Title

Representing and extracting lung cancer study metadata: study objective and study design.

Authors

Garcia-Gathright, Jean I
Oh, Andrea
Abarca, Phillip A
et al.

Publication Date
2015-03-01

DOI
10.1016/j.compbiomed.2015.01.004

Peer reviewed

Published in final edited form as:
Comput Biol Med. 2015 March 1; 0: 63–72. doi:10.1016/j.compbiomed.2015.01.004.

Representing and Extracting Lung Cancer Study Metadata: Study Objective and Study Design

Jean I. Garcia-Gathright[a,*], **Andrea Oh**[b], **Phillip A. Abarca**[c], **Mary Han**[c], **William Sago**[c], **Marshall L. Spiegel**[c], **Brian Wolf**[c], **Edward B. Garon**[c], **Alex A.T. Bui**[a,b], and **Denise R. Aberle**[a,b]
University of California, Los Angeles

Abstract

This paper describes the information retrieval step in Casama (Contextualized Semantic Maps), a project that summarizes and contextualizes current research articles on driver mutations in non-small cell lung cancer. Casama's representation of lung cancer studies aims to capture elements that will assist an end-user in retrieving studies and, importantly, judging their strength. This paper focuses on two types of study metadata: study objective and study design. 430 abstracts on EGFR and ALK mutations in lung cancer were annotated manually. Casama's support vector machine (SVM) automatically classified the abstracts by study objective with as much as 129% higher F-scores compared to PubMed's built-in filters. A second SVM classified the abstracts by epidemiological study design, suggesting strength of evidence at a more granular level than in previous work. The classification results and the top features determined by the classifiers suggest that this scheme would be generalizable to other mutations in lung cancer, as well as studies on driver mutations in other cancer domains.

Keywords

automatic summarization; quality of evidence; information retrieval

1. Introduction

The Lung Cancer Mutation Consortium, the National Cancer Institute's effort to identify and target driver mutations in lung cancer, found that driver mutations were present in 64% of lung adenocarcinomas, and that patients who were treated with targeted therapy lived longer than those who did not receive such treatment [1]. Currently, treatments approved by the Federal Drug Administration are available for cancers with epidermal growth factor receptor

[*]Corresponding author: jigarcia@ucla.edu, Phone: +1 (310) 794-8977, Fax: +1 (310) 794-3546.
[a]Department of Bioengineering 924 Westwood Boulevard, Suite 420 Los Angeles, CA 90024 USA
[b]Department of Radiological Sciences 924 Westwood Boulevard, Suite 420 Los Angeles, CA 90024 USA
[c]Department of Medicine - Division of Hematology-Oncology 924 Westwood Boulevard, Suite 200 Los Angeles, CA 90024 USA

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

(EGFR) mutations and anaplastic lymphoma kinase (ALK) gene rearrangement. As new treatments continue to be identified, it is important for clinicians to stay up-to-date on new research developments in this field.

To illustrate, a clinician may wish to answer the following questions: how likely is it that my patient has this specific mutation? What treatments are available for this mutation? Is my patient likely to respond? This project aims to assist a clinician in answering these questions as well as deeper queries concerning the strength of the claims found in published literature. For instance, were conclusions reached in a prospective clinical trial or a retrospective study? What was the study's sample size? Were the results published in a high-impact journal? Aggregated summaries of biomedical research can help inform a clinician's thinking on treatment strategies and assist in applying research findings to specific patients. Moreover, by utilizing natural language processing (NLP) techniques for automatic summarization, a model of current knowledge can be produced in a tractable fashion.

The work described here is the initial step in a larger project, Casama (Contextualized Semantic Maps), which aims to summarize and contextualize current research articles on driver mutations in cancer. Casama's representation focuses on a specific set of metadata that is geared toward the initial information retrieval task, as well as assisting the user in judging the strength of the studies retrieved. This paper describes the representation and automatic extraction of two types of metadata: study objective and study design. These efforts are demonstrated in the domain of non-small cell lung cancer (NSCLC). Casama's information retrieval performance is compared to that of PubMed. Given the domain-specific approach in which the representation is organized, the generalizability of this scheme to other domains is also investigated.

The major contribution of this work is a framework for improved information retrieval and summarization through a detailed representation of study context. This work also provides an annotated gold standard of study objective and study design as applied to driver mutations in lung cancer, as well as a first pass at automatic extraction of these data elements.

2. Background

Traditional work in information retrieval from PubMed has relied on the use of search filters, such as PubMed's own Clinical Queries, a set of Boolean filters derived by empirically discovering combinations of search terms that yield optimal sensitivity and specificity [2]. However, evaluations of such search filters have shown high specificity but low precision [3, 4]. This problematic performance results in the user having to manually filter through a significant number of irrelevant studies to meet his or her information needs. Many have achieved improved performance by using machine learning for automatic text classification in the biomedical domain [5–12]. Furthermore, more data-driven approaches that provide finer levels of granularity tailored to the domain of interest can provide a richer representation and enrich retrieval. For instance, while PubMed's Clinical Description filter searches for terms explicitly related to phenotype and prevalence, a more detailed information model that includes specific clinical and pathologic features can improve retrieval within that domain. In addition, metadata and attributes of a reported study can be used to judge the strength of evidence in an investigation. Discriminating between experimental studies, observational studies, and sub-types of observational studies can provide potentially useful information. For example, an intervention with promising results in retrospective studies (and no completed prospective studies) may point a clinician to search for open clinical trials for that treatment.

Previous work in classifying studies by strength of evidence relies on independently established standards of evidence, often reduced to two or three classes of evidence level. Aphinyanphongs et al. designated their input articles as ACP+ or ACP- depending on whether they were listed in the American College of Physicians Journal Club [13]. Kilicoglu et al. used the Clinical Hedge Database, the manually-annotated input set used to produce PubMed’s Clinical Queries filters; articles were tagged with regard to their “scientific rigor” (a binary yes/no assessment) [14–16]. Mollá and Gyawali used strength of recommendation scores (A, B, or C) as a metric of evidence [17, 18]. In the domain of neuroscience research, Landreth proposes a graphical summary of published literature in which study reproducibility and convergence are used to weight evidence [19]. In contrast, Casama aims to define objective and specific metrics, such as study design, study size, date of publication, journal impact factor, and outcome measures (e.g., overall survival, progression-free survival, quality of life) that can provide a measure of the strength of the study.

3. Methods

3.1. Representation

Casama’s representation combines top-down and bottom-up strategies to identify key classes and elements that inform clinical decisions. The top-down aspect identifies clinical information needs by means of expert opinion. For NSCLC, a thoracic oncologist (EG) and thoracic radiologist (DA) specializing in lung cancer clinical trials were both asked to identify a set of patient-oriented questions perceived as being important in a clinical study. The questions were: 1) how likely is it that my patient has this mutation; 2) is there a treatment available for this mutation; and 3) is my patient likely to respond?

The bottom-up approach subsequently employs information gathered manually from the literature to suggest ways to stratify the document collection to enable retrieval of studies that answer these questions. Four study objective classes were consequently identified – mutation characterization (relevant to question 1), mutation detection (question 1), treatment (question 2), and prognosis (question 3).

Representation of study designs was informed by a hierarchy of epidemiological study designs identified by the Oxford Centre for Evidence-Based Medicine [20]. Experimental studies provide the highest level of evidence, followed by several observational study types. In descending order of strength of evidence, the study types are: prospective cohort studies, retrospective cohort studies, case control studies, and case series. Cross-sectional studies, which are used for determining prevalence and assessing accuracy of diagnostic tests, are also included in the representation.

Figure 1 illustrates Casama's representation for lung cancer studies. This representation defines the classes Casama aims to automatically extract and visualize for the purposes of contextualized retrieval and summarization. To limit the scope of discussion, this paper focuses on the extraction of the study- objective and study design classes.

3.2. Data Collection and Annotation

The initial retrieval step took place in September 2013. All subsequent tasks (annotation, classification, and evaluation) were performed against this snapshot. PubMed was searched for "EGFR" and "lung" in the titles of articles published between January 2012 and August 2013. Restricting the search to titles ensured that the retrieved abstracts belonged to the domain of lung cancer (as opposed to a study in another cancer domain that cites previous work on lung cancer in the abstract). Excluded from the search were empty abstracts, case reports, reviews, and pre-clinical studies. 211 studies on EGFR mutation in lung cancer were retrieved via PubMed. A similar query replacing "EGFR" with "ALK" resulted in 61 articles.

Also included in the data set were abstracts from the American Society of Clinical Oncologists (ASCO) annual meetings from 2011–2013. This data source was chosen because of its high value as a source of information on current, clinically-oriented cancer research. Similar to the PubMed query the ASCO archive was searched for abstracts not containing "cell lines" whose titles contained "EGFR" or "ALK." 124 studies on EGFR and 34 studies on ALK were retrieved.

Four study objective categories were identified based on a manual investigation of the retrieved corpus and vetted by an expert in the area of lung cancer, a thoracic oncologist (EG). The categories are as follows:

1. Mutation characterization studies. These are studies that aim to discover phenotypic (e.g., clinical and pathologic) features of a driver mutation, such as age, sex, smoking status, and histology. Also belonging to this category are mutation prevalence studies and reports that aim to identify biomarkers for a driver mutation.

2. Mutation detection studies. These types of studies demonstrate a molecular analysis method for detecting driver mutations.

3. Treatment studies. This third set of studies examines the effect of a drug regimen in the treatment of lung cancer.

4. Prognostic studies. These studies associate driver mutations or clinical-pathologic features with outcomes such as survival, tumor response, or adverse events.

Abstracts were further annotated as belonging to one of the following epidemiological study designs:

1. Experimental studies. These types of studies apply an intervention to a set of patients and assess the results. Clinical trials fall into this category.

2. Cohort studies. In a cohort study, no intervention is applied by the investigator. Various cohorts (groups of patients differing by the variable in question) are

defined and compared. Observations are made at more than one time point; thus, temporal outcomes such as survival can be assessed. If possible, cohort studies are further divided into the following sub-types:

   a. Prospective cohort studies. A study is prospective if the outcome of the study is not known at the beginning of the study.

   b. Retrospective cohort studies. A retrospective study looks back on old data where the outcome has already occurred.

3. Cross-sectional studies. These type of studies make an observation of the population at a single timepoint. Prevalence studies fall into this category.

4. Case-control studies. These studies differ from cohort studies in that patients are selected based on having the outcome/event in question. These “cases” are compared to a group that did not have the outcome/event (these are the “controls”). The investigators look back in time to determine factors leading to that outcome/ event.

5. Case series. These studies are descriptive rather than analytical, and describe the experiences of a group of patients (perhaps who share a common clinico-pathologic feature or treatment history). There is no control group.

A set of annotation guidelines was developed to enable annotation by multiple readers. One physician and four non-physicians with 0.5 – 2 years of clinical lung cancer research experience (PA, MH, WS, MS, BW) annotated the document collection. The document collection was divided into five sets of 86 abstracts each. Each annotator reviewed two sets; thus, each abstract was read by two annotators. The annotators placed each abstract into one or more study objective categories, and identified the epidemiological design of the study. If the full-text of an article was available, annotators were permitted to consult the entire study to classify study objectives and study designs.

Annotation was performed iteratively. After each round of annotation, agreement was calculated by Kappa analysis. Classes with low Kappa scores were targeted for discussion. The annotators met to identify differing interpretations of the guidelines, developing strategies for unifying their interpretations by talking through difficult cases.

The annotation guidelines were updated to remove ambiguities identified during the discussion. For instance, one point of disagreement involved whether naming the percentage of patients in a study who were EGFR-positive constituted a prevalence/characterization study. After a period of discussion, the annotators agreed that a study should only be considered a prevalence study if one of its aims was to identify the rate of mutation within a population, selecting its study population carefully for this purpose. Thus, the annotation guidelines were modified to specify this distinction.

Readers then re-annotated their sets of abstracts according to the revised annotation guidelines, and the process was repeated until sufficient agreement across the collection was reached. The Kappa scores presented here were obtained after three rounds of annotation. In order to produce a gold standard, two annotators were selected to resolve discrepancy. They viewed the annotations provided by the first pair of readers, and provided a tie-breaking vote. The two annotators were selected such that no annotator performed tie-breaking on a study for which he or she was one of the original annotators.

The gold standard produced by this process is available online at: http://jigarcia.bol.ucla.edu/casama/. The counts in the gold standard for each category are summarized in Table 1.

3.3. Information Retrieval

A baseline for information retrieval performance was calculated by evaluating PubMed’s filters against the manually-annotated input set of EGFR PubMed abstracts. Filters analogous to Casama’s categories were applied to the original PubMed query, resulting in a subset of retrieved documents. For each filter, the retrieved documents were matched by PMID (PubMed identifier) to the annotated set; the number of results in each Casama category was then tabulated to calculate precision and recall. Newly added studies that were not found in the original set (i.e., studies that were added between the time of retrieval in September 2013 and the time of evaluation) were excluded. The PubMed queries examined are summarized in Tables 2, 3, and 4.

3.4. Automatic Document Classification

The document classification algorithm as developed by Joachims [21] was implemented using Python’s natural language toolkit (NLTK) and machine learning package scikit-learn [22, 23]. If full-text was available for an article, the patient-selection portion of the Methods section (determined by matching regular expressions to the section headings) was concatenated with the abstract in order to improve detection of study design. NLTK preprocessed the text by stemming and removing stop words. Unigram and bigram frequency distributions over the document collection were calculated; a binary feature vector indicating whether each unigram or bigram appeared in the text was created for each abstract. Scikit-learn then trained a set of two-class linear-kernel support vector machines (SVMs) to classify study objective; each SVM in the set corresponded to one of the study objective classes. The hyperplane constructed by each SVM was used to decide whether the document belonged in the corresponding study objective class or not.

A multi-class, one-versus-rest SVM was trained to classify documents by study design. The multiple study design classes were reduced to a set of binary SVMs; each abstract was classified according to the SVM that produced the highest output score. For study design classes with very few training examples (case-control studies, case series, sub-types of cohort studies), documents were classified by a set of hand-crafted rules, as described in Table 5.

5-fold cross validation was performed on the EGFR PubMed training set; precision and recall across folds were calculated. To test the performance of the classifier to previously unseen data, the SVMs were then trained on the entire EGFR PubMed set and tested on the ALK PubMed, EGFR ASCO, and ALK ASCO sets.

The generalizability of these classifiers was further assessed by examining the most discriminative features of the linear-kernel SVM. Features with the highest-magnitude coefficients were considered highly discriminative. Features that are not domain-specific suggest that the classifier could be used in other domains without retraining. Study design classes that were classified by rules were not included in the analysis of top features.

4. Results

4.1. Annotator Agreement

Table 6 details the inter-annotator agreement after three iterations of annotation. Cohen's Kappa agreement for study objectives over all document subsets ranged from 0.518 to 0.846, indicating moderate to substantial agreement. Standard deviations over each category ranged from 0.061 to 0.109. Detection studies had the highest Kappa agreement at 0.792, while prognostic studies had a Kappa of 0.604. Over the entire document space and all study objectives, Kappa agreement was 0.684.

For the major classes of study design (experimental, cohort, cross-sectional), Kappa agreement ranged from 0.518 to 0.860, with intraclass standard deviations ranging from 0.031 to 0.128. Experimental studies had the highest overall Kappa score (0.728) while cohort studies had the lowest (0.608). Overall, the Kappa agreement for this subset of study design classes was 0.688.

Kappa agreement for the smaller study design types (subtypes of cohort studies, case control, case series) was significantly lower, with greater deviations from the mean. Of these, retrospective studies had the best agreement, ranging from 0.352 to 0.634, indicating fair to substantial agreement. For the study design classes that had less than 0.5 Kappa agreement, the gold standard was reviewed by an informatician familiar with the representation (JG), confirming that the value in the gold standard was in agreement with the annotation guidelines.

4.2. Study Objective Classification

Table 7 presents the results of Casama's automatic classification of its four study objective categories (characterization, detection, treatment, prognosis), and compares them to PubMed's results with analogous filters. Casama outperformed PubMed in all categories based on 5-fold cross validation.

Classification of study objectives had better F-scores (balanced precision and recall) than PubMed's narrow filters (high precision, low recall) and its broad filters (high recall, low precision). As shown in Table 8, there was a decrease in performance on the test sets compared to the training set.

Receiver operating characteristic (ROC) curves for study objective classification are presented in Figure 2.

4.3. Study Design Classification

Tables 9 and 10 summarize the results of Casama's study design classifier. In Table 9, retrieval performance is compared to that of PubMed's filters (if available). Receiver operating characteristic (ROC) curves for study design classification are presented in Figure 3.

Casama outperformed PubMed in retrieval of cross-sectional studies, cohort studies, and prospective cohort studies. Casama's performance was similar to PubMed in retrieval of experimental and retrospective cohort studies. PubMed slightly outperformed Casama in retrieval of case-control studies. Rule-based classification worked best for retrospective studies; for the remaining classes, F-scores were less than 0.50. There was no degradation in performance between the training and test sets.

4.4. Representational Class Features

Tables 11 and 12 specify the top features used to discriminate between each pair of classes. Characterization studies aim to find *correlations* with mutation *status*; mutation detection studies *evaluate sensitivity* of *detection methods* in *DNA samples.* Top features for treatment studies include explicit references to treatment (*chemotherapy*, *mg* (dosage)). Prognostic studies usually explicitly mention *prognosis* and examples of outcomes such as *overall survival.*

Discriminative features for the study design classifier indicate that experimental studies describe the details of the intervention (*mg, toxicity*). Top features for the other study design classes reveal that there is a relationship between study objective and study design – cohort studies tend to overlap with prognostic studies; detection or prevalence studies tend to be cross-sectional. In both cases, this relationship is unsurprising. Cohort studies by definition include follow-up and enable assessment of outcomes, as in a prognostic study. No follow-up is required to demonstrate a mutation detection technique, so these studies are often cross-sectional.

5. Discussion

5.1. Kappa agreement

Inter-rater agreement (per the Kappa score) for study objectives was moderate to substantial. One source of disagreement between annotators stemmed from the fact that studies could have more than one objective. Indeed, 86% of studies had at least one study objective that was agreed upon by both annotators; thus, primary objectives were "easy" to annotate whereas it was more difficult to determine secondary aims of a study. Also, some study objective classes differed from each other in subtle ways, such as characterization and prognostic studies, which both aim to characterize various aspects of a mutation. This subtle difference is reflected in the lower Kappa score for prognostic studies.

Kappa scores for study design were moderate to substantial for the main study design classes. Experimental studies and cross-sectional studies had better Kappa agreement within this set of classes, as these are clearly associated with certain study types (clinical trials and detection studies, respectively) and therefore were easier to agree upon. More granular study design types were more difficult to annotate. In particular, the difference between retrospective and prospective study designs was not often communicated clearly in abstracts. Annotators had varying levels of confidence in annotating cohort studies as prospective rather than unknown, whereas retrospective studies often stated their study design explicitly. These observations are reflected in both the Kappa scores and the classification results.

5.2. Document classification

Casama's automatic classification performance was comparable to or better than PubMed's retrieval in every category. Notably, Casama automatically classified experimental studies with similar F-score compared to PubMed's manual tagging of clinical trials.

For study objective classification, a decrease in performance was observed between the training set and the test sets. The ALK PubMed test set had the smallest decrease in performance, and the decrease was greatest in the "treatment" category. A manual review of the incorrectly classified abstracts revealed that many errors could be attributed to differing stages of research between EGFR and ALK (e.g., ALK treatment studies were missed because they were descriptive rather than analytical).

In contrast, the ASCO test sets had a more dramatic drop in performance compared to the training set. In this case, a major was source of error was the difference in vocabulary between PubMed and ASCO. Due to character limits (rather than word count limits), ASCO abstracts use more abbreviations than PubMed (such as "pts" for "patients", or "C" for "chemotherapy"), contributing to error because such abbreviations are not found in the training set. These effects could be mitigated with efforts toward vocabulary standardization and abbreviation replacement via lookup tables and regular expressions. Another solution would be to train an SVM on the EGFR ASCO set. The EGFR ASCO set does not contain enough data to perform 5-fold cross validation, but we were able to train on the entire EGFR ASCO set and test on the ALK ASCO set. As shown in Table 13, classification performance was improved, indicating that performance is indeed sensitive to vocabulary differences between PubMed and ASCO.

For study design classification, performance was preserved between training and test sets. This is a very promising finding, as it suggests that the automatic extraction of study designs is a viable and generalizable strategy. However, rule-based performance was generally poor. Part of this stems from the effect of few examples of prospective cohort studies, case-control studies, and case series in the data set – small *n* results in a large penalty for missed abstracts. The other contributing factor is the fact that most studies do not explicitly name their study design in the abstract. Semantic modeling of study design, including identification of exposures, outcomes, and direction of inquiry for improved study design classification is a possible avenue for future work.

5.3. Top features

An examination of the top features reveals some interesting characteristics of the vocabulary used across studies. Many of these features would be expected (e.g., *chemotherapy* for treatment studies), and some are even included in PubMed's filters (*DNA* for detection studies). The top features also reveal less obvious terms that can be used to discriminate between studies (e.g., *receive* for experimental studies vs. *observe* for cohort studies). However, simply entering a few top features into a PubMed search query is unlikely to produce good retrieval results as the vocabulary is modeled in a high-dimensional feature space via an SVM, going beyond the basic Boolean querying available in PubMed. Indeed, issuing the baseline query to PubMed with the top term for treatment studies (*progression*) results in an F-score of 0.54. AND-ing the two most discriminative terms (*progression*, *advanced*) results in decreased recall; OR-ing them results in decreased precision.

Given the domain-specific nature of this representation, it is important to assess if the classifiers developed here can be applied outside the target domain (i.e., EGFR mutations in lung cancer). Markedly, many of the top features for the study objective classifier are not specific to EGFR mutation. As such, this classifier may be applicable to other driver mutations in NSCLC, especially those with similar treatment strategies. Furthermore, the top features of the study design classifier are not domain dependent and may generalize well to other disease and cancer domains.

5.4. Future work

This classification scheme provides a promising foundation for an automatic summarization system, facilitating the retrieval of studies in the Casama framework. Consider Semantic MEDLINE, a relational framework for automatic summarization [24]. Semantic MEDLINE automatically extracts predications (such as erlotinib TREATS NSCLC) from PubMed search results. These relations are visualized as a graph of interconnected nodes and filtered based on a set of constraints (Figure 4a). Casama aims to build from this foundation, providing more specific filters and weighting metrics to enhance visualization and concept navigation (Figure 4b).

Other future work includes improvements to classification performance, either by retrieving and annotating additional data (especially for sparsely represented study types) or through modifications to the SVM kernel as well as exploration of other classification algorithms such as naïve Bayes and decision trees. Due to their ability to handle high-dimensional feature spaces such as natural language, SVMs are often used in “textbook” examples of text classification [21, 22, 25]; however, the Casama representation is not specific to SVMs and new classification methods can be substituted easily.

Further steps for Casama include: extraction of study metadata such as endpoints, cohort size, and p-values; extraction of cohort attributes for the matching of studies to individual patients; enhanced relation extraction to include relations not covered by Semantic MEDLINE; and dynamic visualization of contextualized semantic networks.

6. Conclusion

In this study, the representation and extraction of study objective and study design in abstracts on EGFR and ALK mutation in lung cancer was explored. A manually-annotated gold standard was produced by multiple expert readers. Good retrieval performance was achieved on the training and test sets compared to PubMed. Study objective classification was sensitive to differences in vocabulary between corpora; however, study design classification was robust to these differences. Based on an examination of top features, both classifiers could generalize outside the lung cancer domain. This study represents a first step in representing and extracting study metadata for contextualized summarization of lung cancer research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NLM T15-LM007356, NIH/NLM R01-LM009961, NIH K23CA149079, and the UCLA Department of Radiological Sciences.

References

1. Kris, Mark G.; Johnson, Bruce E.; Berry, Lynne D.; Kwiatkowski, David J.; John Iafrate, A.; Wistuba, Ignacio I.; Varella-Garcia, Marileila; Franklin, Wilbur A.; Aronson, Samuel L.; Su, Pei-Fang; Shyr, Yu; Ross Camidge, D.; Sequist, Lecia V.; Glisson, Bonnie S.; Khuri, Fadlo R.; Garon, Edward B.; Pao, William; Rudin, Charles; Schiller, Joan; Haura, Eric B.; Socinski, Mark; Shirai, Keisuke; Chen, Heidi; Giaccone, Giuseppe; Ladanyi, Marc; Kugler, Kelly; Minna, John D.; Bunn, Paul A. Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. JAMA. May; 2014 311(19):1998–2006. [PubMed: 24846037]
2. Brian Haynes R, Ann McKibbon K, Wilczynski Nancy L, Walter Stephen D, Werre Stephen R, Team Hedges. Optimal search strategies for retrieving scientifically strong studies of treatment from medline: analytical survey. BMJ (Clinical research ed). May.2005 330(7501):1179.
3. Bachmann, Lucas M.; Coray, Reto; Estermann, Pius; ter Riet, Gerben. Identifying diagnostic studies in MEDLINE: Reducing the number needed to read. Journal of the American Medical Informatics Association : JAMIA. 2002; 9(6):653–658. [PubMed: 12386115]
4. McKibbon, Kathleen Ann; Wilczynski, Nancy Lou; Haynes, Robert Brian; Team, Hedges. Retrieving randomized controlled trials from medline: a comparison of 38 published search filters. Health Information and Libraries Journal. Sep; 2009 26(3):187–202. [PubMed: 19712211]
5. Yu, Wei; Clyne, Melinda; Dolan, Siobhan M.; Yesupriya, Ajay; Wulf, Anja; Liu, Tiebin; Khoury, Muin J.; Gwinn, Marta. GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. BMC bioinformatics. 2008; 9:205. [PubMed: 18430222]
6. Polavarapu, Nalini; Navathe, Shamkant B.; Ramnarayanan, Ramprasad; Haque, Abrarul; Sahay, Saurav; Liu, Ying. Investigation into biomedical literature classification using support vector machines. Proceedings/IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference; 2005. p. 366-374.
7. Donaldson, Ian; Martin, Joel; de Bruijn, Berry; Wolting, Cheryl; Lay, Vicki; Tuekam, Brigitte; Zhang, Shudong; Baskin, Berivan; Bader, Gary D.; Michalickova, Katerina; Pawson, Tony; Hogue, Christopher WV. PreBIND and textomy – mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics. Mar 11.2003 4(1)
8. Yetisgen-Yildiz, Meliha; Pratt, Wanda. The effect of feature representation on MEDLINE document classification. AMIA Annual Symposium Proceedings; 2005; 2005. p. 849-853.
9. Kim, Seunghee; Choi, Jinwook. An SVM-based high-quality article classifier for systematic reviews. Journal of Biomedical Informatics. Feb.2014 47:153–159. [PubMed: 24177318]
10. Wallace, Byron C.; Trikalinos, Thomas A.; Lau, Joseph; Brodley, Carla; Schmid, Christopher H. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics. Jan.2010 11(1):55. [PubMed: 20102628]

11. Chen, David; Müller, Hans-Michael; Sternberg, Paul W. Automatic document classification of biological literature. BMC Bioinformatics. Aug.2006 7(1):370. [PubMed: 16893465]
12. Yin, Lanlan; Xu, Guixian; Torii, Manabu; Niu, Zhendong; Maisog, Jose M.; Wu, Cathy; Hu, Zhangzhi; Liu, Hongfang. Document classification for mining host pathogen protein-protein interactions. Artificial Intelligence in Medicine. Jul; 2010 49(3):155–160. [PubMed: 20472411]
13. Aphinyanaphongs, Yindalon; Tsamardinos, Ioannis; Statnikov, Alexander; Hardin, Douglas; Aliferis, Constantin F. Text categorization models for high-quality article retrieval in internal medicine. Journal of the American Medical Informatics Association: JAMIA. Apr; 2005 12(2): 207–216. [PubMed: 15561789]
14. Kilicoglu, Halil; Demner-Fushman, Dina; Rindflesch, Thomas C.; Wilczynski, Nancy L.; Brian Haynes, R. Towards automatic recognition of scientifically rigorous clinical research evidence. Journal of the American Medical Informatics Association: JAMIA. Feb; 2009 16(1):25–31. [PubMed: 18952929]
15. Fiszman, Marcelo; Bray, Bruce E.; Shin, Dongwook; Kilicoglu, Halil; Bennett, Glen C.; Bodenreider, Olivier; Rindflesch, Thomas C. Combining relevance assignment with quality of the evidence to support guideline development. Studies in Health Technology and Informatics. 2010; 160(Pt 1):709–713. [PubMed: 20841778]
16. Choi, Sungbin; Ryu, Borim; Yoo, Sooyoung; Choi, Jinwook. Combining relevancy and methodological quality into a single ranking for evidence-based medicine. Inf Sci. Dec.2012 214:76–90.
17. Mollá, Diego; Santiago-Martínez, María Elena. Creation of a corpus for evidence based medicine summarisation. The Australasian Medical Journal. Sep; 2012 5(9):503–506. [PubMed: 23115585]
18. Gyawali, Binod; Solorio, Thamar; Benajiba, Yassine. Grading the quality of medical evidence. Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, BioNLP ’12; Stroudsburg, PA, USA. 2012; Association for Computational Linguistics; p. 176-184.
19. Landreth, Anthony; Silva, Alcino J. The need for research maps to navigate published work and inform experiment planning. Neuron. Jul; 2013 79(3):411–415. [PubMed: 23931992]
20. OCEBM Levels of Evidence Working Group et al. The oxford 2011 levels of evidence. 2011
21. Joachims, Thorsten. Learning to Classify Text Using Support Vector Machines. 2002. Springer; Boston: Apr. 2002
22. Bird, Steven; Klein, Ewan; Loper, Edward. Natural Language Processing with Python. 1. O’Reilly Media, Beijing: Cambridge Mass; Jul. 2009
23. Pedregosa, Fabian; Varoquaux, Gaël; Gramfort, Alexandre; Michel, Vincent; Thirion, Bertrand; Grisel, Olivier; Blondel, Mathieu; Prettenhofer, Peter; Weiss, Ron; Dubourg, Vincent; Vanderplas, Jake; Passos, Alexandre; Cournapeau, David; Brucher, Matthieu; Perrot, Matthieu; Duchesnay, Édouard. Scikit-learn: Machine learning in python. J Mach Learn Res. Nov.2011 12:2825–2830.
24. Rindflesch, Thomas C.; Kilicoglu, Halil; Fiszman, Marcelo; Rosemblat, Graciela; Shin, Dongwook. Semantic MEDLINE: An advanced information management application for biomedicine. Inf Serv Use. Jan; 2011 31(1–2):15–21.
25. Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich. Introduction to Information Retrieval. Cambridge University Press; New York, NY, USA: 2008.
26. Nelson, Valerie; Ziehr, Jacqueline; Agulnik, Mark; Johnson, Melissa. Afatinib: emerging next-generation tyrosine kinase inhibitor for NSCLC. OncoTargets and Therapy. 2013; 6:135–143. [PubMed: 23493883]
27. Kirkpatrick, Peter; Graham, Joanne; Muhsin, Mohamed. Cetuximab. Nature Reviews Drug Discovery. Jul; 2004 3(7):549–550.

Summary

Aggregated summaries of biomedical research can help inform a clinician’s thinking on treatment strategies and assist in applying research findings to specific patients. The work described here is the initial step in Casama (Contextualized Semantic Maps), a clinical decision support system which aims to summarize and contextualize current research articles on driver mutations in cancer. Casama’s representation focuses on a set of metadata that is geared toward the initial information retrieval task, as well as assisting the user in judging the strength of the studies retrieved. This paper describes the representation and automatic extraction of two types of metadata: study objective and study design.

Four types of study objectives were identified: mutation characterization, mutation detection, treatment, and prognosis. Study design classes, informed by principles of epidemiology, include: experimental, cohort (prospective or retrospective), and cross-sectional.

Five expert readers annotated a document set of 430 abstracts on EGFR and ALK mutations in lung cancer from PubMed and the American Society of Clinical Oncologists (ASCO). Kappa scores were moderate to substantial for the major study objective and study design classes.

Automatic classification of abstracts was performed with a support vector machine (SVM) classifier and compared to retrieval with PubMed. The SVM classified study objectives with substantially better F-scores compared to PubMed. Classification of study designs was better than or comparable to PubMed. Study objective classification was sensitive to differences in vocabulary across corpora, but study design classification was robust to these differences.

Based on an examination of top features, both classifiers could generalize outside the lung cancer domain. This study represents a first step in representing and extracting study metadata for contextualized summarization of lung cancer research.

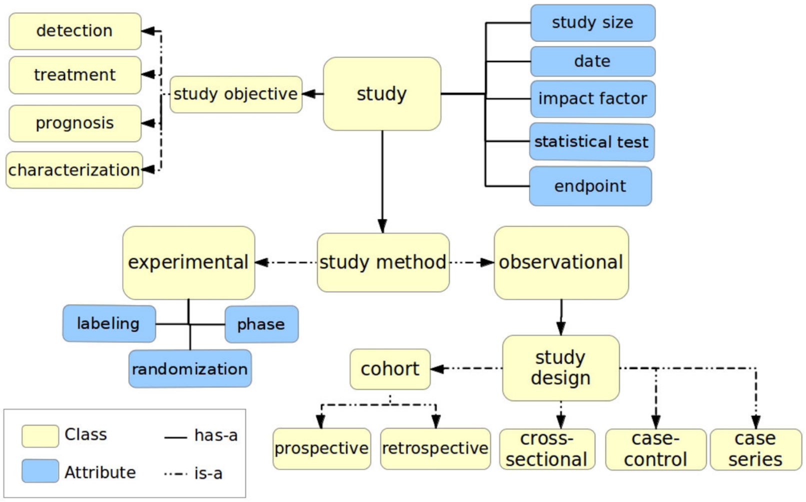
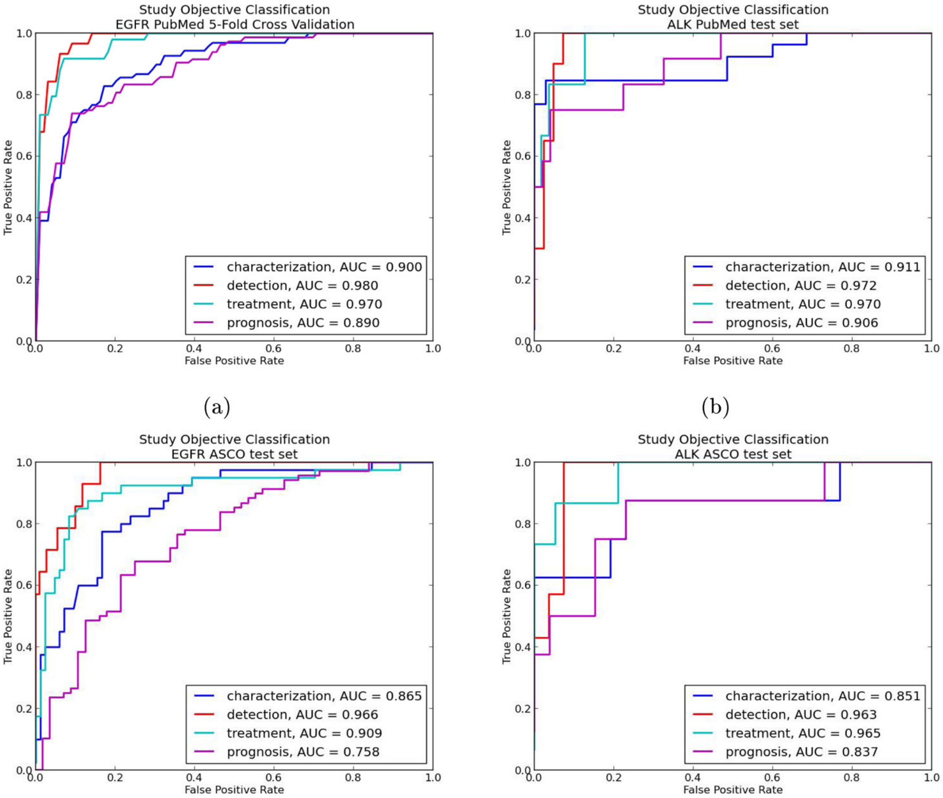


Figure 1.

Casama’s representation of lung cancer studies.



(a)

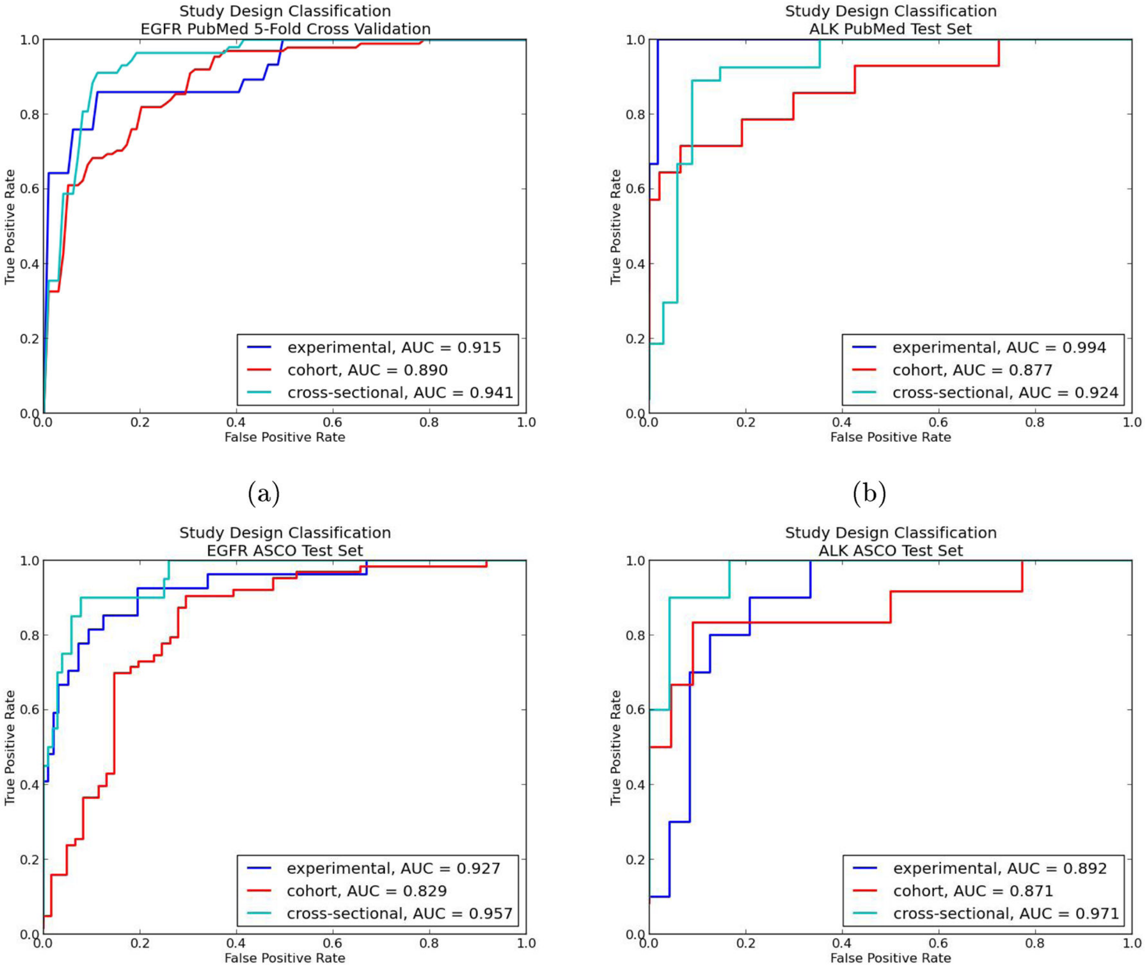
(b)

(c)

(d)

Figure 2.

Receiver operating characteristic and area under the curve for study objective classification on (a) EGFR PubMed, (b) ALK PubMed, (c) EGFR ASCO, (d) ALK ASCO.



(a)

(b)

(c)

(d)

Figure 3.

Receiver operating characteristic and area under the curve for study design classification on (a) EGFR PubMed, (b) ALK PubMed, (c) EGFR ASCO, (d) ALK ASCO.

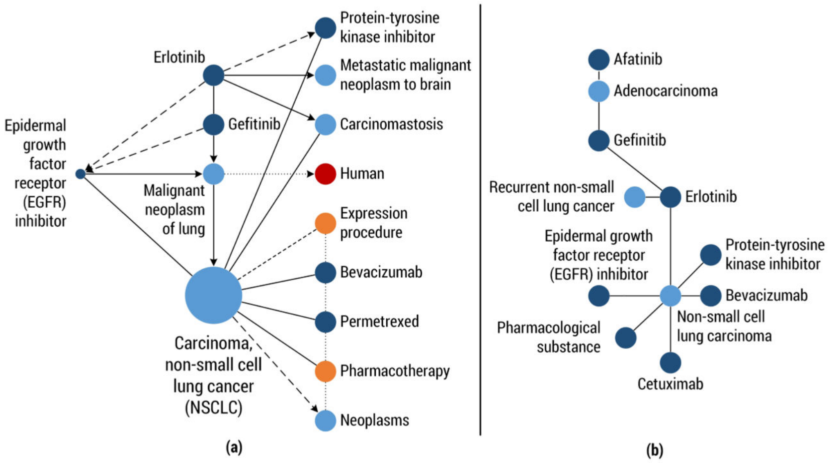

Protein-tyrosine kinase inhibitor
Erlotinib
Metastatic malignant neoplasm to brain
Gefitinib
Carcinomastosis
Epidermal growth factor receptor (EGFR) inhibitor
Human
Malignant neoplasm of lung
Expression procedure
Bevacizumab
Permetrexed
Carcinoma, non-small cell lung cancer (NSCLC)
Pharmacotherapy
Neoplasms
(a)
Afatinib
Adenocarcinoma
Gefinitib
Recurrent non-small cell lung cancer
Erlotinib
Epidermal growth factor receptor (EGFR) inhibitor
Protein-tyrosine kinase inhibitor
Bevacizumab
Non-small cell lung carcinoma
Pharmacological substance
Cetuximab
(b)

Figure 4.

This figure demonstrates the value added by Casama to the Semantic MEDLINE framework in answering the question, “What treatments are available for this mutation?” Figure 4a is Semantic MEDLINE’s visualization of treatments for EGFR-positive NSCLC (node color represents semantic type in the UMLS Semantic Network; edge style represents relation type). One way Semantic MEDLINE reduces the total number of nodes is by salience, including only nodes that appear frequently in the input set. As such, it identifies the treatments erlotinib, gefitinib, pemetrexed, and bevacizumab. In contrast, Figure 4b shows a preliminary Casama visualization with constraints on the number of experimental and prospective cohort studies (*n*≥1). Casama identifies treatment nodes like Semantic MEDLINE (erlotinib, gefitinib, bevacizumab); but pemetrexed is omitted as in this data set, relations with this drug were only found in retrospective studies. Notably, a new node in the graph is afatinib, a relatively new targeted therapy [26]. Because Afatinib has fewer associated studies, this potentially useful knowledge has been removed by Semantic MEDLINE’s salience filter. Cetuximab is a drug approved for colorectal cancer, although this data set includes a study on cetuximab for lung [27]. Such insights could be useful for a clinician seeking information on off-label treatments for lung cancer.

Table 1

Gold standard document counts for training and test sets.

Category	**EGFR PubMed**	**ALK PubMed**	**EGFR ASCO**	**ALK ASCO**
Characterization	74	26	40	8
Detection	35	20	14	7
Treatment	38	5	40	15
Prognosis	81	12	68	8
Experimental	20	3	27	10
Cohort (all)	89	14	63	12
Prospective cohort	7	1	1	0
Retrospective cohort	47	1	35	8
Unknown	35	12	27	4
Cross-sectional	60	27	20	10
Case-control	3	0	0	0
Case series	5	5	4	0

Table 2

Baseline PubMed queries for retrieving abstracts on EGFR mutation in lung cancer.

Original query	egfr [Title] AND lung [Title] AND (“2012/01/01” [PDAT]:”2013/09/01” [PDAT])
Exclusion filter	NOT review [ptyp] AND hasabstract [text] NOT “cells” [title/abstract] NOT “cell lines” [title/abstract] NOT systematic [sb] NOT case reports [ptyp]

Table 3

PubMed Clinical Queries and Medical Genetics filters.

PubMed Filter	**Query**
Clinical Description	Natural History OR Mortality OR Phenotype OR Prevalence OR Penetrance AND Genetics
Genetic Testing	DNA Mutational Analysis OR Laboratory techniques and procedures OR Genetic Markers OR diagnosis OR testing OR test OR screening OR mutagenicity tests OR genetic techniques OR molecular diagnostic techniques AND genetics
Diagnosis (broad)	sensitiv* [Title/Abstract] OR sensitivity and specificity [MeSH Terms] OR diagnose [Title/Abstract] OR diagnosed [Title/Abstract] OR diagnoses [Title/Abstract] OR diagnosing [Title/Abstract] OR diagnosis [Title/Abstract] OR diagnostic [Title/Abstract] OR diagnosis [MeSH:noexp] OR diagnostic * [MeSH:noexp] OR diagnosis, differential [MeSH:noexp] OR diagnosis [Subheading:noexp]
Diagnosis (narrow)	specificity [Title/Abstract]
Therapy (narrow)	randomized controlled trial [Publication Type] OR (randomized [Title/Abstract] AND controlled [Title/Abstract] AND trial [Title/Abstract])
Therapy (broad)	(clinical [Title/Abstract] AND trial [Title/Abstract]) OR clinical trials [MeSH Terms] OR clinical trial [Publication Type] OR random* [Title/Abstract] OR random allocation [MeSH Terms] OR therapeutic use [MeSH Subheading]
Management	Therapy [Subheading] OR treatment [Text Word] OR treatment outcome OR investigational therapies AND Genetics
Etiology (broad)	risk*[Title/Abstract] OR risk*[MeSH:noexp] OR risk *[MeSH:noexp] OR cohort studies [MeSH Terms] OR group [Text Word] OR groups [Text Word] OR grouped [Text Word]
Etiology (narrow)	relative [Title/Abstract] AND risk*[Title/Abstract]) OR (relative risk [Text Word]) OR risks [Text Word] OR cohort studies [MeSH:noexp] OR (cohort [Title/Abstract] AND study [Title/Abstract]) OR (cohort [Title/Abstract] AND studies [Title/Abstract]

Table 4

Map of Casama categories to PubMed queries.

Casama Category	Analogous PubMed Query
Characterization	Original query + Exclusion filter + Clinical Description [filter]
Detection	Original query + Exclusion filter + Genetic Testing [filter] Original query + Exclusion filter + Diagnosis/Broad [filter] Original query + Exclusion filter + Diagnosis/Narrow [filter]
Treatment	Original query + Exclusion filter + Therapy/Broad [filter] Original query + Exclusion filter + Therapy/Narrow [filter] Original query + Exclusion filter + Management [filter]
Prognosis	Original query + Exclusion filter + Prognosis/Broad [filter] Original query + Exclusion filter + Prognosis/Narrow [filter]
Experimental studies	Original query + Exclusion filter + Clinical Trial [ptyp]
Cohort studies	Original query + Exclusion filter + Etiology/Broad [filter] Original query + Exclusion filter + Etiology/Narrow [filter] Original query + Exclusion filter + “cohort studies” [MeSH]
Prospective cohort studies	Original query + Exclusion filter + “cohort studies” [MeSH] AND “prospective studies” [MeSH]
Retrospective cohort studies	Original query + Exclusion filter + “cohort studies” [MeSH] AND “retrospective studies” [MeSH]
Cross-sectional studies	Original query + Exclusion filter + “cross-sectional studies” [MeSH]
Case-control studies	Original query + Exclusion filter + “case-control studies” [MeSH]

Table 5

Rules for extracting sparsely-represented study designs.

Study design	Extraction rules
Retrospective	title/abstract contains “retrospective” OR “review” OR “data” OR “charts” OR “records” OR “analyze”
Prospective	title/abstract contains “prospective”
Unknown cohort	any cohort study not matching rules for retrospective or prospective study
Case-control	title/abstract contains “case” AND “control”
Case series	title/abstract contains “series”

Table 6

Kappa scores for entire document collection.

Category	Set A	Set B	Set C	Set D	Set E	Mean	Std Dev
Characterization	0.725	0.563	0.65	0.718	0.65	0.661	0.066
Detection	0.846	0.821	0.689	0.813	0.793	0.792	0.061
Treatment	0.634	0.552	0.705	0.649	0.845	0.677	0.109
Prognosis	0.606	0.518	0.643	0.725	0.527	0.604	0.086
Experimental	0.781	0.860	0.649	0.621	0.731	0.728	0.097
Cohort (all)	0.622	0.636	0.573	0.577	0.633	0.608	0.031
Retrospective Cohort	0.519	0.635	0.560	0.438	0.352	0.501	0.109
Prospective Cohort	0.378	0.488	0.312	0	0	0.236	0.224
Unknown Cohort	0.254	0	0.270	0.497	0.239	0.252	0.176
Cross-sectional	0.835	0.673	0.518	0.74	0.569	0.667	0.128
Case control	n/a	0	0	0	0	0	0
Case series	0	0.222	0.271	0.467	0.271	0.246	0.167

Table 7

Precision, recall, and F-scores for study objective classification. Bold: maximum improvement over PubMed (129%).

EGFR PubMed Cross Validation				**PubMed Filter**			
	P	**R**	**F**		**P**	**R**	**F**
Characterization	0.82	0.68	0.74	Clinical description	0.48	0.39	0.43
Detection	0.96	0.69	**0.80**	Genetic testing	0.22	0.94	0.35
				Diagnosis (broad)	0.36	0.89	0.52
				Diagnosis (narrow)	0.92	0.31	0.47
Treatment	0.84	0.71	0.77	Therapy (broad)	0.35	0.95	0.51
				Therapy (narrow)	0.77	0.26	0.39
				Management	0.25	0.71	0.37
Prognosis	0.76	0.77	0.76	Prognosis (broad)	0.58	0.78	0.67
				Prognosis (narrow)	0.71	0.51	0.59

Table 8

Precision, recall, and F-scores on test sets for study objective classification.

	ALK PubMed			**EGFR ASCO**			**ALK ASCO**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Characterization	0.80	0.62	0.69	0.66	0.63	0.64	0.43	0.75	0.54
Detection	0.93	0.65	0.76	0.75	0.64	0.69	0.75	0.43	0.55
Treatment	0.67	0.67	0.67	0.80	0.83	0.81	1.0	0.67	0.80
Prognosis	0.77	0.58	0.67	0.78	0.62	0.69	1.0	0.25	0.40

Table 9

Precision, recall, and F-scores on training set for study design classification.

EGFR Cross Validation				PubMed Filter			
	P	**R**	**F**		**P**	**R**	**F**
Experimental	0.46	0.65	0.54	Clinical trials	0.48	0.60	0.53
Cross-sectional	0.77	0.80	0.79	Cross-sectional (MeSH)	0	0	0
Cohort (all)	0.81	0.72	0.76	Cohort (MeSH)	0.65	0.48	0.55
				Etiology (broad)	0.68	0.54	0.61
				Etiology (narrow)	0.60	0.10	0.17
Prospective Cohort	0.67	0.29	0.40	Prospective cohort (MeSH)	0.14	0.29	0.19
Retrospective Cohort	0.53	0.66	0.59	Retrospective cohort (MeSH)	0.63	0.57	0.60
Unknown Cohort	0.44	0.23	0.30				
Case-control	n/a	0	n/a	Case-control (MeSH)	0.05	0.67	0.08
Case-series	0.29	0.40	0.33				

Table 10

Study design classification results for test sets.

	ALK PubMed			EGFR ASCO			ALK ASCO		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Experimental	0.60	1.0	0.75	1.0	0.59	0.74	0.75	0.60	0.67
Cross-sectional	0.79	0.85	0.82	0.81	0.93	0.87	0.73	0.80	0.76
Cohort (all)	0.81	0.75	0.78	0.79	0.86	0.82	0.67	0.50	0.57
Prospective Cohort	n/a	0	n/a	n/a	0	n/a	n/a	0	n/a
Retrospective Cohort	0.17	1.0	0.26	0.55	0.80	0.65	0.40	0.25	0.31
Unknown Cohort	0.75	0.25	0.375	0.64	0.33	0.44	0.25	0.25	0.25
Case-control	n/a	0	n/a	n/a	0	n/a	n/a	0	n/a
Case-series	0.33	0.20	0.25	0.50	0.50	0.50	n/a	0	n/a

Table 11

Top features for study objective classification.

Characterization	Detection	Treatment	Prognosis
status	sample	progression	survival
kras	method	advanced	overall survival
higher	serum	mg	prognosis
correlated	detect	median epidermal	overall
conclusive	evaluate	control	analyze
patient	tumour	month	overall prognostic
smoker	dna	symptom	patient egfr
hospital	rearrangement	receive	month
egfr kras	copy	chemotherapy	differ
result	sensitivity	follow	significantly

Table 12

Top features for study design classification.

Experimental	Cohort	Cross-sectional
patient epidermal	cancer patient	exon
toxicity	prognostic	detect
mg	retrospective	result
receive	worse	evaluate
clarify	observe	egfr kras
day	worse	examine
progression	month	prevalence
grade	prognosis	specimen
progression free	differ	pcr
six	significant difference	exon egfr

Table 13

Classification performance on ALK ASCO when trained on EGFR ASCO.

Category	Precision	Recall	F-score
Characterization	0.67	1.0	0.8
Detection	1.0	0.29	0.44
Treatment	1.0	0.87	0.93
Prognosis	0.53	1.0	0.70
Experimental	0.75	0.90	0.82
Cohort (all)	0.86	0.58	0.69
Prospective cohort	n/a	n/a	n/a
Retrospective cohort	0.63	0.63	0.63
Unknown	n/a	0	n/a
Case-control	n/a	0	n/a
Cross-sectional	1.0	0.80	0.89
Case series	n/a	0	n/a