Lever *et al. Genome Medicine* (2019) 11:78
https://doi.org/10.1186/s13073-019-0686-y

Genome Medicine

RESEARCH Open Access

Text-mining clinically relevant cancer biomarkers for curation into the CIViC database

Jake Lever[1,2], Martin R. Jones[1], Arpad M. Danos[3], Kilannin Krysiak[3,4], Melika Bonakdar[1], Jasleen K. Grewal[1,2], Luka Culibrk[1,2], Obi L. Griffith[3,4,5,6*], Malachi Griffith[3,4,5,6*] and Steven J. M. Jones[1,2,7*]

Abstract

Background: Precision oncology involves analysis of individual cancer samples to understand the genes and pathways involved in the development and progression of a cancer. To improve patient care, knowledge of diagnostic, prognostic, predisposing, and drug response markers is essential. Several knowledgebases have been created by different groups to collate evidence for these associations. These include the open-access Clinical Interpretation of Variants in Cancer (CIViC) knowledgebase. These databases rely on time-consuming manual curation from skilled experts who read and interpret the relevant biomedical literature.

Methods: To aid in this curation and provide the greatest coverage for these databases, particularly CIViC, we propose the use of text mining approaches to extract these clinically relevant biomarkers from all available published literature. To this end, a group of cancer genomics experts annotated sentences that discussed biomarkers with their clinical associations and achieved good inter-annotator agreement. We then used a supervised learning approach to construct the CIViCmine knowledgebase.

Results: We extracted 121,589 relevant sentences from PubMed abstracts and PubMed Central Open Access full-text papers. CIViCmine contains over 87,412 biomarkers associated with 8035 genes, 337 drugs, and 572 cancer types, representing 25,818 abstracts and 39,795 full-text publications.

Conclusions: Through integration with CIVIC, we provide a prioritized list of curatable clinically relevant cancer biomarkers as well as a resource that is valuable to other knowledgebases and precision cancer analysts in general. All data is publically available and distributed with a Creative Commons Zero license. The CIViCmine knowledgebase is available at http://bionlp.bcgsc.ca/civicmine/.

Keywords: Precision oncology, Text mining, Information extraction, Machine learning, Cancer biomarkers

Background

The ability to stratify patients into groups that are clinically related is an important step towards a personalized approach to cancer. Over time, a growing number of biomarkers have been developed to select patients who are more likely to respond to certain treatments. These biomarkers have also been valuable for prognostic purposes and for understanding the underlying biology of the disease by defining different molecular subtypes of cancers that should be treated in different ways (e.g., *ERBB2*/*ESR1*/*PGR* testing in breast cancer [1]). Immunohistochemistry techniques are a primary approach for testing samples for diagnostic markers (e.g., CD15 and CD30 for Hodgkin's disease [2]). Recently, the lower cost and increased speed of genome sequencing have also allowed the DNA and RNA of individual patient samples to be characterized for clinical applications [3]. Throughout the world, this technology is beginning to inform clinician decisions on which treatments to use [4]. Such efforts are dependent on a comprehensive and current understanding of the clinical relevance of variants. For example, the

* Correspondence: obigriffith@wustl.edu; mgriffit@wustl.edu; sjones@bcgsc.ca
[3] McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA
[1] Canada's Michael Smith Genome Sciences Centre, Vancouver, BC, Canada
Full list of author information is available at the end of the article

Personalized Oncogenomics project at BC Cancer identifies somatic events in the genome such as point mutations, copy number variations, and large structural changes and, in conjunction with gene expression data, generates a clinical report to provide an 'omic picture of a patient's tumor [5].

The high genomic variability observed in cancers means that each patient sample includes a large number of new mutations, many of which may have never been documented before [6]. The phenotypic impact of most of these mutations is difficult to discern. This problem is exacerbated by the driver/passenger mutation paradigm where only a fraction of mutations are essential to the cancer (drivers) while many others have occurred through mutational processes that are irrelevant to the progression of the disease (passengers). An analyst trying to understand a patient sample typically performs a literature review for each gene and specific variant which is needed to understand its relevance in a cancer type, characterize the driver/passenger role of its observed mutations, and gauge the relevance for clinical decision making.

Several groups have built in-house knowledgebases, which are developed as analysts examine increasing numbers of cancer patient samples. This tedious and largely redundant effort represents a substantial interpretation bottleneck impeding the progress of precision medicine [7]. To encourage a collaborative effort, the CIViC knowledgebase (https://civicdb.org) was launched to provide a wiki-like, editable online resource where community-contributed edits and additions are moderated by experts to maintain high-quality variant curation [8]. The resource provides information about clinically relevant variants in cancer described in the peer-reviewed literature. Variants include protein-coding point mutations, copy number variations, epigenetic marks, gene fusions, aberrant expression levels, and other 'omic events. It supports four types of evidence associating biomarkers with different classes of clinical relevance (also known as evidence types).

Diagnostic evidence items describe variants that can help a clinician diagnose or exclude a cancer. For instance, the *JAK2* V617F mutation is a major diagnostic criterion for myeloproliferative neoplasms to identify polycythemia vera, essential thrombocythemia, and primary myelofibrosis [9]. Predictive evidence items describe variants that help predict drug sensitivity or response and are valuable in deciding further treatments. Predictive evidence items often explain mechanisms of resistance in patients who progressed on a drug treatment. For example, the *ABL1* T315I missense mutation in the *BCR-ABL* fusion predicts poor response to imatinib, a tyrosine kinase inhibitor that would otherwise effectively target *BCR-ABL*, in patients with chronic myeloid leukemia [10]. Predisposing evidence items describe germline variants that increase the likelihood of developing a particular cancer, such as *BRCA1* mutations for breast/ovarian cancer [11] or *RB1* mutations for retinoblastoma [12]. Lastly, prognostic evidence items describe variants that predict survival outcome. As an example, colorectal cancers that harbor a *KRAS* mutation are predicted to have worse survival [13].

CIViC presents this information in a human-readable text format consisting of an "evidence statement" such as the sentence describing the ABL1 T315I mutation above together with data in a structured, programmatically accessible format. A CIViC "evidence item" includes this statement, ontology-associated disease name [14], evidence type as defined above, drug (if applicable), PubMed ID, and other structured fields. Evidence items are manually curated and associated in the database with a specific gene (defined by Entrez Gene) and variant (defined by the curator).

Several groups have created knowledgebases to aid clinical interpretation of cancer genomes, many of whom have joined the Variant Interpretation for Cancer Consortium (VICC, http://cancervariants.org/). VICC is an initiative that aims to coordinate variant interpretation efforts and, to this end, has created a federated search mechanism to allow easier analysis across multiple knowledgebases [15]. The CIViC project is co-leading this effort along with OncoKB [16], the Cancer Genome Interpreter [17], Precision Medicine Knowledge base [18], Molecular Match, JAX-Clinical Knowledge base [19], and others.

Most of these projects focus on clinically relevant genomic events, particularly point mutations, and provide associated clinical information tiered by different levels of evidence. Only CIViC includes RNA expression-based biomarkers. These may be of particular value for childhood cancers which are known to be "genomically quiet, " having accrued very few somatic mutations. Consequently, their clinical interpretation may rely more heavily on transcriptomic data [20]. Epigenomic biomarkers will also become more relevant as several cancer types are increasingly understood to be driven by epigenetic misregulation early in their development [21]. For example, methylation of the MGMT promoter is a well-known biomarker in brain tumors for sensitivity to the standard treatment, temozolomide [22].

The literature on clinically relevant cancer mutations is growing at an extraordinary rate. For instance, only 5 publications in PubMed mentioned BRAF V600E in the title or abstract in 2004 compared to 454 papers in 2017. In order to maintain a high-quality and up-to-date knowledgebase, a curation pipeline must be established. This typically involves a queue for papers, a triage system, and then assignment to a highly experienced curator. This prioritization step is important given the limited time of curators and the potentially vast number of papers to be reviewed. Prioritization must identify papers that contain knowledge that is of current relevance to users of the knowledgebase. For instance, selecting papers for drugs that are no longer clinically approved would not be valuable to the knowledgebase.

Text mining methods have become a common approach to help prioritize literature curation. These methods fall broadly into two main categories, information retrieval (IR) and information extraction (IE). IR methods focus on paper-level information and can take multiple forms. Complex search queries for specific terms or paper metadata (helped by the MeSH term annotations of papers in biomedicine) are common tools for curators. More advanced document clustering and topic modeling systems can use semi-supervised methods to predict whether a paper would be relevant to curation. Examples of this approach include the document clustering method used for the ORegAnno project [23].

IE methods extract structured knowledge directly from the papers. This can take the form of entity recognition, by explicitly tagging mentions of biomedical concepts such as genes, drugs, and diseases. A further step can involve relation extraction to understand the relationship discussed between tagged biomedical entities. This structured information can then be used to identify papers relevant to the knowledgebase. IE methods are also used for automated knowledgebase population without a manual curation step. For example, the miRTex knowledgebase, which collates microRNAs and their targets, uses automated relation extraction methods to populate the knowledgebase [24]. Protein-protein interaction networks (such as STRING [25]) are often built using automatically generated knowledgebases. Our previous work has used information extraction methods to extract the role of genes in cancer but did not identify specific aberrations or the clinical relevance of them [26].

The main objective of this project was to identify frequently discussed cancer biomarkers that fit the CIViC evidence model but are not yet included in the CIViC knowledgebase. We developed an information extraction-based method to extract key parts of the evidence item: cancer type, gene, drug (where applicable), and the specific evidence type from published literature. This allows us to count the number of mentions of specific evidence items in abstracts and full-text articles and compare against the CIViC knowledgebase. We present our methods to develop this resource, known as CIViCmine (http://bionlp.bcgsc.ca/civicmine/). The main contributions of this work are an approach for knowledgebase construction that could be applied to many areas of biology and medicine, a machine learning method for extracting complicated relationships between four entity types, and extraction of relationships across the largest possible publically accessible set of abstracts and full-text articles. This resource, containing 87,412 gene-cancer associations with clinical relevance, is valuable to all cancer knowledgebases to aid their curation and also as a tool for precision cancer analysts searching for evidence supporting biomarkers not yet included in any other resource.

Methods

Corpora

The full PubMed, PubMed Central Open Access (PMCOA) subset, and PubMed Author Manuscript Collection (PMCAMC) corpora were downloaded from the NCBI FTP website using the PubRunner infrastructure [27]. These documents were converted to the BioC format for processing with the Kindred package [28]. HTML tags were stripped out and HTML special characters converted to Unicode. Metadata about the papers were retained including PubMed IDs, titles, journal information, and publication date. Subsections of the paper were extracted using a customized set of acceptable section headers such as "Introduction," "Methods," "Results," and many synonyms of these (accessible through the GitHub repository). The corpora were downloaded in bulk in order to not overload the EUtils RESTFUL service that is offered by the NCBI. The updated files from PubMed were processed to identify the latest version of each abstract to process.

Term lists

Term lists were curated for genes, diseases, and drugs based on several resources. The cancer list was curated from a section of the Disease Ontology [14]. All terms under the "cancer" (DOID:162) parent term were selected and filtered for nonspecific names of cancer (e.g., "neoplasm" or "carcinoma"). These cancer types were then matched with synonyms from the Unified Medical Language System (UMLS) Metathesaurus [29] (2019AA), either through existing external reference links in the Disease Ontology or through exact string-matching on the main entity names. The additional synonyms in the UMLS were then added through this link. The gene list was built from the Entrez gene list and complemented with UMLS terms. Terms that overlapped with common words found in scientific literature (e.g., ice) were removed.

The drug list was curated from the WikiData resource [30]. All Wikidata entities that are medication instances (Wikidata identifier: Q12140) were selected using a SPARQL query. The generic name, brand name, and synonyms were extracted where possible. This list was complemented by a custom list of general drug categories (e.g., chemotherapy, tyrosine kinase inhibitors) and a list of inhibitors built using the previously discussed gene list. This allowed for the extraction of terms such as "EGFR inhibitors." This was done because analysts are often interested in and publications often discuss biomarkers associated with drug classes that target a specific gene.

All term lists were filtered with a stopword list. This was based on the stopword list from the Natural Language Toolkit [31] and the most frequent 5000 words found in the Corpus of Contemporary American English [32] as well as a custom set of terms. It was then merged with common words that occur as gene names (such as ICE).

A custom variant list was built that captured the main types of point mutations (e.g., loss of function), copy number variation (e.g., deletion), epigenetic marks (e.g., promoter methylation), and expression changes (e.g., low expression). These variants were complemented by a synonym list.

The word lists and tools used to generate them are accessible through the BioWordlists project (https://github.com/jakelever/biowordlists) and data can be found in the Zenodo repository (https://doi.org/10.5281/zenodo.1286661).

Entity extraction

The BioC corpora files were processed by the Kindred package. This NLP package used Stanford CoreNLP [33] for processing in the original published version [28]. For this work, it was changed to Spacy [34] for the improved Python bindings in version 2 for this project. This provided easier integration and execution on a cluster without running a Java subprocess. Spacy was used for sentence splitting, tokenization, and dependency parsing of the corpora files. Furthermore, we use the Scispacy parsing model [35].

Exact string matching was then used against the tokenized sentences to extract mentions of cancer types, genes, drugs, and variants. Longer terms were prioritized during extraction so that “non-small cell lung cancer” would be extracted instead of just “lung cancer.” Variants were also extracted with a regular expression system for extracting protein-coding point mutations (e.g., V600E).

Gene fusions (such as *BCR-ABL1*) were detected by identifying mentions of genes separated by a forward slash, hyphen, or colon. If the two entities had no overlapping HUGO IDs, then it was flagged as a possible gene fusion and combined into a single entity. If there were overlapping IDs, it was deemed likely to be referring to the same gene. An example is *HER2/neu* which is frequently seen and refers to a single gene (*ERBB2*) and not a gene fusion. We used the 24 gene fusions associated with acute myeloid leukemia from MyCancerGenome (https://www.mycancergenome.org/) as a sanity check and found that 23 were found in the literature using this method with only RPN1-MECOM missing.

Acronyms were also detected, where possible, by identifying terms in parentheses and checking the term before it, for instance, “non-small cell lung carcinoma (NSCLC).” This was done to remove entity mistakes where possible. The acronym detection method takes the short-form (the term in brackets) and iterates backward through the long-form (the term before brackets) looking for potential matches for each letter. If the long-form and short-form have overlapping associated ontology IDs, they likely refer to the same thing and can be combined, as in the example above. If only one of the long-form or short-form has an associated ontology ID, they are combined and assigned the associated ontology ID. If both long-form and short-form have ontology IDs but there is no overlap, the short-form is disregarded as the long-form has more likelihood of getting the specific term correct.

Gene mentions that are likely associated with signaling pathways and not specific genes (e.g., “MTOR signaling”) are also removed using a simple pattern based on the words after the gene mention. One final post-processing step merges neighboring terms with matching terms. So “HER2 neu” would be combined into one entity as the two terms (*HER2* and *neu*) refer to the same gene.

Sentence selection

With all biomedical documents parsed and entities tagged, all sentences were selected that mention at least one gene, at least one cancer, and at least one variant. A drug was not required as only one (predictive) of the four evidence types involves a drug entity. We evaluated 100 randomly selected sentences and found that only 10 contained information potentially relevant to CIViC, with 7 of the sentences referring to prognostic associations. Many of the sentences report genetic events found in cancer types, methods, and other irrelevant information. Manual annotation of a dataset with only 10% relevance would be hugely inefficient and frustrating for expert annotators. Furthermore, any machine learning system would face a large challenge dealing directly with a class balance of 10%. Therefore, we elected to use a keyword search to enrich the sentences with CIViC relevant knowledge.

Through manual review of a subset of the sentence combined with knowledge of the requirement of CIViC, we selected the keywords found in Table 1. Most of the keywords target a specific association type (e.g., survival for prognostic). This set was not designed to be exhaustive but to keep a reasonable balance of relevant sentences that could be later filtered by a machine learning system. In selecting each keyword, the filtered sentences were evaluated for relevance and the keyword was added if at least half of the sentences seemed relevant to CIViC. The five groups were treated separately such that 20% of the corpus comes from each of the five groups. This was done to provide coverage for the rarer types

Table 1 The five groups of search terms used to identify sentences that potentially discussed the four evidence types. Strings such as “sensitiv” are used to capture multiple words including “sensitive” and “sensitivity”

General	Diagnostic	Predictive	Predisposing	Prognostic
marker	diagnostic	sensitiv	risk	survival
		resistance	predispos	prognos
		efficacy		DFS
		predict		

such as diagnostic that were not found at all in the initial 100 sentences evaluated.

Annotation platform

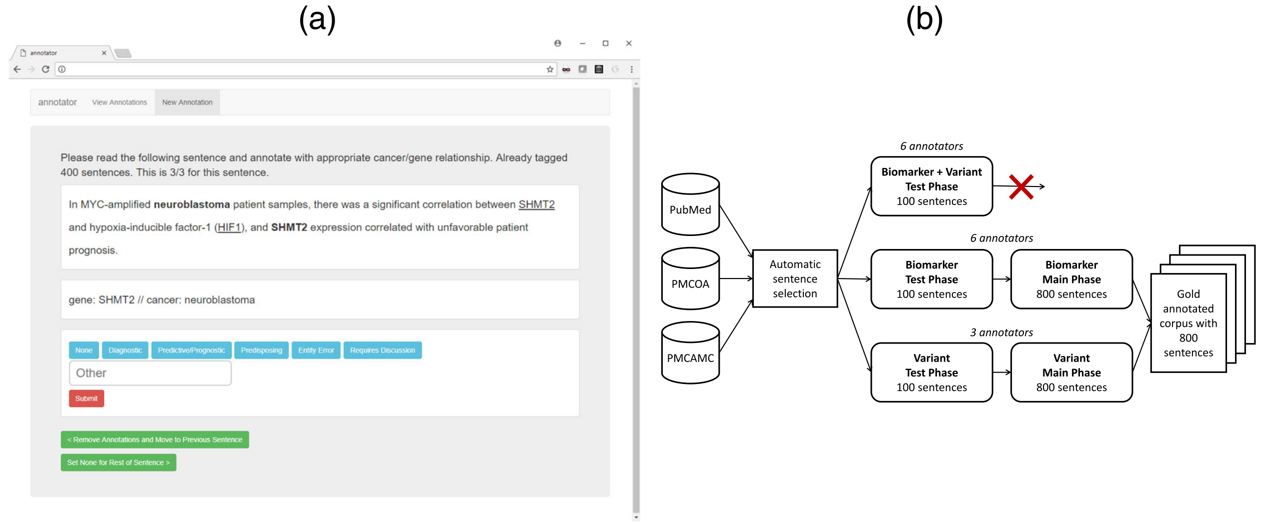
A web platform for simple relation annotation was built using Bootstrap (https://getbootstrap.com/). This allowed annotators to work using a variety of devices, including their smartphones. The annotation system could be loaded with a set of sentences with entity annotations stored in a separate file (also known as standoff annotations). When provided with a relation pattern, for example, “Gene/Cancer,” the system would search the input sentences and find all pairs of the given entity types in the same sentence. It would make sure that the two entities are not the same term, as in some sentences a token (or set of tokens) could be annotated as both a gene name and a cancer type (e.g., “retinoblastoma”). For a sentence with two genes and two cancer types, it would find all four possible pairs of gene and cancer type.

Each sentence, with all the possible candidate relations matching the relation pattern, would be presented to the user, one at a time (Fig. 1a). The user can then select various toggle buttons for the type of relation that these entities are part of. They can also use these to flag entity extraction errors or mark contentious sentences for discussion with other annotators.

Annotation

For the annotation step (outlined in Fig. 1b), the annotated data set (known as the gold set) was constructed using a consensus of multiple annotators. An equal number of sentences were selected from each of the groups outlined in Table 1. This guaranteed coverage of all four evidence types as otherwise the prognostic type dominated the other groups. If this step was not done, 100 randomly selected filtered sentences would only contain 2 (on average) from the diagnostic group. However, this sampling provided poor coverage of sentences that describe specific point mutations. Many precision oncology projects only focus on point mutations and so a further requirement was that 50% of sentences for annotation include a specific point mutation. Altogether, this sampling provides better coverage of the different omic events and evidence types that were of interest. Special care is required when evaluating models built on this customized training set as an unweighted evaluation would not be representative of the real literature.

Sentences that contain many permutations of relationships (e.g., a sentence with 6 genes and 4 cancer types mentioned) were removed. An upper limit of 5 possible



Please read the following sentence and annotate with appropriate cancer/gene relationship. Already tagged 400 sentences. This is 3/3 for this sentence.

In MYC-amplified **neuroblastoma** patient samples, there was a significant correlation between SHMT2 and hypoxia-inducible factor-1 (HIF1), and **SHMT2** expression correlated with unfavorable patient prognosis.

gene: SHMT2 // cancer: neuroblastoma

None
Diagnostic
Predictive/Prognostic
Predisposing
Entity Error
Requires Discussion
Other
Submit
< Remove Annotations and Move to Previous Sentence
Set None for Rest of Sentence >
(b)
PubMed
PMCOA
PMCAMC
Automatic sentence selection
6 annotators
Biomarker + Variant Test Phase 100 sentences
6 annotators
Biomarker Test Phase 100 sentences
Biomarker Main Phase 800 sentences
3 annotators
Variant Test Phase 100 sentences
Variant Main Phase 800 sentences
Gold annotated corpus with 800 sentences

Fig. 1 a A screenshot of the annotation platform that allowed expert annotators to select the relation types for different candidate relations in all of the sentences. The example sentence shown describes a prognostic marker. **b** An overview of the annotation process. Sentences are identified from literature that describes cancers, genes, variants, and optionally drugs before being filtered using search terms. The first test phase tried complex annotation of biomarker and variants together but was unsuccessful. The annotation task was split into two separate tasks for biomarkers and variants separately. Each task had a test phase and then the main phase on the 800 sentences that were used to create the gold set relations was enforced for each sentence. This was done with the knowledge that the subsequent relation extraction step would have a greater false positive rate for sentences with a very large number of possible relations. It was also done to make the annotation task more manageable. An annotation manual was constructed with examples of sentences that would and would not match the four evidence types. This was built in collaboration with CIViC curators and is available in our Github repository (https://github.com/jakelever/civicmine). Each annotation task began with a test phase of 100 sentences. This allows the annotators to become accustomed to the annotation platform and make adjustments to the annotation manual to clarify misunderstandings.

The first test phase (Biomarker + Variant) involved annotating sentences for ternary (gene, cancer, variant) or quaternary (gene, cancer, variant, drug) relationships. The ternary relationships included diagnostic, prognostic, and predisposing, and the quaternary relationship was predictive. As many sentences contain multiple mentions of the same gene or variant, we found there was a combinatorial problem as different annotators found it challenging to decide which variants should be associated with which gene. The annotators were trying to decide linguistically which of the mentions was part of the biomarker being described. For example, in a sentence that mentioned the same variant five times, different annotators chose different mentions of the same variant. These were flagged as differences and reduced the annotator agreement. This led to the low F1-score inter-annotator agreement (average of 0.52) and forced us to reconsider the annotation approach.

To reduce the possible combinations, we split the task into two separate tasks, the biomarker annotation, and the variant annotation. The biomarker annotation involved binary (gene, cancer) and ternary (gene, cancer, drug) relations that described one of the evidence types. The variant annotation task (gene, variant) focused on whether a variant (e.g., deletion) was associated with a specific gene in the sentence. For a sentence containing two genes, two cancer types, and three variants, the original combined task would have 12 combinations that would require annotation. By splitting it into the two tasks, the biomarker task would have four combinations and the variant task would also have four combinations. We hypothesized that a smaller number of combinations would reduce the cognitive load for the annotators and increase inter-annotator agreement. To further reduce complexity, the predictive and prognostic evidence types were merged (as shown in Fig. 2), to further reduce the annotation complexity. The predictive/prognostic annotations could be separated after tagging as relationships containing a drug would be predictive and those without would be prognostic. A further postprocessing step to generate the gold set involved identifying prognostic relationships that overlapped with predictive relationships (i.e., shared the same gene and cancer type in a sentence) and removing them.

With the redefined annotation task, six annotators were involved in biomarker annotation, all with knowledge of the CIViC platform and having experience interpreting patient cancer variants in a clinical context. Three annotators (one of whom was involved in the biomarker annotation) were involved in variant annotation and they all had experience in cancer genomics. Both annotation tasks started with a new 100-sentence test phase to evaluate the redefined annotation tasks and resolve any ambiguity within the annotation manuals. Good inter-annotator agreement was achieved at this stage for both the biomarker annotation (average F1-score = 0.68) and variant annotation (average F1-score = 0.95). The higher agreement scores validated our reasoning to split the annotation task in two. In fact, the very high variant annotation score suggests that this task was made relatively easy by separating it. These 100 sentences were discarded as they exhibited a learning curve as annotators become comfortable with the task. Between each annotation stage, the annotators discussed through video conference the difficulties that had been encountered. These comments were used to improve the annotation manuals with the aim to capture sentences with greater relevance to CIViC and also increase inter-annotator agreement.

To generate the highest possible annotations, each sentence would be annotated by three different annotators and a majority voting system used to resolve conflicts. As there were six annotators for the biomarker annotation task, we split them into two groups who would work on each half of the 800-sentence corpus. Separately, three annotators worked on variant annotation with the 800-sentence set. Table 2 shows the inter-annotator agreement for these tasks for the full 800 sentences. The inter-annotator agreement is even higher for the biomarker task than the initial 100-sentence test suggesting that the refinements to the annotation manual and the video conference discussions helped. The biomarker and variant annotations are then merged to create the gold corpus of 800 sentences used for the machine learning system.

Relation extraction

The sentences annotated with relations were then processed using the Kindred relation extraction Python package. Relation extraction models were built for all five of the relation types: the four evidence types (diagnostic, predictive, predisposing, and prognostic) and one associated variant relation type. Three of the four evidence type relations are binary between a gene entity and a

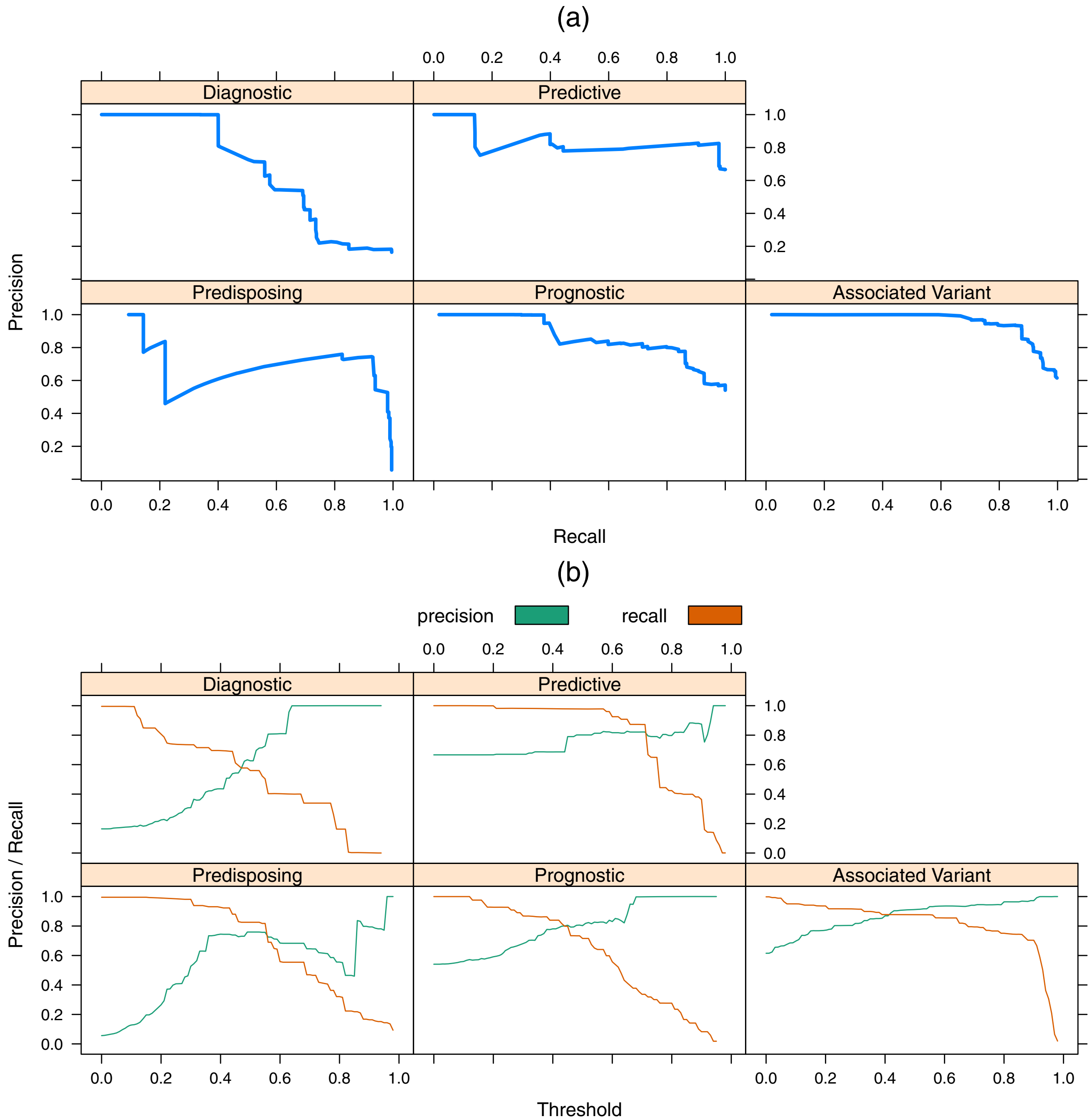


Fig. 2 **a** The precision-recall curves illustrate the performance of the five relation extraction models built for the four evidence types and the associated variant prediction. **b** This same data can be visualized in terms of the threshold values on the logistic regression to select the appropriate value for high precision with reasonable recall

cancer entity. The associated variant relation type is also binary between a gene entity and a variant entity. The predictive evidence item type was ternary between a gene, a cancer type, and a drug.

Most relation extraction systems focus on binary relations [36, 37] and use features based on the dependency path between those two entities. The recent BioNLP Shared Task 2016 series included a subtask for non-binary relations (i.e., relations between three or more entities), but no entries were received [38]. Relations between 2 or more entities are known as n-ary relations where $n \geq 2$. The Kindred relation extraction package, based on the VERSE relation extraction tool [39], which won part of the BioNLP Shared Task 2016, was enhanced to allow prediction of n-ary relations. First, the candidate relation builder was adapted to search for relations of a fixed n which may be larger than 2. This meant that sentences with 5 non-overlapping tagged

Table 2 The inter-annotator agreement for the main phase for 800 sentences, measured with F1-score, showed good agreement in the two sets of annotations for biomarkers as well as very high agreement in the variant annotation task. The sentences from the multiple test phases are not included in these numbers and were discarded from further analysis

	Annotator 2	Annotator 3
Annotator 1	0.74	0.73
Annotator 2	NA	0.74
Annotator 1	0.78	0.85
Annotator 2	NA	0.79
Annotator 1	0.96	0.96
Annotator 2	NA	0.96

Table 3 Number of annotations in the training and test sets

Annotation	Train	Test
Associated variant	768	270
Diagnostic	156	62
Predictive	147	43
Predisposing	125	57
Prognostic	232	88

entities would generate 60 candidate relations with $n = 3$. These candidate relations would then be pruned by entity types. Hence, for the predictive relation type (with $n = 3$), the first entity must be a cancer type, the second a drug, and the third a gene. Two of the features used are based on the path through the dependency graph between the entities in the candidate relation. For relations with more than two entities, Kindred made use of a minimal spanning tree within the dependency graph. The default Kindred features were then constructed for this subgraph and the associated entities and sentences. All features were represented with 1-hot vectors or bag-of-word representations.

During training, candidate relations are generated with matching n-ary to the training set. Those candidate relations that match a training example are flagged as positive examples with all others as negative. These candidate relations are vectorized, and a logistic regression classifier is trained against them. The logistic regression classifier outputs an interpretable score akin to a probability for each relation, which was later used for filtering. Kindred also supports a Support Vector Machine classifier (SVM) or can be extended with any classifier from the scikit-learn package [40]. The logistic regression classifier was more amenable to adjustment of the precision-recall tradeoff.

For generation of the knowledgebase, the four evidence type relations were predicted first which provided relations including a gene. The associated variant relation was then predicted and attached to any existing evidence type relation that included that gene.

Evaluation

With the understanding that the annotated sentences were selected randomly from customized subsets and not randomly from the full population, care was taken in the evaluation process.

First, the annotated set of 800 sentences was split 75%/25% into a training and test set that had similar proportions of the four evidence types (Table 3). Each sentence was then tracked with the group it was selected from (Table 1). Each group has an associated weight based on the proportion of the entire population of possible sentences that it represents. Hence, the prognostic group, which dominates the others, has the largest weight. When comparing predictions against the test set, the weighting associated with each group was then used to adjust the confusion matrix values. The goal of this weighting scheme was to provide performance metrics which would be representative for randomly selected sentences from the literature and not for the customized training set.

Precision-recall tradeoff

Figure 2a shows precision-recall curves for all five of the relation types. The diagnostic and predisposing tasks are obviously the most challenging for the classifier. This same data can be visualized by comparing the threshold values used against the output of the logistic regression for each metric (Fig. 2b).

To provide a high-quality resource, we decided on a trade-off of high precision with low recall. We hypothesized that the most commonly discussed cancer biomarkers, which are the overall goal of this project, would appear in many papers using different wording. These frequently mentioned biomarkers would then be likely picked up even with lower recall. This also reduces the burden on CIViC curators to sift through false positives. With this, we selected thresholds that would give as close to 0.9 precision given the precision-recall curves for the four evidence types. We require a higher precision for the variant annotation (0.94). The thresholds and associated precision-recall tradeoffs are shown for all five extracted relations in Table 4.

Application to PubMed, PMCOA, and PMCAMC with updates

With the thresholds selected, the final models were applied to all sentences extracted from PubMed, PMCOA, and PMCAMC. This is a reasonably large computational problem and was tasked to the compute cluster at the Canada's Michael Smith Genome Sciences Centre.

To manage this compute and provide infrastructure for easy updating with new publications in all three corpora, we made use of the updated PubRunner infrastructure

Table 4 The selected thresholds for each relation type with the high precision and lower recall trade-off

Extracted relation	Threshold	Precision	Recall
Associated variant	0.70	0.941	0.794
Diagnostic	0.63	0.957	0.400
Predictive	0.93	0.891	0.141
Predisposing	0.86	0.837	0.218
Prognostic	0.65	0.878	0.414

(paper in preparation - https://github.com/jakelever/pubrunner). This allows for easy distribution of the work across a compute cluster. The resulting data was then pushed to Zenodo for perpetual and public hosting [41]. The data is released with a Creative Commons Public Domain (CC0) license so that other groups can easily make use of it.

The PubRunner infrastructure enables the easy update of the resource. We plan to update the resource every month. It manages the download and execution of the tool as well as the upload of the data to the Zenodo repository.

CIViC matching

To make comparisons with CIViC, we downloaded the nightly data file from CIViC (https://civicdb.org/releases – downloaded on 24 September 2019) and matched evidence items against items in CIViCmine. The evidence type and IDs for genes and cancers were used for matching. Direct string matching was used to compare drug names for predictive biomarkers. The exact variant was not used for comparison in order to find genes that contain any biomarkers that match between the two resources.

Some mismatches occurred with drug names. For example, CIViCmine may capture information about the drug family while CIViC contains information on specific drugs or a list of drugs. Another challenge with matching with CIViCmine is related to the similarity of cancer types in the Disease Ontology. Several pairs of similar cancers types are used interchangeably by some researchers and not by others, e.g., stomach cancer and stomach carcinoma. CIViC may contain a biomarker for stomach cancer and CIViCmine matches all the other details except it relates it to stomach carcinoma.

User interface

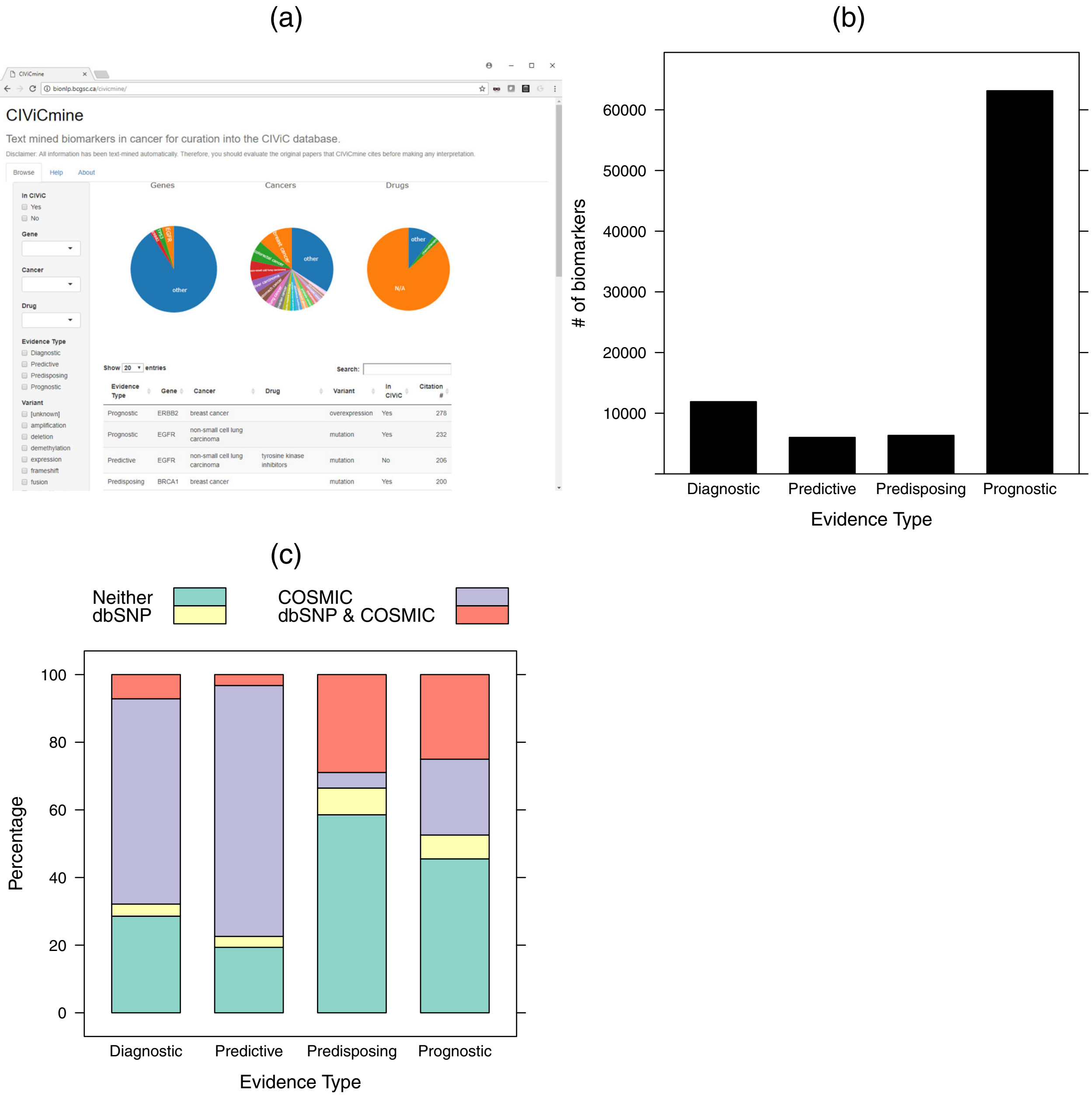
To make the data easily explorable, we provide a Shiny-based front-end (Fig. 3a) [42]. This shows a list of biomarkers extracted from abstracts and papers, which can be filtered by the Evidence Type, Gene, Cancer Type, Drug, and Variant. To help prioritize the biomarkers, we use the number of unique papers in which they are mentioned as a metric. By default, the listed biomarkers are shown with the highest citation count first. Whether the biomarker is found in CIViC is also shown as a column and is an additional filter. The CIViC information is updated daily by downloading the latest nightly release. This allows CIViC curators to quickly navigate to biomarkers not currently discussed in CIViC and triage them efficiently.

With filters selected, the user is presented with pie charts that illustrate the representation of different cancer types, genes, and drugs. When the user clicks on a particular biomarker, an additional table is populated with the citation information. This includes the journal, publication year, section of the publication (e.g., title, abstract or main body), subsection (if cited from the main body), and the actual text of the sentence from which the relationship was extracted. This table can further be searched and sorted, for example, to look for older citations or citations from a particular journal. The PubMed ID is also provided with a link to the citation on PubMed.

Results

From the full PubMed corpus and all downloadable papers from PubMed Central, we extracted 87,412 biomarkers with a breakdown into the four types (Fig. 3b). As expected, based on our preliminary analysis, there are many more prognostic evidence items than the other three types. Table 5 outlines examples of all four of these evidence types. 34.8% of sentences (42,363/121,589) contain more than one evidence item, such as the predictive example which relates *EGFR* as a predictive marker in NSCLC to both erlotinib and gefitinib. In total, we extracted 186,659 mentions of biomarkers from 67,210 unique papers. These biomarkers relate to 8035 genes, 572 cancer types, and 337 drugs. We further delved into the variants extracted for each of the evidence types. For extracting protein-coding mutations, we are unable to ascertain directly from the text if they are germline or somatic. Instead, we compared them with entries in COSMIC [43] that were tagged as somatic and dbSNP [44] that were found above 1% in the population. Figure 3c shows that, as expected, the predisposing type is most strongly associated with germline variants. Interestingly, many of the prognostic variants are also germline while diagnostic and predictive variants are more likely to be somatic.

EGFR and *TP53* stand out as the most frequently extracted genes in different evidence items (Fig. 4a). Over 50% of the *EGFR* evidence items are associated with lung cancer or non-small cell lung carcinoma (NSCLC). *CDKN2A* has a larger proportion of diagnostic biomarkers associated with it than most of the other genes in the top 20. *CDKN2A* expression is a well-established marker for distinguishing HPV+ versus HPV− cervical



CIViCmine

Text mined biomarkers in cancer for curation into the CIViC database.

Disclaimer: All information has been text-mined automatically. Therefore, you should evaluate the original papers that CIViCmine cites before making any interpretation.

In CIViC
□ Yes
□ No

Gene

Cancer

Drug

Evidence Type
□ Diagnostic
□ Predictive
□ Predisposing
□ Prognostic

Variant
□ [unknown]
□ amplification
□ deletion
□ demethylation
□ expression
□ frameshift
□ fusion

Show 20 entries Search:

Evidence Type	Gene	Cancer	Drug	Variant	In CIViC	Citation #
Prognostic	ERBB2	breast cancer		overexpression	Yes	278
Prognostic	EGFR	non-small cell lung carcinoma		mutation	Yes	232
Predictive	EGFR	non-small cell lung carcinoma	tyrosine kinase inhibitors	mutation	No	206
Predisposing	BRCA1	breast cancer		mutation	Yes	200

Fig. 3 a A Shiny-based web interface allows for easy exploration of the CIViCmine biomarkers with filters and overview pie charts. The main table shows the list of biomarkers and links to a subsequent table showing the list of supporting sentences. **b** The entirety of PubMed and PubMed Central Open Access subset were processed to extract 87,412 biomarkers distributed between the four different evidence types shown. **c** Protein-coding variants extracted for each evidence item are compared against somatic variants in COSMIC and > 1% prevalence SNPs in dbSNP

cancers. Its expression or methylation states are discussed as diagnostic biomarkers in a variety of other cancer types including colorectal cancer and stomach cancer.

Breast cancer is, by far, the most frequently discussed cancer type (Fig. 4b). A number of the associated biomarkers focus on predisposition, as breast cancer has one of the strongest hereditary components associated with germline mutations in *BRCA1* and *BRCA2*. NSCLC shows the largest relative number of predictive biomarkers, consistent with the previous figure showing the importance of *EGFR*.

For the predictive evidence type, we see a disproportionally large number associated with the general term chemotherapy and specific types of chemotherapy including cisplatin, paclitaxel, and doxorubicin (Fig. 4c). Many targeted therapies are also frequently discussed such as the *EGFR* inhibitors, gefitinib, erlotinib, and cetuximab. More

Table 5 Four example sentences for the four evidence types extracted by CIViCmine. The associated PubMed IDs are also shown for reference

Type	PMID	Sentence
Diagnostic	29214759	JAK2 V617F is the most common mutation in myeloproliferative neoplasms (MPNs) and is a major diagnostic criterion.
Predictive	28456787	In non-small cell lung cancer (NSCLC) driver mutations of EGFR are positive predictive biomarkers for efficacy of erlotinib and gefitinib.
Predisposing	28222693	Our study suggests that one BRCA1 variant may be associated with increased risk of breast cancer.
Prognostic	28469333	Overexpression of Her2 in breast cancer is a key feature of pathobiology of the disease and is associated with poor prognosis.

general terms such as “tyrosine kinase inhibitor” capture biomarkers related to drug families.

Lastly, we see that expression related biomarkers dominate the variant types (Fig. 4d). Markers based on expression are more likely to be prognostic than those using non-expression data (83.3% versus 45.2%). The popular approach to exploring the importance of a gene in a cancer type is to correlate expression levels with patient survival. With the extended historical use of immunohistochemical methods as well as the accessibility of large transcriptome sets and survival data (e.g., TCGA), such associations have become very common. The “mutation” variant type has a more even split across the four evidence types. The mutation term covers very general phrasing without a mention of a specific mutation. The substitution variant type does capture this information but there are far fewer than biomarkers with the “mutation” variant type. This reflects the challenge of extracting all of the evidence item information from a single sentence. It is more likely for an author to define a mutation in another section of the paper or aggregate patients with different mutations within the same gene and then use a general term (e.g., *EGFR* mutation) when discussing its clinical relevance. There are also a substantial number of evidence items where the variant cannot be identified and are flagged as “[unknown].” These are still valuable but may require more in-depth curation to identify the actual variant.

Of all the biomarkers extracted, 21.4% (18,709/ 87,412) are supported by more than one citation. The most cited biomarker is *BRCA1* mutation as a predisposing marker in breast cancer with 682 different papers discussing this. The initial priority for CIViC annotation is on highly cited biomarkers that have not yet been curated into CIViC, to eliminate obvious information gaps. However, the single citations may also represent valuable information for precision cancer analysts and CIViC curators focused on specific genes or diseases.

We compared the 87,412 biomarkers extracted by CIViCmine with the 2518 in the CIViC resource as of 24 September 2019. The first Venn diagram in Fig. 5a shows the overlap of exact evidence items between the two resources. The overlap is quite small and the number evidence extracted in CIViCmine not yet included in CIViC is very large. The associations that are unique to CIViCmine would likely contain curatable associations that should be added to CIViC. The associations that are unique to CIViC indicate limitations of this method. Many of these associations are likely not described within a single sentence or are in publications for which the full-text is inaccessible. Furthermore, this approach is most successful with variants that are mentioned multiple times in the literature and will have a harder time with associations mentioned only a single time.

We next compare the cited publications using PubMed ID. Despite not having used CIViC publications in training CIViCmine, we find that a substantial number of papers cited in CIViC (294/1474) were identified automatically by CIViCmine. The remaining ~ 1100 papers were likely not identified as they did not contain a single sentence that contained all the information necessary for extraction. Future methods that can identify biomarkers discussed across multiple sentences would likely identify more of these papers. Altogether, CIViCmine includes 6600 genes, 443 cancer types, and 251 drugs or drug families not yet included in CIViC.

We further compared CIViCmine with the Cancer Genome Interpreter (CGI) and OncoKB resources, two more resources that are part of the VICC consortium. We compare the CGI biomarkers dataset against CIViCmine predictive variants and the CGI cancer genes marked as predisposing against CIViCmine predisposing genes in Fig. 5a. While we find reasonable overlap with the small set of predisposing genes, the overlap with predictive biomarkers is very small. While there are challenges mapping one knowledgebase to another (e.g., making sure that disease identifiers match up), a manual inspection suggested that this was only a minor issue and that the two datasets do not overlap well. Furthermore, the overlap of biomarkers from OncoKB and CIViCmine predictive variants is also very small. The CIViCmine system is designed to best capture biomarkers that are mentioned multiple times in the literature within a single sentence. This suggests that many of the biomarkers in the Cancer Genome Interpreter and OncoKB are not mentioned many times in the literature. Finally, it strongly suggests that the CIViCmine resource is valuable to the broader community as it contains vast numbers of associations that should be added to these other resources.

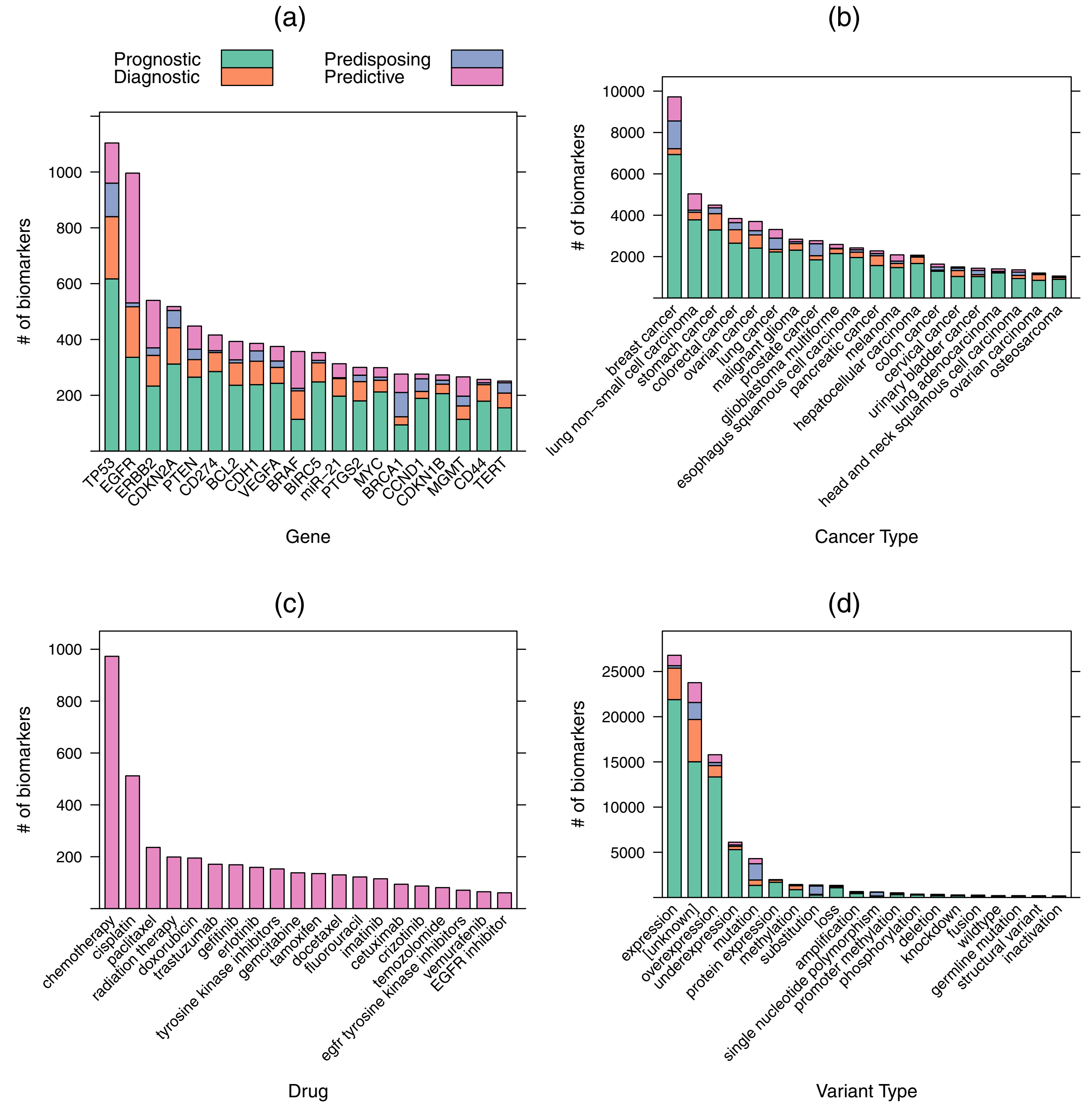


Fig. 4 The top 20 **a** genes, **b** cancer types, **c** drugs, and **d** variants extracted as part of evidence items

Use cases

There are two use cases of this resource that are already been realized by CIViC curators at the McDonnell Genome Institute and analysts at BC Cancer.

Knowledgebase curation use case: The main purpose of this tool is to assist in curation of new biomarkers in CIViC. A CIViC curator, looking for a frequently discussed biomarker, would access the CIViCmine Shiny app through a web browser. This would present the table, pie charts, and filter options on the left. They would initially filter the CIViCmine results for those not already in CIViC. If they had a particular focus, they may filter by evidence type. For example, some CIViC curators may be more interested in diagnostic, predictive, and prognostic biomarkers than predisposing. This is due to the relative importance of somatic events in many cancer types. They would then look at the table of biomarkers, already sorted by citation count in descending order, and select one of the top ones. This would then populate a table further down the page. Assuming

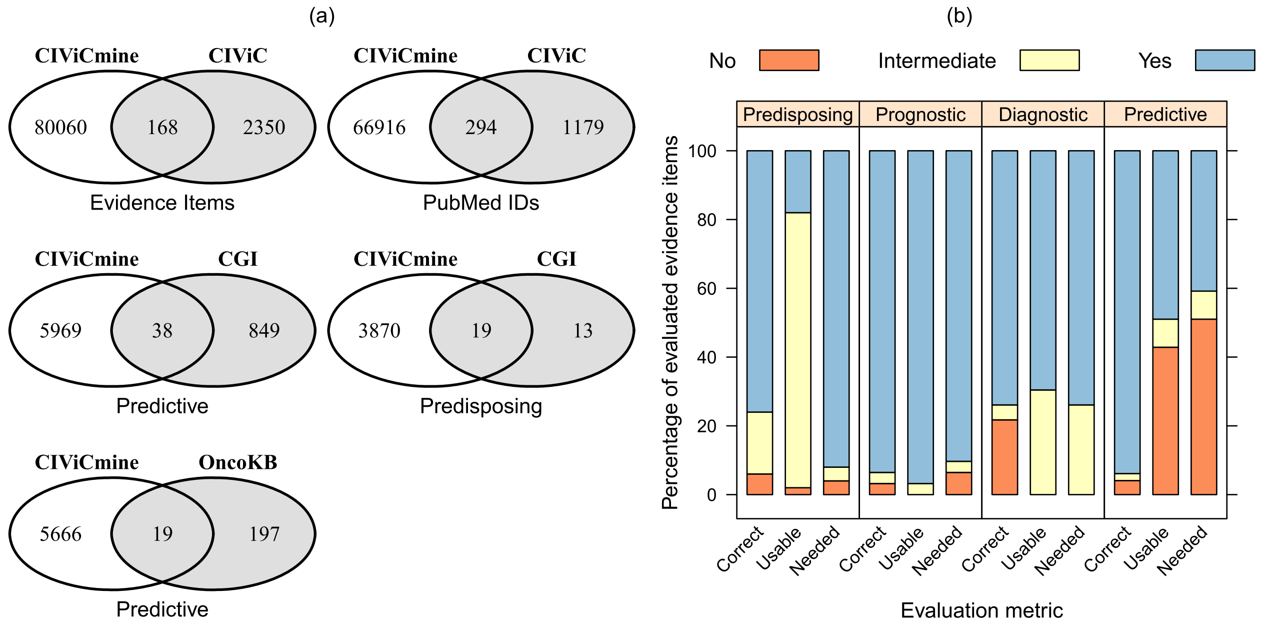


Fig. 5 **a** A comparison of the associations and papers in CIViCmine with CIViC, the Cancer Genome Interpreter and OncoKB. **b** The top results in CIViCmine were evaluated by a CIViC curator and measured for three categories (correctness, usability, and need). Percentages are shown for each metric and evidence type for no, intermediate, and yes

that this is a frequently cited biomarker, there would be many sentences discussing it, which would quickly give the curator a broad view of whether it is a well-supported association in the community. They might then open multiple tabs on their web browser to start looking at several of the papers discussing it. They might select an older paper, close to when it was first established as a biomarker, and a more recent paper from a high-impact journal to gauge the current view of the biomarker. Several of the sentences may cite other papers as being important to establishing this biomarker. The curator would look at these papers in particular, as they may be the most appropriate to curate. Importantly, the curator can use this to identify the primary literature source(s), which includes the experimental data supporting this biomarker.

Personalized cancer analyst use case: While interpreting an individual patient tumor sample, an analyst typically needs to interpret a long list of somatic events. Instead of searching PubMed for each somatic event, they can initially check CIViC and CIViCmine for existing structured knowledge on the clinical relevance of each somatic event. First, they should check CIViC given the high level of pre-existing curation there. This would involve searching the CIViC database through their website or API. If the variant does not appear there, they would then progress to CIViCmine. By using the filters and search functionality, they could quickly narrow down the biomarkers for their gene and cancer type of interest. If a match is found, they can then move to the relevant papers that are listed below to understand the experiments that were done to make this assertion. As they evaluate this biomarker, they could enter this evidence and all of the structured fields that may be spread throughout the publication into the CIViC database. Both CIViC and CIViCmine reduce curation burden by aggregating likely applicable data across multiple synonyms for the gene, disease, variant, or drug not as easily identified through PubMed searches.

Evaluation by CIViC curator

To evaluate the curation value of the data provided by CIViCmine, a CIViC curator evaluated the top biomarkers identified by CIViCmine that were not found in CIViC. Biomarkers with high citation counts were selected for each evidence type and filtered for those which the variant was also extracted. They were then evaluated for correctness (whether the sentences matched the extracted structured data), usability (whether there was enough information for curation into CIViC contained within the sentence), and need (whether this information was lacking in CIViC). Each biomarker was marked in all three categories with yes, intermediate, and no. Intermediate scores are used to identify cases where additional information (e.g., reading the full paper or its citations) was needed. Figure 5b shows the summary of the results as percentages for each of the three metrics across the four evidence types.

Overall, the results are very positive with 73% of evaluated biomarkers being deemed needed by CIViC. The predictive evidence type was found to have a larger proportion of unneeded evidence items. This was due to the catch-all groups (e.g., *EGFR* inhibitors) that were deemed to be too vague for inclusion into CIViC but might provide valuable information for other clinical researchers. The high percentage of intermediate for the usability of predisposing biomarkers was due to the general variant terms identified (such as mutation) where the exact variant was unclear and further curation would be needed. Overall, these results show that CIViCmine provides valuable data that can be curated into CIViC and other knowledgebases.

Discussion

This work provides several significant contributions to the fields of biomedical text mining and precision oncology. Firstly, the annotation method is drastically different from previous approaches. Most annotation projects (such as the BioNLP Shared Tasks [45, 46] and the CRAFT corpus [47]) have focused on abstracts or entire documents. The biomarkers of interest for this project appear sparsely in papers so it would have been inappropriate to annotate full documents and a focus on individual sentences was necessary. In selecting sentences, we aimed for roughly half the sentences to contain positive relations. This would enable better classifier training with a more even class balance. Therefore, we filtered the sentences with a series of keywords after identifying those that contain the appropriate entities. This approach could be applied to many other biomedical topics.

We also made use of a simpler annotation system than the often used brat [48] which allowed for fast annotation by restricting the possible annotation options. Specifically, annotators did not select the entities but were shown all appropriate permutations that matched the possible relation types. Issues of incorrect entity annotation were reported through the interface, collated, and used to make improvements to the underlying wordlists for gene, cancer types, and drugs. We found that once a curator became familiar with the task, they could curate sentences relatively quickly with approximately 1–2 min spent on each sentence. Expert annotation is key to providing high-quality data to build and evaluate a system. Therefore, reducing the time required for expert annotators is essential.

The supervised learning approach differs from methods that used co-occurrence based (e.g., STRING [25]) or rule-based (e.g., mirTex [24]) methods. Firstly, the method can extract complex meaning from the sentence providing results that would be impossible with a co-occurrence method. A rule-based method would require enumerating the possible ways of describing each of the diverse evidence types. Our approach can capture a wide variety of biomarker descriptions. Furthermore, most relation extraction methods aim for optimal F1-score [38], placing an equal emphasis on precision and recall. To minimize false positives, our approach of high precision and low recall would be an appropriate model for other information extraction methods applied to the vast PubMed corpus.

Apart from the advantages outlined previously, several other factors lead to the decision to use a supervised learning approach to build this knowledgebase. The CIViC knowledgebase could have been used as training data in some form. The papers already in CIViC could have been searched for the sentences discussing the relevant biomarker, which could then have been used to train a supervised relation extraction system. An alternative approach to this problem would have been to use a distant supervision method using the CIViC knowledgebase as seed data. This approach was taken by Peng et al., who also attempted to extract relations across sentence boundaries [49]. They chose to focus only on point mutations and extracted 530 within-sentence biomarkers and 1461 cross-sentence biomarkers. These numbers are substantially smaller than the 70,655 extracted in CIViCmine.

The reason to not use the CIViC knowledgebase in the creation of the training data was taken to avoid any curator-specific bias that may have formed in the selection of papers and biomarkers already curated. Avoiding this approach was key to providing a broad and unbiased view of the biomarkers discussed in the literature. CIViC evidence items include additional information such as directionality of a relationship (e.g., does a mutation cause drug sensitivity or resistance), whether the variant is germline or somatic, the level of support for it (from preclinical models up to FDA guidelines) and several other factors. It is highly unlikely that all this information will be included within a single sentence. Therefore, we did not try to extract this information concurrently. Instead, it is an additional task for the curator as they process the CIViCmine prioritized list. While single gene biomarkers are the most commonly discussed findings, there are an increasing number of multi-gene markers or more complex interactions involving multiple variants or treatments. Our system focuses on mapping a single gene, with a single variant (where possible) with a single cancer type and a single drug (for predictive evidence items). Further research would be needed to extract these complex associations, especially as they are more likely to span multiple sentences. It is also challenging to judge the immediate clinical utility of the extracted biomarkers as their use would rely on the data accessible to a clinician (e.g., whether they have panel, whole-genome sequencing, or expression data).

A robust named entity recognition solution does not exist for a custom term list of cancer types, drugs, and variants. For instance, the DNorm tool [50] does not capture many cancer subtypes. A decision was made to go for high recall for entity recognition, including genes, as the relation extraction step would then filter out many incorrect matches based on context. This decision is further supported by the constant evolution of cancer type ontologies, as demonstrated by workshops at recent Biocuration conferences.

CIViCmine has two limitations that are shared by almost all text-mined knowledgebases, access to the published literature for text-mining, and the focus on sentences as the unit of discovery. PubMed contains over 20 million abstracts but PubMed Central only contains approximately 2 million full-text articles. It has been shown many times that the full-text contains the majority of text-mineable information but over 90% of papers are behind paywalls. Furthermore, the supplementary materials may also provide further text for text mining, but the lack of standardization in accessing this text is a large obstacle. Text mining methods are also broadly limited to focusing on single sentences due to the huge challenges that remain in coreference resolution to link pronouns to entities in other sentences. It is incredibly difficult to quantify how much knowledge is lost due to this limitation, but as the associations become more complicated and include more entities, the recall will drop substantially. The limitation is likely one of the main reasons for the poor overlap with the other knowledgebases.

Conclusions

The CIViCmine resource, accessible at http://bionlp.bcgsc.ca/civicmine, and freely available associated data provide a valuable addition to the precision oncology informatics community. CIViCmine can be used to assist curation of other precision cancer knowledgebases and can be used directly by precision cancer analysts to search for biomarkers of interest. As this resource will be updated monthly with the latest research, it will constantly change as new cancer types and drug names enter the lexicon. We anticipate that the methods described can be used in other biomedical domains and that the resources provided will be valuable to the biomedical text mining and precision oncology fields.

Acknowledgements
The authors would like to thank Compute Canada for the use of compute infrastructure for this research.

Authors' contributions
JL, MRJ, OLG, MG, and SJMJ developed the project idea. JL developed the methods, did the analyses, and wrote the initial paper draft. JL, MRJ, AMD, KK, MB, JG, LC, and MG annotated text data. OLG, MG, and SJMJ supervised the project. All authors read and approved the final manuscript.

Funding
CIViC is supported by the National Cancer Institute (NCI) of the National Institutes of Health (NIH) under award number U01CA209936. JL is supported by a Vanier Canada Graduate Scholarship. MG is supported by the NHGRI under award number R00HG007940.

Availability of data and materials
The complete datasets are available in the Zenodo repository (https://doi.org/10.5281/zenodo.1472826) with the September 2019 release used for this paper [41]. The data can also be viewed through the web viewer (http://bionlp.bcgsc.ca/civicmine/) and subsets of the data can be downloaded there. The code for the analysis and web viewer is available at GitHub (https://github.com/jakelever/civicmine/).

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Canada's Michael Smith Genome Sciences Centre, Vancouver, BC, Canada. [2]University of British Columbia, Vancouver, BC, Canada. [3]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA. [4]Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO, USA. [5]Division of Oncology, Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA. [6]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA. [7]Simon Fraser University, Burnaby, BC, Canada.

Received: 17 June 2019 Accepted: 7 November 2019
Published online: 03 December 2019

References

1. Onitilo AA, Engel JM, Greenlee RT, Mukesh BN. Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival. Clin Med Res. 2009;7(1–2):4–13.
2. Rüdiger T, Ott G, Ott MM, Müller-Deubert SM, Müller-Hermelink HK. Differential diagnosis between classic Hodgkin's lymphoma, T-cell-rich B-cell lymphoma, and paragranuloma by paraffin immunohistochemistry. Am J Surg Pathol. 1998;22(10):1184–91.
3. Prasad V, Fojo T, Brada M. Precision oncology: origins, optimism, and potential. Lancet Oncol. 2016;17(2):e81–6.
4. Shrager J, Tenenbaum JM. Rapid learning for precision oncology. Nat Rev Clin Oncol. 2014;11(2):109–18.
5. Laskin J, Jones S, Aparicio S, Chia S, Ch'ng C, Deyell R, et al. Lessons learned from the application of whole-genome analysis to the treatment of patients with advanced cancers. Mol Case Stud. 2015;1(1):a000570.
6. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol. 2016;34(2):155.
7. Good BM, Ainscough BJ, McMichael JF, Su AI, Griffith OL. Organizing knowledge to enable personalization of medicine in cancer. Genome Biol. 2014;15(8):438.
8. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49(2):170.
9. Mesa R, Jamieson C, Bhatia R, Deininger MW, Gerds AT, Gojo I, et al. Myeloproliferative neoplasms, version 2.2017, NCCN clinical practice guidelines in oncology. J Natl Compr Cancer Netw. 2016;14(12):1572–611.
10. Branford S, Rudzki Z, Walsh S, Parkinson I, Grigg A, Szer J, et al. Detection of BCR-ABL mutations in patients with CML treated with imatinib is virtually always accompanied by clinical resistance, and mutations in the ATP phosphate-binding loop (P-loop) are associated with a poor prognosis. Blood. 2003;102(1):276–83.
11. King M-C, Marks JH, Mandell JB. Others. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science. 2003;302(5645):643–6.

12. Harbour JW. Overview of rb gene mutations in patients with retinoblastoma: implications for clinical genetic screening1. Ophthalmology. 1998;105(8):1442–7.
13. Phipps AI, Buchanan DD, Makar KW, Win AK, Baron JA, Lindor NM, et al. KRAS-mutation status in relation to colorectal cancer survival: the joint impact of correlated tumour markers. Br J Cancer. 2013;108(8):1757.
14. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, et al. Disease ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2011;40(D1):D940–6.
15. Wagner AH, Walsh B, Mayfield G, Tamborero D, Sonkin D, Krysiak K, et al. A harmonized meta-knowledgebase of clinical interpretations of cancer genomic variants. bioRxiv. 2018:366856. https://doi.org/10.1101/366856.
16. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol. 2017;1:1–16.
17. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018;10(1):25.
18. Huang L, Fernandes H, Zia H, Tavassoli P, Rennert H, Pisapia D, et al. The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. J Am Med Inform Assoc. 2017;24(3):513–9.
19. Patterson SE, Liu R, Statz CM, Durkin D, Lakshminarayana A, Mockus SM. The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. Hum Genomics. 2016;10(1):4.
20. Adamson PC, Houghton PJ, Perilongo G, Pritchard-Jones K. Drug discovery in paediatric oncology: roadblocks to progress. Nat Rev Clin Oncol. 2014;11(12):732.
21. Baylin SB, Ohm JE. Epigenetic gene silencing in cancer–a mechanism for early oncogenic pathway addiction? Nat Rev Cancer. 2006;6(2):107.
22. Hegi ME, Diserens A-C, Gorlia T, Hamou M-F, de Tribolet N, Weller M, et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med. 2005;352(10):997–1003.
23. Aerts S, Haeussler M, Van Vooren S, Griffith OL, Hulpiau P, Jones SJ, et al. Text-mining assisted regulatory annotation. Genome Biol. 2008;9(2):R31.
24. Li G, Ross KE, Arighi CN, Peng Y, Wu CH, Vijay-Shanker K. miRTex: a text mining system for miRNA-gene relation extraction. PLoS Comput Biol. 2015;11(9):e1004391.
25. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45(D1):D362-D368. https://doi.org/10.1093/nar/gkw937.
26. Lever J, Zhao EY, Grewal J, Jones MR, Jones SJ. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat Methods. 2019;16:505-507.
27. Anekalla KR, Courneya J, Fiorini N, Lever J, Muchow M, Busby B. PubRunner: a light-weight framework for updating text mining results. F1000Res. 2017;6.
28. Lever J, Jones S. Painless relation extraction with kindred. BioNLP. 2017;2017:176–83.
29. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(suppl_1):D267–70.
30. Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase. Commun ACM. 2014;57(10):78–85.
31. Bird S. NLTK: the natural language toolkit. In: Proceedings of the coling/acl on interactive presentation sessions. Sydney: Association for Computational Linguistics; 2006. p. 69–72.
32. Davies M. The 385+ million word Corpus of Contemporary American English (1990–2008+): design, architecture, and linguistic insights. Int J Corpus Linguist. 2009;14(2):159–90.
33. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations; 2014. p. 55–60.
34. Honnibal M, Johnson M. An improved non-monotonic transition system for dependency parsing. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Lisbon: Association for Computational Linguistics; 2015. p. 1373–8. Available from: https://aclweb.org/anthology/D/D15/D15-1162.
35. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing; 2019.
36. Björne J, Salakoski T. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task. In: Proceedings of the BioNLP Shared Task 2013 Workshop; 2013. p. 16–25.
37. Bui Q-C, Campos D, van Mulligen E, Kors J. A fast rule-based approach for biomedical event extraction. In: Proceedings of the BioNLP Shared Task 2013 Workshop; 2013. p. 104–8.
38. Chaix E, Dubreucq B, Fatihi A, Valsamou D, Bossy R, Ba M, et al. Overview of the Regulatory Network of Plant Seed Development (SeeDev) Task at the BioNLP Shared Task 2016. In: Proceedings of the 4th BioNLP Shared Task Workshop; 2016. p. 1–11.
39. Lever J, Jones SJ. VERSE: Event and relation extraction in the BioNLP 2016 Shared Task. In: Proceedings of the 4th BioNLP Shared Task Workshop; 2016. p. 42–9.
40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12(Oct):2825–30.
41. Lever J, Jones MR, Danos AM, Krysiak K, Bonakdar M, Grewal J, et al. CIViCmine dataset: Zenodo; 2019. Available from: https://doi.org/10.5281/zenodo.3441694
42. RStudio, Inc. Easy web applications in R. 2013.
43. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2014;43(D1):D805–11.
44. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.
45. Kim J-D, Ohta T, Pyysalo S, Kano Y, Tsujii J. Overview of BioNLP'09 shared task on event extraction. In: Proceedings of the workshop on current trends in biomedical natural language processing. Boulder, Colorado: Shared task: Association for Computational Linguistics; 2009. p. 1–9.
46. Kim J-D, Pyysalo S, Ohta T, Bossy R, Nguyen N, Tsujii J. Overview of BioNLP shared task 2011. In: Proceedings of the BioNLP shared task 2011 workshop. Portland, Oregon: Association for Computational Linguistics; 2011. p. 1–6.
47. Bada M, Eckert M, Evans D, Garcia K, Shipley K, Sitnikov D, et al. Concept annotation in the CRAFT corpus. BMC Bioinformatics. 2012;13(1):161.
48. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. BRAT: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon: Association for Computational Linguistics; 2012. p. 102–7.
49. Peng N, Poon H, Quirk C, Toutanova K, Yih W-T. Cross-Sentence N-ary relation extraction with graph LSTMs. Trans Assoc Comput Linguist. 2017;5:101–15.
50. Leaman R, Islamaj Doğan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics. 2013;29(22):2909–17.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.