A Software Application for Mining and Presenting Relevant Cancer Clinical Trials per Cancer Mutation

Cancer Informatics
Volume 16: 1–8

DOI: 10.1177/1176935117711940

Lisa M Gandy[1], Jordan Gumm[1], Amanda L Blackford[2], Elana J Fertig[2] and Luis A Diaz Jr[2]

[1]Department of Computer Science, Central Michigan University, Mt Pleasant, MI, USA.
[2]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA.

ABSTRACT: ClinicalTrials.org is a popular portal which physicians use to find clinical trials for their patients. However, the current setup of ClinicalTrials.org makes it difficult for oncologists to locate clinical trials for patients based on mutational status. We present CTMine, a system that mines ClinicalTrials.org for clinical trials per cancer mutation and displays the trials in a user-friendly Web application. The system currently lists clinical trials for 6 common genes (ALK, BRAF, ERBB2, EGFR, KIT, and KRAS). The current machine learning model used to identify relevant clinical trials focusing on the above gene mutations had an average 88% precision/recall. As part of this analysis, we compared human versus machine and found that oncologists were unable to reach a consensus on whether a clinical trial mined by CTMine was "relevant" per gene mutation, a finding that highlights an important topic which deems future exploration.

KEYWORDS: clinicaltrials.gov, gene-specific therapies, information retrieval, natural language processing, machine learning

RECEIVED: November 15, 2016. **ACCEPTED:** May 3, 2017.

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 1601 words, excluding any confidential comments to the academic editor.

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Lisa M Gandy, Central Michigan University, Pearce Hall 119, Mt Pleasant, MI 48858, USA. Email: gandy1l@cmich.edu

Introduction

ClinicalTrials.gov is a database of clinical trials hosted by the National Library of Medicine and the National Institutes of Health. The Food and Drug Administration Modernization Act of 1997 required the US Department of Health and Human Services (USDHHS) to create a registry of clinical trials. To meet this demand, the USDHHS created ClinicalTrials.gov.

ClinicalTrials.gov also features a results database, which provides a summary of study outcomes, including adverse events. The database is a useful portal in which patients, doctors, and clinical researchers can connect to recruit candidates for clinical trials. ClinicalTrials.gov lists studies on a global scale; 190 countries currently include studies on the Web site. The Web site is heavily used; as of May 2015, the online portal receives more than 179 million page views per month and 61 000 unique visitors per day.[1]

ClinicalTrials.gov does, however, have limitations in how it presents clinical studies to patients with cancer. First, regarding cancer trials, oncologists must search through "clutter" as the clinical trials featured on the site not only focus on cancer but also on several types of diseases. Second, with the advent of gene therapies and genome evaluation, cancers are not only being treated according to the origin of cancer but also by the specific gene mutations present. Many genes consist of several hundred mutations, and therefore, searching for all variant names in ClinicalTrials.gov is onerous and time-consuming.

Another drawback of using ClinicalTrials.gov is that its search functionality is a text search that is not case specific. This unrestricted search can be problematic as search terms (such as mutation names) often have more than 1 meaning. For example, if a physician searches the clinical trials database for the EGFR mutation, which is a common cause of lung cancer, the search results will include clinical trials for the EGFR mutation mixed in with trials for eGFR, a drug to treat kidney disease.

A final example of search confusion occurs as clinical trials often exclude patients with certain mutations. For example, a text search for the KRAS mutation will retrieve "A Phase II Study of Perioperative Chemotherapy Plus Panitumumab in Patients With Colorectal Cancer Liver Metastases"[2] This study excludes patients with the KRAS mutation by including text in the clinical trial description that says "[patients are included which] . . . have a non-mutant (wild-type) K-ras gene." Although the clinical trial specifically excludes patients with the KRAS mutation, the ClinicalTrials.org search engine still retrieves the trial.

To remedy these search problems, the authors present a software system, CTMine, which begins with a list of common gene names (in this case, 6 genes). The system then collects clinical trials per gene/mutation by searching ClinicalTrials.gov. Occasionally, a search for a particular gene or mutation will retrieve a clinical trial which does not contain the appropriate text. As an additional verification step, the system uses regular expressions to verify that gene/mutation names correctly appear in the wording of each clinical trial retrieved by ClinicalTrials.org. The authors use Catalogue

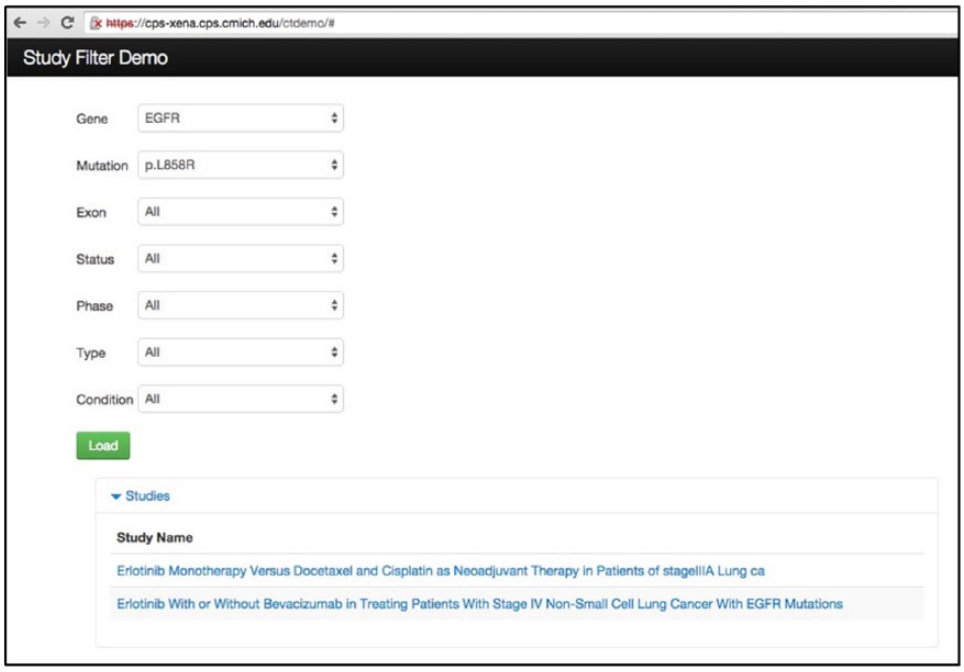

Figure 1. CTMine application user interface.

Table 1. Results for existing clinical trial classification and recommendation systems.

AUTHOR	DATA SET	ACCURACY, %	F SCORE, %
Bui et al	SMOKE, PAIN	83.0	85.7
Frenz et al	PubMed	77.0	NA
Meric-Bernstam et al	MD Anderson Clinical Trials	28.4	NA
Sarker et al	*Journal of Family Practice*	68.7	NA
Kilicoglu et al	MEDLINE	NA	66

Abbreviation: NA, Not Available.

of Somatic Mutations in Cancer (COSMIC)[3] to find all mutations associated with a gene.

To further refine the search of clinical trials by mutation, the authors create a machine learning model to rule out clinical trials which are not pertinent to a search for cancer mutation–focused clinical trials (such as double use or in exclusion circumstances documented above). The authors created a training set of clinical trials, where oncologists judged the relevance of each trial. The authors created an online interactive software to streamline the judging process.

Finally, the authors create a cancer mutation–oriented interface (Figure 1) which begins with a list of genes which the user can choose from and then allows the user to select the particular mutation related to the gene. The CTMine system only includes genes and mutations which appear in the clinical trials available to the system. The system allows oncologists to quickly search for clinical trials which focus on cancers with specific gene mutations, a function which is currently not available to the public.

Methods

Prior work

Clinical trial classification. This section focuses on systems which search or classify clinical trials using techniques drawn from natural language processing and machine learning. The accuracy of each system is given in Table 1.

Bui et al[4] built a system (regular expression discovery [RED]) which uses a top-down approach involving both natural language processing and text matching to categorize smoking status and pain status data sets. The authors use the matching text common to a set of clinical trials to create regular expressions to use in search of new trials of the same type. They developed a new classifier called "novel regular expression classifier" (RED) and created 2 text classifiers based on RED. The accuracy and F score for the system are 83.0% and 85.7%, respectively.

Frenz[5] created a Perl Regular (PREP) which allows clinicians to search PubMed for clinical trials which focus on mutations that cause deafness. The system makes use of regular

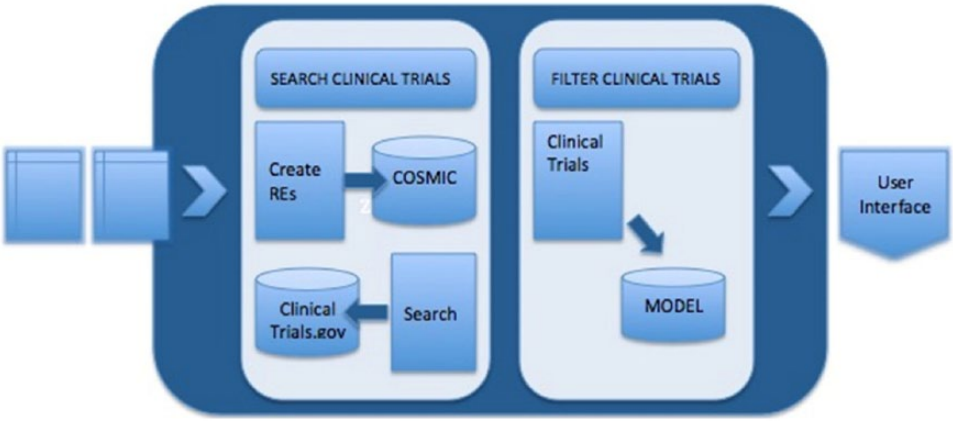

Figure 2. CTMine system diagram.

expressions. The system largely mirrors the procedures implemented by state-of-the-art information retrieval systems. Although this system uses regular expressions in their search, it does not check results for false-positive results, resulting in search terms with 2 meanings to "slip" through the system. The accuracy of the PREP system is 77%.

Meric-Bernstam et al[6] match clinical trials to gene-level alterations by a combination of natural language processing of gene names and using therapeutics known a priori to target those genes. The authors hoped to automatically match patients with certain mutations to trials at MD Anderson Medical Center which focused on the same mutation. In final results, 28.4% of patients who matched with clinical trials went on to enroll in the proposed trials.

Clinical trial recommendation strategies. There are several studies which are not directly related to text search, but the work helps inform and augment current search strategies. Sarker et al[7] explore several machine learning algorithms such as Support Vector Machines (SVMs), Bayes Net, and K-Nearest Neighbors, among others, to predict the strength of evidence in clinical trials. The authors used the Strength of Recommendation Taxonomy grading criteria as classifier values. The machine learning models reported the highest F score as 68.2%. This work could be used to inform search recommendations based on the strength of evidence.

In the same vein, Kilicoglu et al[8] create an ensemble of supervised machine learners to leverage manually annotated MEDLINE citations and then proceed to find scientifically rigorous treatment-related studies. The machine learning techniques used were Naive Bayes, SVMs, and Boosting. The classifiers were trained on 10 000 manually annotated MEDLINE citations and tested on an additional 2000 citations. The machine learning models created did better than chance (F score of 66%).

System description

The CTMine system currently collects clinical trials for the following 6 genes: ALK, BRAF, EGFR, ERBB2, KIT, and KRAS. The system includes these genes because they are highly represented among patients with cancer, resulting in their inclusion in a significant number of clinical trials. A system diagram of the CTMine system is given in Figure 2.

Table 2. Regular expressions used for gene name search in clinical trial documents.

[a-zA-Z]+[\d]+[a-zA-Z/]*[GENE MUT]+
p.[a-zA-Z]+[\d]+[a-zA-Z/]*[GENE MUT]+
[a-zA-Z]+[\d]+[a-zA-Z/]+[GENE MUT]+
p.[a-zA-Z]+[\d]+[a-zA-Z/]+[GENE MUT]+

Clinical trial collection. A Web crawler was built in the Python programming language library to extract the list of mutations returned from the COSMIC search engine API (application programming interface) per gene. Catalogue of Somatic Mutations in Cancer is an online repository of all known cancer mutations for a particular gene. The repository also provides the identifiers of studies in which researchers study certain mutations.

After COSMIC was used to extract a list of mutations and alternate IDs for each gene, a single query was constructed per mutation and passed to ClinicalTrials.gov. Each constructed query was used to retrieve a list of clinical trials from ClinicalTrials.org. The text of each clinical trial was subsequently checked to ensure that the particular search mutation existed in the wording of each retrieved trial, which was not always guaranteed. Regular expressions used for mutation search are given in Table 2. Note that regular expression 1 appears to subsume regular expression 2. However, the authors used regular expression 2 to match mutations and then retrieve the mutation number so that they could perform an additional check and ensure that the gene and mutation were in fact related.

In the next step, the system stores the text and title from each clinical trial along with a link to the relevant gene mutation in a PostgreSQL relational database. Per clinical trial, the

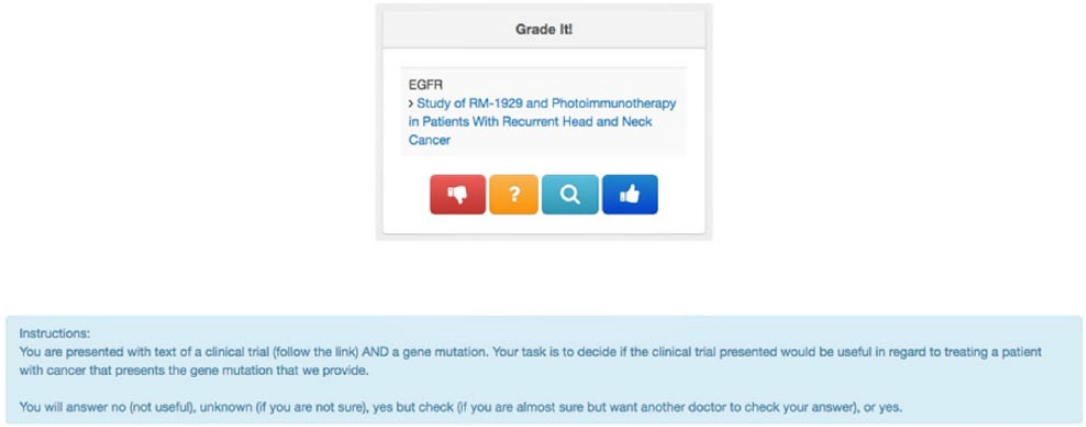

Figure 3. Clinical trial grading application.

system also stores additional metadata attributes. These include trial status, the phase of the trial, study type, and study condition.

As a working example, consider the ALK gene. The CTMine system would first collect the names of alternate IDs and mutation names related to ALK from COSMIC. The system would then generate a query for each mutation name and gene/alternate ID. One such clinical trial returned would be "Study of Oral RXDX-101 in Adult Patients With Locally Advanced or Metastatic Cancer Targeting NTRK1, NTRK2, NTRK3, ROS1, or ALK Molecular Alterations (STARTRK-1)"[9] which when using the regular expressions outlined in Table 2 which when checked does, in fact, contain the gene ALK. However, ClinicalTrials.gov also returns the clinical trial "A Long-Term Safety Study of ALKS 5461"[10] which does not reference the ALK gene but the drug ALKS 5461. The regular expressions below would not match in this case, and the system discards the clinical trial.

Clinical trial refinement methods. After collecting clinical trials, a second phase was established to exclude trials which exhibited characteristics such as the double use of gene names and exclusion of patients with certain gene mutations. To this end, the authors created a machine learning model that excluded clinical trials which were not relevant to the particular gene mutation of interest. To build the classifier, experts were needed to read each clinical trial and manually score if the trial was related to a specific gene. The authors developed a clinical trial voting application for this purpose. Four oncology residents from Johns Hopkins University to were recruited to score each trial.

Clinical trial grading online application. In total, 429 clinical trials, which reference 6 genes, were collected by the system. There was some overlap between genes as clinical trials occasionally reference more than 1 gene. A Web application was built using the Django Web framework in Python to facilitate voting. Per clinical trial, the system presents a study and a given mutation. The voter then scores if the particular study would be valuable to a patient with the given gene mutation. The voter selected from the following choices: (1) not useful, (2) unknown (if they are unsure), (3) useful but check (if they are slightly unsure but would like another oncologist to check on their answer), or (4) useful. Voters were given the following instructions via the grading Web site:

> *You are presented with text of a clinical trial (follow the link) AND a gene mutation. Your task is to decide if the clinical trial presented would be useful in regard to treating a patient with cancer that presents the gene mutation that we provide.*
>
> *You will answer no (not useful), unknown (if you are not sure), yes but check (if you are almost sure but want another doctor to check your answer), or yes.*

A screenshot of the voting application is given in Figure 3.

Results

Classification

The four oncologists were unable to vote on all 429 studies due to time constraints. A breakdown of studies voted on per gene/choice and studies voted on per grader is given in Tables 3 and 4.

Weka,[11] a commonly used machine learning package written in Java, was used to create a predictive model. The authors tested both the SVM and Naive Bayes classifiers. The features used to create the model were the text of the clinical study and the class was the vote given by most of the graders per study. All text was transformed into a feature vector (the bag-of-words model). Table 5 provides a list of Weka parameters, with a description, which were used to transform the text into the bag of words.

The authors used 3 training sets for classification. The first training set consisted of studies scored by all 4 doctors (n = 427). The second training set consisted of studies graded by 3 or more doctors (n = 402). The third training set consisted of studies graded by 2 or more doctors (n = 166). In all the training

Table 3. Number of studies listed by consensus grade/gene.

GENE/GRADE	NO	UNKNOWN	USEFUL BUT CHECK	YES	TOTAL	%
ALK	19	0	0	12	31	7
BRAF	22	0	0	35	57	13
EGFR	80	2	2	65	149	35
ERBB2	31	1	0	30	62	15
KIT	68	0	1	8	77	18
KRAS	28	0	0	24	52	12
Total	248	3	3	174	428	100

Table 4. Number of studies per voter.

VOTER ID	NO. OF VOTES
Voter A	427
Voter B	173
Voter C	407
Voter D	417

Table 5. Weka string to bag of word transformation attributes and values.

WEKA ATTRIBUTE	DESCRIPTION	VALUE
IDFTransform	Word frequency converted into inverse document frequency	True
TFTransform	Word frequency converted into term frequency—inverse document frequency when used with IDFTransform = True (see above)	True
lowerCaseTokens	All words are transformed to lower case	True
Stemmer	A collection of stemming software provided by Weka	LovinsStemmer
Stopwords	A list of stopwords provided by Weka or user defined	Weka default list
Tokenizer	A collection of tokenizers provided by Weka	WordTokenizer
useStopList	A variable to use the provided stop list (see stopwords above)	True

sets, the authors marked the class as the majority consensus with ties broken in the following order: yes, no, unknown, and yes but check. A further breakdown by percentage and number of word features is given in Table 6.

The F scores for all studies, when classified with Naive Bayes, SVMs, and Random Forests, are given in Table 7. All results were statistically significant at the $P < .05$ level when compared with chance. It is evident that as the number of studies graded increases, the F score decreases, thought the starkest contrast is between consensus data featuring 3 or more graders and 4 or more graders. The Naive Bayes model yielded the poorest results, whereas SVM results and Random Forest results are nearly identical.

Table 8 gives results for the SVM classifier on data with 2 or more voter consensus. A Pearson correlation statistic was performed to evaluate the correlation between the F score per gene and the percentage of studies per gene. The R score was not significant ($P > .05$). However, there are significant differences between F score results (note the F score of 57.1 for KIT and 57.1 for ERBB2). It is currently unclear why the SVM classifier was able to correctly predict the validity of a study in relation to a gene and warrants further investigation.

Figure 4A is a heat map of the Cohen interrater reliability between all 4 graders and all graded studies. Values on the upper triangle are the Cohen correlation coefficient, and the lower triangle reflects the number of studies reviewed by both graders. Values along the diagonal are the total number of studies reviewed by both graders. As evident in the figure, the interrater reliability between the 4 graders was low. Of the 4 graders, grader 3 had the highest average interrater reliability, thought it only ranged from 0.39 to 0.5. For further analysis, we assumed that grader 3 was the "gold standard" and calculated the interrater reliability of each grader and grader 3 for each gene (see Figure 4B). However, in this case, the interrater reliability only slightly improved (ranging from 0.48 to 0.6).

It is noteworthy that the weak consensus described above still yields high classification results by the machine learning algorithm. It is likely that the accuracy of the algorithm could improve by providing more training data. The additional training data would make it easier to define concrete features of

Table 6. Distribution of training model studies and attributes.

NO. OF CONSENSUS MODEL	NO. OF STUDIES	% OF ALL STUDIES	NO. OF ATTRIBUTES
⩾2	427	99	7861
⩾3	402	94	7460
=4	166	39	3868

Table 7. Machine learning results per number of graders.

	NAIVE BAYES			SVM			RANDOM FOREST		
	P	R	F	P	R	F	P	R	F
2	76.8	76.9	76.7	87.9	87.9	87.8	77.0	75.7	74.4
3	81.3	81.3	81.3	87.0	86.9	86.9	78.1	77.8	77.4
4	87.5	88.0	87.5	87.4	88.0	87.6	80.5	80.7	79.7

Abbreviations: P, precision; R, recall; F, F score.

Table 8. Machine learning results per gene.

GENE	PRECISION	RECALL	F SCORE	%
ALK	63.1	64.5	61.9	7
BRAF	75.1	75.4	75.2	13
EGFR	65.2	67.1	66.0	35
ERBB2	57.1	58.1	57.1	15
KIT	78.0	88.3	82.8	18
KRAS	65.3	65.4	65.2	12
				100

cancer trials that lead to an oncologist rejecting (or conversely accepting) a clinical trial.

The authors assume that the low interrater agreement would resolve with more graders. Figure 4C is a plot of the width of an exact binomial 95% confidence interval for the probability of selecting a trial as a function of the number of graders. In this case, a sharp drop occurs in the curve, and then the curve starts to level out after approximately 50 graders. This curve exists even a mutation is erroneously associated with a trial at a probability of 50%.

Discussion

As this study seeks to improve the search functionality of the ClinicalTrials.gov system in relation to mutation-specific clinical trials, it is worthwhile to compare the functionality of both systems. Table 9 gives a point-by-point comparison of both systems.

As stated previously, an important feature of the CTMine system is its method of searching by regular expression. ClinicalTrials.gov does not have this functionality, and therefore, text search is limited. Also, CTMine only returns clinical trials that directly refer to cancer studies, whereas ClinicalTrials.gov also includes trials which mention medications that treat the side effects of chemotherapy drugs, as well as clinical trials for diseases other than cancer. Regarding "double use," several mutation names (such as EGFR) are also names of particular types of other diseases. The CTMine system detects and excludes most trials with these features from its search results.

Regarding search features of both systems, status, phase, type, and condition are search options available on both platforms. However, ClinicalTrials.gov also provides search functionality regarding interventions, outcomes, sponsors, location, sex, age, funder, safety, and data, whereas CTMine does not include these search parameters.

Future Work and Conclusions

In this article, the authors discuss difficulties that occur in mutation-specific search as a result of ClinicalTrials.org Web site's limited search functionality. The authors offer a solution in the form of the CTMine system. Results demonstrate that the CTMine system assigns articles to the associated gene with high accuracy (88%). The authors also demonstrate that CTMine has a fully functional workflow, beginning with article download and ending with an interactive Web application.

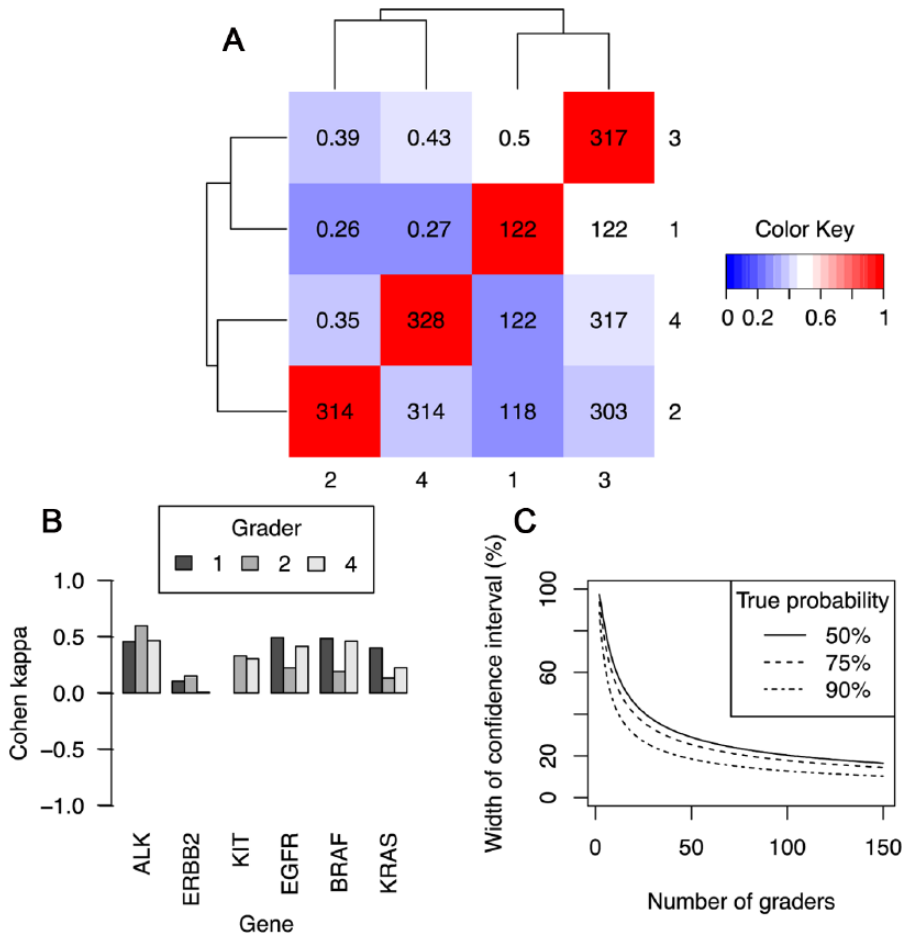

Figure 4. (A) Heatmap of Cohen correlation between all graders in all studies, colored according to the color key. Values on the upper triangle are the Cohen correlation coefficient, and lower triangle the number of studies reviewed by both graders. Values along the diagonal are a total number of studies reviewed by each grader. (B) Correspondence between graders relative to a reference grader treated as the "gold standard" for each gene. (C) Precision (width of the confidence interval) for estimated probability of selecting a trial vs number of graders.

Table 9. Comparison of CTMine and ClinicalTrials.gov.

CTMINE	CLINICALTRIALS.GOV
Regular expression search	No regular expression used in search
Clinical trials only related to cancer are mined and available to user	Not available
Clinical trials with double use of mutation name not included in results	Does not differentiate between different uses of mutation names
Results sorted by mutation name	Not available
Status, phase, type, condition	Available
Not available	Interventions, outcomes, sponsors, location, sex, age, funder, safety, date

In the future, the authors plan to incorporate more genes into the CTMine system (all known genes) and thus expand the underlying machine learning model. The authors also intend to include additional machine learning techniques, such as convolutional networks, which have shown great promise regarding text classification.

The authors find that 30 graders would optimize the uncertainty estimates of the suitability of a trial to a patient. Therefore, the authors have released the grading software used to create the data classes to the public at http://cps-xena.cps.cmich.edu/ctdemo. The authors hope in this way to increase the number of graders accessible for training data in future studies. Finally, regarding search features, the authors plan to add interventions, outcomes, sponsors, location, sex, age, funder, safety, and date so that the CTMine search functionality is as nuanced as that of ClinicalTrials.gov.

Author Contribution

Conception and Design: LAD, EJF, LMG. Data Collection: JG, LMG. Data Analysis and Interpretation: LMG, JG, EJF, ALB. Article Draft: LMG, EJF, LAD, ALB. Critical Revision of the Article: LMG, EJF.

REFERENCES

1. About ClinicalTrial.gov. https://clinicaltrials.gov/ct2/about-site/background. Accessed August 11, 2015 (Archived by WebCite® http://www.webcitation.org/6ahzRK1u4).
2. *A Phase II Study of Perioperative Chemotherapy Plus Panitumumab in Patients With Colorectal Cancer Liver Metastases*. https://clinicaltrials.gov/ct2/show/NCT01260415. Accessed August 3, 2015 (Archived by WebCite® http://www.webcitation.org/6aW12rVGR).
3. COSMIC. http://cancer.sanger.ac.uk/cosmic. Accessed August 3, 2015 (Archived by WebCite® http://www.webcitation.org/6aW1CtNMd).
4. Bui DDA, Zeng-Treitler Q. Learning regular expressions for clinical text classification. *J Am Med Inform Assn*. 2014;21:850–857.
5. Frenz CM. Deafness mutation mining using regular expression based pattern matching. *BMC Med Inform Decis Mak*. 2007;7:32.
6. Meric-Bernstam F, Brusco L, Shaw K, et al. Feasibility of large-scale genomic testing to facilitate enrollment onto genomically matched clinical trials. *J Clin Oncol*. 2015;33:2753–2762.
7. Sarker A, Mollá-Aliod D, Paris C. Towards automatic grading of evidence. In: Proceedings of LOUHI 2011 Third International Workshop on Health Document Text Mining and Information Analysis; 2011.
8. Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. *J Am Med Inform Assoc*. 2009;16:25–31.
9. *Study of Oral RXDX-101 in Adult Patients With Locally Advanced or Metastatic Cancer Targeting NTRK1, NTRK2, NTRK3, ROS1, or ALK Molecular Alterations (STARTRK-1)*. https://clinicaltrials.gov/ct2/show/NCT02097810?term=ALK&recr=Open&rank=27. Accessed August 3, 2015 (Archived by WebCite® http://www.webcitation.org/6pD7vGIdi).
10. A Long-Term Safety Study of ALKS 5461. https://clinicaltrials.gov/ct2/show/NCT02141399?term=ALK&recr=Open&rank=24. Accessed March 24, 2017 (Archived by WebCite http://www.webcitation.org/6pD77OhWO).
11. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD Explorat Newslett*. 2009;11:10–18.