component of the framework against manually reviewed sets of clinical trials and generated descriptive statistics of the knowledge base.

Results and Discussion: The automated cancer treatment trial identification system achieved a high precision of 0.9944. Together with the manual review process, it identified 20 193 cancer treatment trials from ClinicalTrials.gov. The automated gene-alteration extraction system achieved a precision of 0.8300 and a recall of 0.6803. After validation by manual review, we generated a knowledge base of 2024 cancer trials that are labeled with specific genetic alteration information. Analysis of the knowledge base revealed the trend of increased use of targeted therapy for cancer, as well as top frequent gene-alteration pairs of interest. We expect this knowledge base to be a valuable resource for physicians and patients who are seeking information about personalized cancer therapy.

Keywords: personalized cancer therapy, natural language processing, clinical trial

INTRODUCTION

Personalized cancer therapy, which provides tailored treatments based on a patient's specific characteristics (eg, genetic status), has shown great promise for improving outcomes for cancer patients. With the advent of next-generation sequencing, sequencing of tumor and normal tissue has become increasingly available and thus there is increasing interest in genomically informed therapy with approved and investigational agents. Hundreds of clinical trials are investigating drugs that target specific genetic alterations in tumors. Health care providers and patients who want to participate in such trials need to search trials of targeted therapies. Unfortunately, details about genetic information in cancer trials are often embedded in narrative clinical trial documents or protocols and are not directly searchable. This study aims to unlock genetic information in cancer trials to meet an important information need related to personalized cancer therapy. We developed and evaluated a semi-automated framework that identifies cancer trials from ClinicalTrials.gov and extracts genetic alteration information from the Title and Eligibility Criteria sections of clinical trial documents. By applying the framework to all trials at ClinicalTrials.gov, we built a knowledge base that contains 20 193 cancer clinical trials (covering 10 years from 2005 to 2014), of which 2024 are labeled with specific genetic alteration information.

BACKGROUND

Cancer is the second leading cause of death in the United States. While precision medicine has the potential to impact many conditions,

oncology is a particular area of emphasis. As one example, the president's recently announced Precision Medicine Initiative allocated \$70 million of \$215 million to the National Cancer Institute "to scale up efforts to identify genomic drivers in cancer and apply that knowledge in the development of more effective approaches to cancer treatment."¹

Much effort has been devoted to developing knowledge bases to support personalized cancer therapy. For example, the Catalogue of Somatic Mutations in Cancer (<http://cancer.sanger.ac.uk/cosmic>), a database of genes involved in the development of cancers and related information, has contributed greatly to research in personalized cancer therapy. However, in order to efficiently translate research findings to clinical practice, one critical step is to summarize genetic information of the tumor that is actionable and clinically significant. Several research teams have worked in this area. Two such examples are MyCancerGenome.org (initiated by Vanderbilt University) and PersonalizedCancerTherapy.org (initiated by MD Anderson). These websites serve as personalized cancer medicine knowledge resources for physicians, patients, caregivers, and researchers. The sites' authors collect information from multiple sources and give up-to-date information on what mutations make cancers grow and related therapeutic implications, including available clinical trials. We participate in the latter initiative, led by the MD Anderson Cancer Center Sheikh Khalifa Bin Zayed Al Nahyan Institute for Personalized Cancer Therapy (IPCT). The IPCT mission is to "provide personalized cancer therapy for all of our patients and define the new standard of patient care by improving outcomes and reducing costs."² Although there are

Correspondence to Hua Xu, PhD, School of Biomedical Informatics, University of Texas Health Science Center at Houston, 7000 Fannin St, Suite 870, Houston, TX 77030, USA. Phone: 713-500-3924; E-mail: hua.xu@uth.tmc.edu For numbered affiliations see end of article.

© The Author 2016. Published by Oxford University Press on behalf of the American Medical Informatics Association. All rights reserved. For Permissions, please email: journals.permissions@oup.com

FDA-approved targeted therapies for cancer (eg, BRAF inhibitors for BRAF mutant melanoma), many more targeted therapies are currently available via clinical trials. Therefore, identifying genomically relevant clinical trials is critically important for personalized cancer therapy.

One of the biggest challenges when building knowledge bases for personalized cancer therapy such as MyCancerGenome and PersonalizedCancerTherapy.org is that much of the detailed information is embedded in narrative documents. For example, ClinicalTrials.gov, a publicly available registry hosted by the National Library of Medicine at the National Institutes of Health, provides documents in compliant EXtensible Markup Language (XML) format about clinical trials for all diseases, including cancer. The registry is the largest clinical trial database, which currently contains over 200 000 research studies conducted in more than 190 countries. Although controlled terminologies of clinical trials such as Medical Subject Headings (MeSH) are suggested for data entry, data in the XML fields are still often entered as textual strings. Data fields that are relevant to this study include the Title and Description fields, which record the study title and description, respectively; Condition, which states the indication of the treatment; and others, such as Primary Purpose. Other narrative sections such as Eligibility Criteria are also used in this study. Efforts have been made to provide partially structured information about clinical trials;^{3,4} detailed information about genetic alterations that may make a patient eligible (or ineligible) for a particular cancer trial is still available only in the narrative text. It is time consuming to manually extract such information from ClinicalTrials.gov; therefore it is important to develop informatics approaches such as information extraction to facilitate curation of genetic alteration information in clinical trials.

Many attempts have been made to curate structure information from the narrative text. An Interactive Task in the BioCreative III studied the utility and usability of text-mining tools for real-life biocuration tasks including gene normalization.⁵ Wei et al.⁶ developed a web-based assisting tool, PubTator, which shows the capability of enhancing both efficiency and accuracy of manual curation. However, automatically extracting genetic alteration information from ClinicalTrials.gov records is also challenging. First, finding trials about cancer treatment is not straightforward. Searching for cancer in the Condition field on ClinicalTrials.gov returns many trials that are not about cancer therapeutics, but mention the term cancer somewhere in the document. Besides, for a given cancer trial document, mentions of gene names may be ambiguous. For example, the gene symbol “MET” could also mean the English word “met” (eg, “Patient has met the inclusion criteria.”). Moreover, gene symbols could be mentioned as part of other biomedical entities such as drugs. For example, “EGFR” in the sentence “Patients may not have had prior EGFR tyrosine kinase inhibitors” should not be identified as a gene. Instead, the phrase “EGFR tyrosine kinase Inhibitors” should be identified as a drug class. Further, identification of gene names only is not sufficient. More specific gene alteration status such as gene mutation, deletion, or amplification needs to be determined to facilitate searches by physicians or other advanced users.

In our previous studies, we developed machine learning–based methods to detect genetic status from cancer trials by working with MyCancerGenome.org and IPCT data.^{7,8} However, our previous studies were relatively small pilot projects that focused on methodology development and evaluation for intermediate tasks such as word sense disambiguation. In this study, we developed an end-to-end system that takes ClinicalTrials.gov documents as inputs and generates annotations of genetic alterations in all cancer trials. It consists of 2 main components: (1) a new cancer treatment trial identification system,

and (2) a revised genetic alteration extraction system built on previous methods.⁷ By combining the automated approaches and manual review into a single framework, we built a knowledge base of cancer trials labeled with the relevant genetic alteration information. This knowledge base can facilitate cancer trial enrollment and clinical decisions (eg, selecting trials of targeted therapies for which a specific patient may be eligible).

METHODS

Figure 1 shows an overview of the proposed framework. Documents at ClinicalTrials.gov are the inputs of the system, and a knowledge base about cancer treatment trials with gene alteration annotations is the output of the system. The framework consists of 2 components: (1) a component that identifies cancer treatment trials from ClinicalTrials.gov, and (2) a component that extracts gene alteration information from each trial that is identified in step 1.

Cancer treatment trial identification

ClinicalTrials.gov has implemented various technologies to normalize and standardize data submitted to its system.⁹ For example, the Condition field indicates the diseases (or conditions) that the study drug is intended to treat. Despite the usefulness of such technologies, the function of searching trials that aim to investigate drugs for cancer therapy is still not ideal. False positives are observed when we search “cancer” in the Condition field using the default search engine available at ClinicalTrials.gov. Table 1 shows 2 types of such errors. As our goal is to build an accurate knowledge base of trials for cancer treatment, we developed a new system that can automatically identify cancer treatment trials with high confidence. For ambiguous results, we implemented a manual review process to ensure only trials for cancer treatment are included. Figure 2 shows the workflow of our system for collecting cancer treatment trials. It consists of 3 steps: (1) collect candidate trials from ClinicalTrials.gov, (2) score each candidate trial by integrating information from multiple sections of trials and external knowledge bases, and (3) manually review trials with lower scores. We describe each of these steps as follows:

Figure 1: Overview of the 2-step framework

Table 1: Examples of non-cancer treatment trials returned by ClinicalTrials.gov's default search engine, when searching the Condition field using the keyword "cancer."

Trial ID	Title	Condition	Purpose	Comments
NCT00900757	"Intravenous Palonosetron With Radiotherapy and Concomitant Temozolomide"	Malignant Glioma	"To determine the safety and tolerability of palonosetron in the prevention of radiation induced nausea and vomiting (RINV) in primary glioma patients receiving radiation (RT) and concomitant temozolomide (TMZ)"	The trial is not about glioma but about complications following the treatment of glioma
NCT00380406	"Protecting Ovaries and Fertility During Chemotherapy — The PROOF Trial"	Cancer, Fertility Preservation	"The purpose of this study is to determine whether gonadotropin releasing hormone agonists (medical therapy) will protect against ovarian failure in reproductive aged women undergoing sterilizing chemotherapy"	Although cancer is specified as in the Condition field, the trial is not about cancer treatment

Figure 2: The 3 steps for identifying trials about cancer treatment

Collect candidate trials

This is an initial step in fetching potential trials for cancer treatment. By reviewing the "See Conditions by Category" page at ClinicalTrials.gov, we constructed a list of 487 cancer terms, ranging from general terms such as "cancer" to specific ones such as "non-Hodgkin lymphoma." We then queried ClinicalTrials.gov, by specifying the Conditions field as one of the 487 cancer names. Moreover, we further limited returned trials to those wherein the Primary Purpose field is either "Treatment" or "Prevention," and whose Intervention field is either "Drug" or "Biological" types of substances.

Score candidate trials

For each candidate trial, we extracted information from 4 sections of the trial document, Title, Purpose, Condition, and Intervention, to determine whether the trial was about cancer treatment. MetaMap¹⁰ was used to extract disease concepts and a dictionary lookup program was used to extract drug names. PubChem¹¹ and DrugBank¹² were used to build the lexicon for the drug lookup program. We then developed a scoring system to determine the likelihood of a clinical trial being about cancer treatment, based on 2 assumptions: (1) the more cancer terms mentioned in Title, Purpose, and Condition, the more likely that it was a cancer trial, and (2) the more known cancer drugs mentioned in Intervention, the more likely that it was a cancer trial. The system calculated the ratios between cancer mentions to non-cancer disease mentions in the Title, Purpose, and Condition sections, as well as the ratio between known cancer drugs (based on drug indication knowledge bases such as MEDication Indication resource (MEDI)¹³ and noncancer drugs in the Intervention section. Then, a weighted linear sum of the 4 features was calculated to produce the final score for a trial, using empirically chosen weights. If the score was larger than the

cutoff value, which was determined empirically, we included the trial as a cancer treatment trial without further manual review.

Manually review trials with lower scores

For trials with scores lower than the cutoff value, we presented the trial document to reviewers, who manually determined whether the trial was about drugs to treat cancers. Four reviewers with medical backgrounds were recruited to review these uncertain trials.

Genetic alteration status extraction

After a trial was determined to be a cancer treatment trial, we further processed it by the second component in Figure 1, which is to extract gene alteration information from the Title and Eligibility Criteria sections. Figure 3 shows the workflow of the gene alteration annotation system, which consists of 4 steps:

Pre-processing

This component includes section detection, sentence splitting, and tokenization. Simple rules based on XML tags of trial documents at ClinicalTrials.gov were used to extract Title, Inclusion Criteria, and Exclusion Criteria sections. Regular expression-based sentence boundary detection and tokenization programs were developed to break each section into sentences and tokens.

Gene Identification

The goal of this step is to determine whether a gene name was mentioned in the Eligibility Criteria and Title sections of a trial. The gene identification problem has been extensively studied in other data sources such as biomedical literature. Many rule-based approaches¹⁴ as well as machine learning-based methods^{15,16} have been proposed and shown reasonable performance, with a focus on optimizing F-measure. For the gene identification problem in clinical trial documents here, we proposed a hybrid approach, with the goal of achieving a higher recall for following manual review. This task was further divided into 2 tasks: (1) find all possible gene names, and (2) disambiguate gene names that may refer to English words or other entities (eg, drugs), as explained in the Introduction section. For the first task, we developed a dictionary lookup program that implements a simple maximum length string-matching algorithm to find gene names based on a lexicon of gene names. The gene name lexicon was built by collecting human genes from the HUGO Gene Nomenclature Committee¹⁷ and all cancer genes in the Catalogue of Somatic Mutations in Cancer

Figure 3: The workflow of genetic alteration status extraction

database, with rich synonyms from additional resources such as the EntrezGene database.¹⁸ The comprehensive gene synonym list assures that we capture gene names with a high recall; however, it also includes many ambiguous names (eg, the gene synonym “MET” could be the English word “met”). In this project, we adopted a word sense disambiguation system developed in our previous work to determine ambiguous gene mentions in clinical trial documents.⁷ We leveraged the previous training samples and modified the existing system to classify candidate gene mentions into 3 categories: “Gene-related” (eg, PTEN gene mutation), “Drug” (eg, no prior EGFR-inhibitor therapy), and “Others” (eg, patient met the criteria).

Automated gene alteration status extraction

This component extracts the alteration status of each gene mention identified in the previous step. By working with experts at MD Anderson Cancer Center, we defined 6 different categories of gene alteration status, as shown in Table 2. A rule-based system was developed for gene alteration status detection. We used TOKENSREGEX,¹⁹ a framework for defining cascade patterns over token sequences, to develop the rules. We built up patterns over gene mentions and their surrounding context words to extract pairs of genes and associated alteration mentions. Five hundred trials annotated by the IPCT group at MD Anderson Cancer Center were used as the development set for generating rules. The system grouped gene mention(s) first by identifying the parallel structure of multi-gene mentions, then recognized the genetic alteration status regarding to the gene group. We built 10 rules on grouping gene mentions and 85 rules on determining genetic alteration status. Figure 4 illustrates the extraction of gene alteration information from a sentence with a MUTATION status. In addition, we used the NegEx algorithm²⁰ to identify negated gene-alteration pairs, eg, “with EGFR mutation negative.”

Manual Annotation of gene alteration

The automated gene alteration extraction system does not achieve 100% accuracy. To build an accurate knowledge base of cancer trials with gene alteration labels, we implemented a manual review process on the top of the automated system. We developed an annotation system that highlights the predicted gene mentions in a trial document and summarizes the predicted gene alteration status. Reviewers can read the original trial documents and decide whether to accept or reject the gene alteration status predicted by our system. In addition, they can also add new entries of gene alteration pairs if our system misses any. Figure 5 shows a screen shot of the manual review system. Six annotators with biomedicine backgrounds were recruited to perform this annotation task.

Evaluation

Cancer treatment trial identification

To evaluate the scoring system for identifying cancer treatment trials, we constructed a gold standard dataset of 1500 trials, which were randomly selected from the candidate trials and manually reviewed by a domain expert. To reduce the annotation cost, we randomly selected

Table 2: Categories of genetic alteration status defined in this study (Gene mentions are highlighted in bold)

Category	Definition	Examples
GENERAL	The trial generalizes genomic alterations (ie, mutations, amplifications/deletions, or translocations/fusions/rearrangements are not specified)	“Advanced solid tumor with diagnosed alteration in one or more of the following genes (PTEN , BRAF , KRAS , NRAS , PI3KCA , ErbB1 , ErbB2 , MET , RET , c-KIT , GNAQ , GNA11)”
WILDTYPE	Tumors that are wild type for a specific gene	“The tumor tissue must have been determined to be KRAS , NRAS , BRAF , PIK3CA wild-type by central CLIA testing”
MUTATION	Tumors with mutations in a specific gene	“Patients must have tumor harboring PTEN loss, PIK3CA mutation, and/or EGFR mutation”
AMPLIFICATION	Tumors with amplifications of a specific gene (including tumors with protein overexpression as determined by immunohistochemistry (IHC))	“Documentation of amplified PDGFRA ”
DELETION	Tumors with deletion of a specific gene (including tumors with loss of protein expression as determined by IHC)	“Patients with ATM deficient tumors”
FUSION	Tumors with fusions/translocations/rearrangements of a specific gene	“Mixed-lineage leukemia (MLL) gene rearranged Acute Lymphoblastic Leukemia” “Ph-negative CML allowed with presence of BCR-ABL rearrangement”

100 trials and asked another domain expert to double-annotate. The Kappa score between the 2 annotators is 0.914, denoting the high quality of this gold standard. We measured the precision, recall, and F-measure of the system at various cutoff values. Our goal was to identify a cutoff value that yields a high precision, as we would not review positive trials predicted by the system. For trials with scores lower than the cutoff, we recruited 4 annotators with medical backgrounds to annotate each trial as either a cancer treatment trial or not. We evaluated the inter-annotator agreement among the annotators using 200 trials by calculating the Kappa statistic. We divided the remaining low-scored trials into 4 sets and assigned each set to an annotator to produce the final set of cancer treatment trials.

Figure 4: An illustration of the rule-based genetic alteration status extraction system

Figure 5: An example of the annotation interface for predicted genetic alteration status

Trial: nct00988169.txt >>Next Trial
1148 of 1210

Status: Complete

Annotations from INCLUSION_CRITERIA:

	MUT	AMP	DEL	FUS	WT	GENERAL	NEGATED	
EGFR	☒	☐	☐	☐	☐	☐	☐	<input type="button" value="X"/>
	☐	☐	☐	☐	☐	☐	☐	<input type="button" value="+"/>

All Saved

INCLUSION_CRITERIA
Inclusion Criteria :
.....
- Presence of exon 19 or exon 21 **EGFR** activating mutation .
.....

Genetic alteration status extraction

We collected all cancer treatment trials from 2005 to 2014 and processed them through the genetic alteration status extraction system. To evaluate the performance of the Gene Identification component, we randomly selected 1600 cancer treatment trials from the set of 25 530 trials and manually annotated gene mentions in these trials into 3 categories: GENE, DRUG, and OTHER. The standard evaluation metrics including precision (P), recall (R), and F-measure (F1) were then reported in this newly created dataset. For the rule-based Genetic Alteration Status Extraction system, we created a gold standard dataset of 200 randomly selected, manually annotated trials. Precision, recall, and F-measures of the Genetic Alteration Status Extraction system were then reported in this dataset. To assess the manual annotation process, all 6 annotators were asked to annotate the same 200 trials for genetic alteration information and the inter-annotator agreement was measured using the Kappa statistic.

Descriptive analysis of the final knowledge base

Once the final knowledge base was constructed, we conducted several descriptive analyses, including the growth of gene-related cancer trials over time, the distribution of trials by genetic alteration category, and the most frequent gene-alteration pairs.

RESULTS

Cancer treatment trial identification

As of May 11, 2015, we retrieved 47 544 trials by querying ClinicalTrials.gov with the 487 cancer type names. We further filtered the retrieved trials based on their Primary Purpose and Intervention fields, which produced 29188 candidate trials. Then, we scored the candidate trials by using the proposed trial scoring system. Table 3 shows the performance of the system assessed at various cutoff scores. Although the system showed the best F-measure of 0.8662 at the cutoff of 0, we selected 4.5 as the cutoff value, as it yielded a high

Table 3: The performance of the trial scoring system for identifying cancer treatment trials at various cutoff values (The employed cutoff value and corresponding performance are highlighted in bold)

Cutoff	Precision	Recall	F-measure	No. of trials with score < cutoff
−2.5	0.7515	0.9784	0.8501	2055
−1.5	0.7795	0.9598	0.8603	2467
−0.5	0.8205	0.9098	0.8629	2894
0	0.8365	0.8980	0.8662	3020
0.5	0.8643	0.8490	0.8566	3434
1.5	0.9012	0.7784	0.8354	4205
2.5	0.9425	0.6588	0.7755	5725
3.5	0.9653	0.5176	0.6739	8130
4.5	0.9944	0.3451	0.5124	13 153
5.5	1.000	0.1157	0.2074	23 152

precision of 0.9944. There were 16 035 trials with scores higher than or equal to 4.5, and all of them were classified as cancer treatment trials without further review. The remaining 13 153 trials with scores lower than 4.5 were manually reviewed by 4 annotators. Note that by choosing 4.5 as the cutoff instead of 5.5, which yielded 1.00 precision, we were able to reduce the amount of manual annotation work to 56.8% (13 153/23 152) at the expense of 0.6% reduction in precision. The average Kappa value among the 4 annotators was 0.670. As a result of the manual review, 9495 additional trials were found to be cancer treatment trials. Together with 16 035 trials that were automatically identified, a total of 25 530 cancer treatment trials were collected in this study.

Genetic alteration status extraction

From the 1600 cancer treatment trials, 12 339 potential gene mentions were recognized and annotators identified 2089 true gene mentions and 10 250 non-gene mentions. Evaluation using this annotated dataset showed that our Gene Identification system achieved a high recall of 0.9914 and a low precision of 0.3404, which met our requirement to capture as many gene mentions as possible. We then applied the Gene Identification system to 25 530 cancer treatment trials and 15 083 trials that had at least 1 gene mention. Among them, 14 033 trials were conducted during the study period (2005 to 2014) and used for the following genetic alteration extraction. Evaluation of the rule-based genetic alteration status extraction system using the 200 manually annotated trials showed that the system achieved a precision of 0.8300 and a recall of 0.6803. In the same dataset, the average Kappa value among the 6 annotators was 0.604, indicating substantial but not perfect agreement. After manual review of the predicted genetic alteration status, 2024 cancer trials were identified with at least 1 genetic alteration mention in the eligibility criteria. The average speed for manual review was about 2 min/trial. The 6 annotators spent about 2 weeks to manually review all 14 033 trials.

Descriptive statistics of the knowledge base

The knowledge base of cancer trials with extracted genetic alteration information was released at <https://sbmi.uth.edu/ccb/resources/>. Figure 6 shows the results of some descriptive analysis of genetic

alterations in cancer trials. Figure 6A plots the number of cancer trials mentioning genetic alterations over time, and it is very clear that the number of cancer trials mentioning genetic alterations in the eligibility criteria has been increasing over the past 10 years. As shown in Figure 6B, among cancer trials mentioning genetic alterations, more than 1000 trials mentioned the MUTATION of a cancer gene. Cancer gene amplification and fusion are also the focus of many clinical studies. As shown in Figure 6C, top gene-alteration pairs in cancer trials include <HER2, AMPLIFICATION>, <EGFR, MUTATION>, and <BRAF, MUTATION>.

DISCUSSION

We developed a semi-automated framework for annotating genetic alteration status in cancer treatment trials. The framework combines automated text-processing techniques with manual review to assure high-quality annotations with minimal manual effort. Based on the tasks, the automated text-processing components could be adjusted to achieve either high recall (eg, the Gene Identification system) or high precision (eg, the scoring system for cancer treatment trials). A semi-automatic approach such as the one proposed here could be a promising alternative for other knowledge base curation tasks, as it minimizes human workload but still produces data in a precise manner. For example, the high-precision cancer treatment trial scoring system automatically identified 16 395 cancer treatment trials from 29 188 candidate trials, indicating a 56% reduction in the annotation time. Another significant contribution of this work is the knowledge base of cancer trials with genetic alteration annotations. Physicians or patients who seek treatment options based on specific genetic alterations can now browse or search our knowledge base to quickly find available trials based on specific conditions. Making such information available in a computable format is critical for enabling personalized cancer therapy.

To further improve the automated text-processing components developed here, we analyzed errors returned by the different components. For example, one of the reasons that the Gene Identification system did not achieve 100% recall was the presence of nonstandard gene names, eg, the gene name “BRAF” is a part of the gene mutation name “BRAFV600.” The low precision of the Gene Identification system was due to several issues, such as the small training size and the imbalanced classes. Nevertheless, it was still useful for filtering out about 50% of non-gene mentions. We also manually reviewed 100 errors of genetic alteration extraction. We found that the Genetic Alteration Status Extraction system did not work well with complex or unseen examples. Almost 95% false negative errors were caused by unseen cases, which the rules we developed cannot cover. For example, from the sentence “Patient must have tumor tissue tested for KRAS mutation and should be confirmed to carry a wild type,” the rule-based system missed the correct one, “<KRAS, WILDTYPE>.” The others were caused by misrecognized gene names (5%). Taking the phrase “mutations in SDHB, SDHV, or VHL genes” as an example, only “<SDHB, MUTATION>” genetic alteration was extracted. The parallel structure “SDHB, SDHV, or VHL” was not matched by our rules due to the typo “SDHV.” Among the false positive errors, around 40% were caused by the context limitation of the rule-based system, which did not consider sentence-level information. For example, “Prior receipt of vaccination against EGFRvIII” does not mean the trial focused on tumors with EGFRvIII. However, our system extracted a false positive genetic alteration “<EGFR, MUTATION>” without considering the sentence context. Sixty percent of false positive errors were caused by a genetic condition name containing a gene-alteration pair. For

Figure 6: Descriptive statistics of the genetic alteration knowledge base of cancer trials

example, the rule system extracted `<CD40, MUTATION>` from the condition name “*CD40 ligand deficiency*.”

The Kappa value for genetic alteration status annotation was 0.604, showing substantial but not perfect agreement between annotators. Several issues contributed to the discrepancy of annotations. One is related to the bias introduced by the pre-annotation system. Sometimes an annotator just went with the pre-annotation decisions without careful review. Another issue is about more difficult or rare cases. For example, to annotate the phrase “confirmed UGT1A1 TA indel genotype,” some annotators used “DELETION,” while others identified it as “MUTATION,” due to the confusing abbreviation “indel.”

This study has several limitations. First, we limited the search scope of genetic alteration information to Title and Eligibility Criteria sections (include/exclude criteria) only. However, other sections could also contain genetic alteration information and currently they are not analyzed. Second, we focused on gene mentions in the clinical trial text, thus our system is primarily optimized for identifying genotype-selected trials.²¹ However, it would not be as effective at identifying “genotype-relevant” trials. Such trials contain information of agents targeting a pathway that was altered by a genomic alteration, which is not specifically mentioned in the text. For example, BRAF-activating mutations may confer sensitivity to MEK inhibitors, but MEK inhibitor trials selecting for BRAF mutations may not be retrieved in a search for BRAF, if BRAF is also not mentioned in the clinical trials.org text.

Thus, in addition to this tool, other tools are still needed to provide scientific associations between drugs and genes/pathways. Furthermore, we identified genetic alterations at the category level instead of the more specific variant level. Therefore, our future work will include extending the proposed pipeline to other sections of trial documents and to extract more detailed variant-level information. We are also planning to integrate this framework into the workflow of IPCT at MD Anderson, in order to process MD Anderson trials and to facilitate physicians’ and patients’ information needs in the context of genomically informed trial selection.

The knowledge base that we built here covers 10 years of cancer treatment trials, from 2005 to 2014. Extensive efforts will be made continuously to keep the knowledge base as accurate and up-to-date as possible. Based on our observation, there are about 1000 cancer trials added to ClinicalTrials.gov every 6 months. So we plan to go through the same procedure to update the knowledge base every 6 months, which should be doable based on our experience. In addition to clinical trial documents, biomedical literature is another, much richer resource for gene alterations in cancer therapy. It provides more details about findings and conclusions of personalized cancer therapy research. Thus, another future direction of this study would be to build more sophisticated literature-mining tools to link more detailed evidence from literature to knowledge extracted from clinical trials.

CONCLUSION

In this study, we developed a semi-automated framework for identifying cancer treatment trials and extracting genetic alteration status information from Eligibility Criteria and Title sections of ClinicalTrials.gov. We then successfully applied this system to trials at ClinicalTrials.gov and created a knowledge base of cancer treatment trials with detailed labels of genetic alteration status. We believe this knowledge base will greatly contribute to personalized cancer therapy initiative by allowing users to efficiently identify genetic information in cancer trials.

CONTRIBUTORS

H.X., T.C., E.B., and F.M.B. conceived of the study. J.X., H.J.L., J.Z., Y.W., and H.X. were responsible for the overall design, development, and evaluation of this study. J.Z., Y.Z., A.J., V.H., and A.B. developed the annotation guidelines and provided the original datasets for this study. J.X., H.J.L., and H.X. did the bulk of the writing; T.C., F.M.B., and E.B. also contributed to writing and editing of this manuscript. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval of the manuscript for publication.

ACKNOWLEDGEMENTS

The authors would like to thank Beate Litzenburger, Nora Sanchez, and Yekaterina Khotskaya at MD Anderson Cancer Center, and Guixiao Ding, Xiao Dong, Qiang Wei, Kyle T Nguyen, and Tolulola Dawodu at UTHealth for their annotation work.

FUNDING

This study was supported in part by National Institute of General Medical Sciences (NIGMS) grant 1 R01 GM103859-01, National Cancer Institute (NCI) U01 CA180964, Sheikh Bin Zayed Al Nahyan Foundation, Cancer Prevention Research Institute of Texas (CPRIT) Precision Oncology Decision Support Core RP150535, National Center for Advancing Translational Sciences (NCATS) grant UL1 TR000371 (Center for Clinical and Translational Sciences), the Bosarge Foundation and the MD Anderson Cancer Center Support grant (NIH/NCI P30 CA016672). The first author (J.X.) is partially supported by the National Nature and Science Foundation of China (NSFC 61203378).

COMPETING INTERESTS

None.

References

1. FACT SHEET: President Obama's Precision Medicine Initiative. 2015. <https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative>. Accessed September 1, 2015.
2. Sheikh Khalifa Bin Zayed Al Nahyan Institute for Personalized Cancer Therapy: Transforming Cancer Care Through Research. <http://www.mdanderson.org/education-and-research/research-at-md-anderson/personalized>

-advanced-therapy/sheikh-khalifa-bin-zayed-al-nahyan-institute-for-personalized-cancer-therapy/index.html. Accessed September 1, 2015.

3. Geibel P, Trautwein M, Erdur H, et al. Ontology-based information extraction: identifying eligible patients for clinical trials in neurology. *J Data Semantics*. 2014;4(2):133–147.
4. Li J, Lu Z. Systematic identification of pharmacogenomics information from clinical trials. *J Biomed Informatics*. 2012;45(5):870–878.
5. Arighi CN, Roberts PM, Agarwal S, et al. BioCreative III interactive task: an overview. *BMC Bioinformatics*. 2011;12(Suppl 8):S4.
6. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res*. 2013;41(Web Server issue):W518–W522.
7. Wu Y, Levy MA, Micheel CM, et al. Identifying the status of genetic lesions in cancer clinical trial documents using machine learning. *BMC Genomics*. 2012;13(Suppl 8):S21.
8. Zeng J, Wu Y, Bailey A, et al. Adapting a natural language processing tool to facilitate clinical trial curation for personalized cancer therapy. In *AMIA Summits on Translational Science Proceedings*. 2014:126–131.
9. Gillen JE, Tse T, Ide NC, McCray AT. Design, implementation and management of a web-based data entry system for ClinicalTrials.gov. *Stud Health Technol Informatics*. 2004;1466–1470.
10. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010;17(3):229–236.
11. Bolton EE, Wang Y, Thiessen PA, Bryant SH. PubChem: integrated platform of small molecules and biological Activities. *Annual Reports in Computational Chemistry*. Washington, DC: American Chemical Society. 2008;4:217–241.
12. Wishart DS, Knox C, Guo AC, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*. 2006;34 (Database issue):D668–D672.
13. Wei WQ, Cronin RM, Xu H, Lasko TA, Bastarache L, Denny JC. Development and evaluation of an ensemble resource linking medications to their indications. *J Am Med Inform Assoc*. 2013;20(5):954–961.
14. Hanisch D, Fundel K, Mevissen H-T, Zimmer R, Fluck J. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*. 2005;6(Suppl 1):S14.
15. Lee KJ, Hwang YS, Kim S, Rim HC. Biomedical named entity recognition using two-phase model based on SVMs. *J Biomed Inform* 2004;37(6):436–447.
16. Torii M, Hu Z, Wu CH, Liu H. BioTagger-GM: a gene/protein name recognition system. *J Am Med Inform Assoc*. 2009;16(2):247–255.
17. Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA. genenames.org: the HGNC resources in 2011. *Nucleic Acids Res*. 2011;39(Database issue):D514–D519.
18. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*. 2005;33(Database issue):D54–D58.
19. Chang AX, Manning CD. TokensRegex: defining cascaded regular expressions over tokens. Stanford University Technical Report. 2014.
20. 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(5):301–310.
21. Meric-Bernstam F, Johnson A, Holla V, et al. A decision support framework for genomically informed investigational cancer therapy. *J Natl Cancer Institute*. 2015;107(7):djv098.

AUTHOR AFFILIATIONS

¹ School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA

² Institute for Personalized Cancer Therapy, University of Texas MD Anderson Cancer Center, Houston, TX, USA

³ Department of Investigational Cancer Therapeutics, University of Texas MD Anderson Cancer Center, Houston, TX, USA

⁴ Division of General Internal Medicine, Department of Internal Medicine, Medical School, University of Texas Health Science Center at Houston, Houston, TX, USA