JAMIA Open, 3(3), 2020, 332–337
doi: 10.1093/jamiaopen/ooaa028
Advance Access Publication Date: 29 September 2020
Brief Communications

Brief Communications

Predictive article recommendation using natural language processing and machine learning to support evidence updates in domain-specific knowledge graphs

Bhuvan Sharma, Van C. Willis, Claudia S. Huettner, Kirk Beaty, Jane L. Snowdon, Shang Xue, Brett R. South, Gretchen P. Jackson, Dilhan Weeraratne and Vanessa Michelini

IBM Watson Health, Cambridge, Massachusetts, USA
Corresponding Author: Bhuvan Sharma, IBM Watson Health, 75 Binney St Cambridge, MA 0142, USA (sharmabh@us.ibm.com)

Received 15 April 2020; Revised 26 May 2020; Accepted 19 June 2020

ABSTRACT

Objectives: Describe an augmented intelligence approach to facilitate the update of evidence for associations in knowledge graphs.
Methods: New publications are filtered through multiple machine learning study classifiers, and filtered publications are combined with articles already included as evidence in the knowledge graph. The corpus is then subjected to named entity recognition, semantic dictionary mapping, term vector space modeling, pairwise similarity, and focal entity match to identify highly related publications. Subject matter experts review recommended articles to assess inclusion in the knowledge graph; discrepancies are resolved by consensus.
Results: Study classifiers achieved F-scores from 0.88 to 0.94, and similarity thresholds for each study type were determined by experimentation. Our approach reduces human literature review load by 99%, and over the past 12 months, 41% of recommendations were accepted to update the knowledge graph.
Conclusion: Integrated search and recommendation exploiting current evidence in a knowledge graph is useful for reducing human cognition load.

Key words: machine learning, natural language processing, artificial intelligence, precision medicine

LAY SUMMARY

The volume of knowledge in medicine has grown so much in recent decades that manual search, reading, comprehension, and analysis of this much information is impractical. We have developed a service that uses elements of machine learning and text mining to narrow the number of research articles experts need to review to keep curated genomics knowledge bases current. This service uses the content of the current set of curated articles to predict highly related articles that are most likely to be of interest for including in the knowledge base. Articles recommended by the service are reviewed by experts to determine if in fact they should be included in the knowledge base. In use, our service was highly accurate and reduced human literature review load by 99%. Over a year of use, 41% of service recommendations were used to update the knowledge base.

BACKGROUND AND SIGNIFICANCE

Recent decades have produced dramatic growth of knowledge and digitally available data in medicine. For example, 40% of the global science and engineering publications in 2016 were biomedical,[1] and PubMed[2] currently comprises over 30 million citations with >1 million publications indexed per year since 2011. This immense volume of

data, combined with projections for continued exponential growth, makes manual literature searching, reading, comprehension, and analysis impractical.

In evidence-based cancer precision medicine, thousands of associations are continuously evolving. Tracking new evidence for these associations is infeasible without technology. Consequently, text mining bio-curation tools for cancer precision medicine are becoming increasingly important.[3] Various approaches have been developed utilizing natural language processing and machine learning to aid in precision medicine literature curation (reviewed in reference 4). Methods span article recommendation techniques[5] and the classification of documents for prioritized human review[6,7] to automated identification of gene, mutation, and drug entities, and relationships.[8] SME review of the outputs from such automated systems continues to be an important and necessary step for validating and interpreting[9,10] evidence and determining updates in databases.

Watson™ for Genomics (WfG) uses deidentified patient somatic mutation files from sequenced tumor biopsies to categorize DNA alterations related to cancer and list potential therapeutic options targeting these alterations for consideration by clinicians. WfG performs this analysis by using information extracted from the medical literature and knowledge bases, presenting the findings with supporting evidence.[11–17]

The domain knowledge for WfG is captured in a knowledge graph representing entities such as genes, variants, conditions, and drugs linked by associations including "clinical efficacy" or "preclinical efficacy," backed by evidence. Confidence in such associations is based on the evidence in the unstructured text of peer-reviewed publications. As related evidence is published, associations can evolve from "preclinical efficacy" to "clinical efficacy," or the confidence may change in light of new studies. Thousands of entity associations are maintained in this knowledge graph, and with the field rapidly evolving there is high likelihood of new related evidence being published every week. To facilitate timely evidence updates for WfG's knowledge graph we developed an internal tool, IBM® Predictive Article Recommendation Service (IBM PARSe), that utilizes current knowledge graph evidence to search for and recommend related literature for SME review.

OBJECTIVE

IBM PARSe combines the current evidence corpus with new publications and uses a mix of machine learning, entity detection, semantic dictionary mapping, and unstructured text mining techniques to identify highly related candidate publications for SME review. Without this system, SMEs use a labor-intensive approach requiring manual search for thousands of associations. IBM PARSe allows SMEs to spend more time reviewing (rather than searching for) related, high-quality publications. The objective of this manuscript is to describe the IBM PARSe system architecture, report its performance, and provide examples from its real-world application.

METHODS

Knowledge graph

The WfG knowledge graph consists of more than 700 genes, 10M variants, 1200 drugs, and 300 conditions associated with each other through one or more association types, including functional, clinical, preclinical, resistance, and predisposition as described in Table 1.

Table 1. Example knowledge graph data by association type

Association type	Associated entities	Example	Evidence
Functional	Gene and variant, or fusion genes	TTYH3-BRAF fusion gene	• PMID: 31558800 • This study describes the discovery of a novel, highly unusual TTYH3-BRAF fusion gene that contains an almost full-length BRAF protein including the autoinhibitory domain but is still fully pathogenic.
Clinical	Gene, variant, therapy, condition	EGFR L858R mutation, and response to afatinib in NSCLC	• PMID: 23816960 • The L858R mutation in the EGFR gene represents the most common activating mutation in NSCLC and is associated with response to targeted EGFR inhibitors. The manuscript describes the clinical efficacy of a second-generation inhibitor of EGFR (afatinib) in NSCLC patients with this mutation.
Preclinical	Gene, variant, therapy, condition	Cotargeting of JAK2 and HDAC in MPN using ruxolitinib and vorinostat	• PMID: 31227936 • This study addresses an unmet need medical suggesting a combination therapy of ruxolitinib with an HDAC inhibitor (vorinostat) to increase response and duration in MPN patients with a mutation in JAK2. The preclinical evidence presented in this study showed promising results and the association was added to our knowledge graph to recommend this dual therapy should investigational treatment options become available.
Resistance	Gene, variant, therapy	BCL2A1 G101V mutation as a mechanism of resistance to venetoclax	• PMID: 30514704 • This study describes a novel recurrent resistance mechanism to treatment with venetoclax that was detectable in patients several months prior to disease progression. This can inform the physician of imminent patient relapse.
Predisposition	Gene, variant, condition	Gene ETV6, mutation R418G and condition ALL	• PMIDs: 27365488, 25807284 • Predisposition to cancer may be detected by next-generation sequencing. In this study, the R418G missense mutation has been described as a germline mutation in families with familial platelet disorders with a predisposition to leukemia.

Abbreviations: PMID: PubMed ID; NSCLC: nonsmall-cell lung carcinoma; HDAC, histone deacetylase; MPN, myeloproliferative neoplasm; ALL, acute lymphoblastic leukemia.

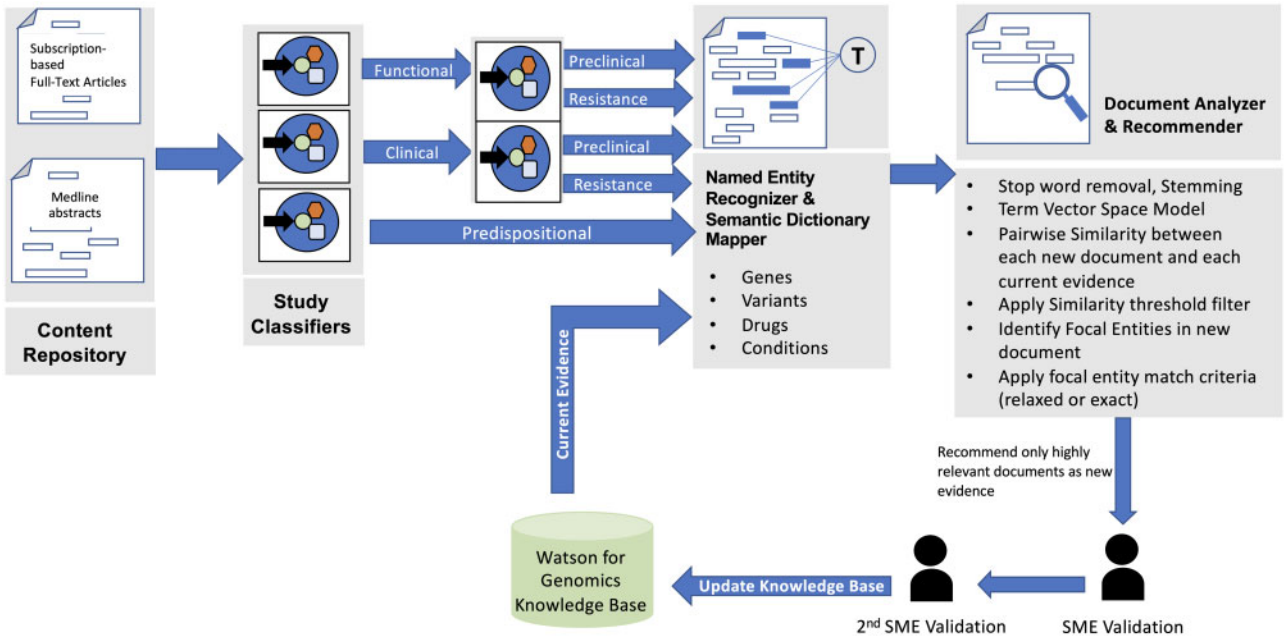

Figure 1. IBM PARSe architecture. SME: subject matter expert.

IBM PARSe

The IBM PARSe system architecture is shown in Figure 1. The key IBM PARSe components are the Content Repository, Study Classifiers, Named Entity Recognizer & Semantic Dictionary Mapper, and Document Analyzer & Recommender. IBM PARSe is configured to run weekly.

Content repository

The Content Repository used by IBM PARSe accesses Medline, PubMed open access, and full-text publications available through subscription from select genomics journals. The initial list of journals was identified using a 90th percentile cutoff on the pareto distribution of count of articles in the WfG knowledge graph available through manual SME updates before implementation of IBM PARSe. This list is refreshed once per year as updates to the knowledge graph and changes in publisher contracts can shift the pareto distribution of journal article counts. All publications (excluding reviews) from this select list of journals in the prior week are subjected to Study Classifiers.

Study classifiers

Distinct binary classifiers were developed for each association type (functional, clinical, preclinical, predisposition, and resistance) using a Naïve Bayes algorithm with unigram and bigram features. During execution, documents outside these study types were discarded; the rest proceeded to the Named Entity Recognizer and Semantic Dictionary Mapper.

Training data consisted of abstracts from articles already included by SMEs in the knowledge graph of WfG. A total of 2200 functional, 300 clinical, 105 predisposition, 150 resistance, and 120 preclinical articles were used for training. While training for a particular study type, articles from all other types were labeled as negative examples, and sampling was performed to ensure an equal number of positive and negative training cases for each classifier.

Named entity recognizer and semantic dictionary mapper

SME-generated dictionaries with synonyms were built for genes, drugs, and conditions using NCI Thesaurus (https://ncit.nci.nih.gov/ncitbrowser/) and terms hand-picked by SMEs. The dictionaries contained about 10 000 synonyms for genes, 12 000 for drugs, and 2000 for conditions. Dictionary-based annotators were used to identify named entities in corpus documents. A rich set of regular expressions was used to identify mentions of variants expressed in the literature such as protein-level substitution, genomic-level substitution, rearrangement, or copy number variation.[18,19] Subsequent to Named Entity Recognition, Semantic Dictionary Mapper replaced all occurrences of named entity synonyms with their base name.

Document analyzer and recommender

The Document Analyzer & Recommender built multiple corpora, one for each association type. Each corpus consisted of all evidence of an association type from the current knowledge graph and all new publications of the same study type as identified by Study Classifiers. A proprietary custom English stop word dictionary was used to remove words that do not add meaning to the text. We built this stop word list from the 400+ most frequent words in our collection of articles after filtering out genomically relevant terms through SME vetting. Words such as “inhibition” and “protein” were filtered out by SME vetting while frequent words such as “of,” “for,” “the,” “done,” “mg,” and “month” were included. This process was followed by stemming to reduce inflectional forms of words to a common base form. These steps preserve semantics by controlling vocabulary size.[20]

A term vector space model[21,22] using term frequency and inverse document frequency would be constructed at run time for each corpus, reflecting the importance of a word in relation to the corpus, followed by a pairwise cosine similarity between each new and prior document. If pairwise similarity exceeded a set threshold, the new document became a candidate to recommend for the association to

which the prior document was used as evidence. All candidate documents were further validated by matching focal entities (gene/variant/drug/condition) to the entities linked by the matched association. If there was a match, the document was recommended for updating the evidence, and if one or more focal entity was different (partial match), the document became a candidate for recommending a new association. Recommended documents, along with identifying information (eg authors, PMID), were sent to SMEs for final review.

SME review

Review of IBM PARSe-recommended documents was conducted with a full-text evaluation by an SME who either accepted or rejected the recommendation. Accepted recommendations were reviewed by a second SME, with discrepancies resolved by consensus with the SME team. Only documents approved by both SMEs were incorporated into the knowledge graph. SME feedback is captured for articles they do not find suitable for the knowledge graph.

Evaluation methodology

Study classifiers were tested for precision, recall, and F-score against a corpus of 500 labeled documents from the WfG knowledge graph.

A retrospective evaluation of IBM PARSe performance was performed on a set of 194 articles used as ground truth for prediction. These articles were published after June 1, 2016 and incorporated by SMEs in the knowledge graph prior to IBM PARSe implementation. Ground truth articles were combined with (1) all articles used as evidence in the knowledge graph but published prior to June 1, 2016 and (2) all articles ($N = 15\,320$) published between June 1 and 30 from select genomics journals to simulate an IBM PARSe monthly run executed on June 30th. This design assumed that all 194 target articles were published in the month prior to the simulation.

RESULTS

Study classifier performance

F-scores for binary study classifiers as measured against a test corpus of 500 labeled documents selected from the WfG knowledge graph ranged from 0.88 to 0.94 (Table 2).

Establishing a similarity threshold

The difference in vocabulary between study types necessitated experiments to identify appropriate similarity thresholds for each study type. All possible document pairs (19 900) from a test corpus of 200 articles used as evidence for clinical associations in the knowledge graph were divided into two groups: one having pairs used as evidence together for an association (Group A, $n = 61$) and another having pairs not used as evidence together for any association (Group B, $n = 19\,839$). False positives (from Group B), false negatives (from Group A), and total error were determined at various cosine similarity thresholds as shown in Figure 2A. Distribution of cosine similarity (where 0 indicates no word match and 1 indicates a duplicate document) for each group is shown in Figure 2B.

The threshold with the lowest total error, 0.21, was selected as an optimal threshold. This same approach was used to identify thresholds for each study type, whose values ranged from 0.20 to 0.25.

Retrospective performance evaluation

A retrospective study was conducted to evaluate if IBM PARSe was able to successfully predict the documents added by SMEs in the WfG knowledge graph after June 1, 2016. The study simulated an IBM PARSe monthly run of the 15 320 documents published between June 1 and 30 in select genomics journals and assumed that all 194 target articles were published in the month prior to the simulation. An F-score of 0.83 (precision = 0.81, recall = 0.85), correctly predicting 166 out of 194 targeted articles, was achieved by the retrospective study.

Examples and ongoing evaluation

To further illustrate IBM PARSe function, examples of knowledge graph changes resulting from IBM PARSe use follow. Per National Comprehensive Cancer Network guidelines, WfG associated crizotinib with treatment of nonsmall-cell lung carcinoma (NSCLC) characterized by MET amplification or exon 14 skipping. IBM PARSe

Table 2. Binary study classifier performance

Study type	Precision	Recall	F-score
Clinical	0.96	0.92	0.94
Functional	0.93	0.91	0.92
Predisposition	0.93	0.89	0.91
Resistance	0.91	0.86	0.88
Preclinical	0.90	0.86	0.88

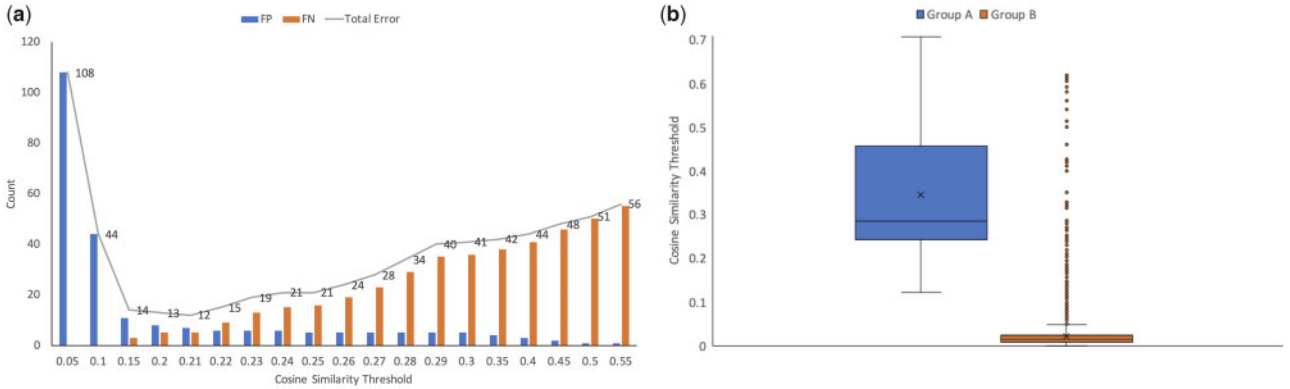

Figure 2. Cosine similarity score threshold evaluation for clinical studies. (A) False positives (FP), false negatives (FN), total error and (B) Distribution of cosine similarity scores for clinical studies at various cosine similarity thresholds

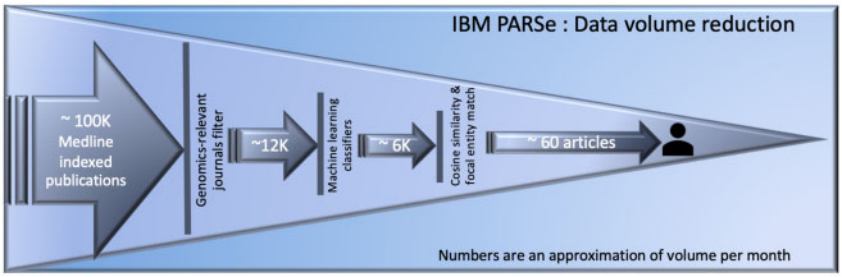

Figure 3. Monthly IBM PARSe data volume reduction.

identified a publication describing a clinical trial of crizotinib for difficult-to-treat NSCLC patients with ALK fusion and concomitant MET amplification.[23] Pyrotinib was already associated in WfG with amplification of ERBB2 (HER2) for the treatment of HER2-positive breast cancer.[24] New studies suggested by IBM PARSe[25,26] provided additional evidence of the efficacy of pyrotinib in this indication. WfG had icotinib marked as futile for EGFR exon 20 mutations in treatment of lung adenocarcinoma.[27] IBM PARSe identified a new study[28] showing efficacy of icotinib in patients harboring EGFR-RAD51 fusion, resulting in a new association and evidence in the knowledge graph. Similarly, another study describing V561M mutation in FGFR1 as conferring resistance to the drug AZD4547[29] was identified from prior evidence[30] in the WfG knowledge graph.

We observe 95% concordance between first and second SME reviews. The average monthly volume reduction with IBM PARSe over 12 months is depicted in Figure 3. Estimated reduction in volume for SME review is 99% on a search space already narrowed by focusing on select journals.

Out of 248 studies recommended by IBM PARSe, 102 (41%) of them have been used by SMEs to update the evidence in WfG's knowledge graph.

DISCUSSION

This manuscript describes IBM PARSe, an automated system that utilizes machine learning classifiers and document similarity to prior evidence, to identify related scientific publications for review by experts to maintain the WfG knowledge graph. IBM PARSe classifiers demonstrated F-scores greater than 0.87, and a retrospective evaluation predicted 166 of 194 articles with a 0.83 F-score. These results are similar to a classification study published recently.[6] Real-world system usage over 12 months resulted in an estimated 99% reduction in volume for SME review, comparable to related work[7] but higher than more stringent classifications for systematic reviews,[31–34] and a yield rate (mean ratio of studies ultimately accepted after review) of 41% which is a significant improvement over 2.94% as reported in pure systematic reviews.[35]

Early error analysis revealed a number of false positives from review studies, so we modified the process to exclude review papers from classifiers. Articles with mention of multiple gene/drug/condition/variants were often recommended for multiple associations, hence we devised a strategy for identifying focal gene/drug/condition/variants to reduce such false positives. We observed that recall for our Study Classifiers is consistently lower than precision, likely because of a lack of differentiating vocabulary in some documents. This occurs more frequently with resistance and preclinical article types which often contain vocabulary and semantics of other study types in the background section of the abstract. Future work includes improving classifier accuracy for articles with confusing semantics using an ensemble of classifiers each built on different parts of the abstract and improving new association recommendation accuracy via integration of models for relation assertion, developed and used in another system within our group. Finally, to help identify false negatives, we plan to incorporate appropriate sampling techniques to achieve a good distribution of articles filtered out by classifiers and similarity thresholds.

CONCLUSION

The results from this study indicate that knowledge graph evidence similarity to new documents is a viable and relevant strategy for automatically narrowing human literature review load. The approach described in this paper demonstrates an integrated solution to support SMEs in updating evidence amidst the unrelenting flow of new studies. Artificial intelligence technology support for SME literature review will be of increasing importance in any science domain with complex networks of entities and associations, particularly in rapidly evolving fields such as cancer genomics and precision medicine.

AUTHORS' CONTRIBUTIONS

Conceptualization: B.S., D.W.; Methodology: B.S., C.H., K.B., S.X., V.M.; Software: B. S., C.H., K.B., S.X., V.M.; Writing—Original Draft: B.S., V.W.; Writing—Review & Editing: B.S., V.W., C.H., K.B., J.S., S.X., B.S., G.J., D.W., V.M.; Visualization: B.S., V.W.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest statement. All authors were employed by IBM when this study was conducted.

REFERENCES

1. Science & Engineering Indicators 2018. https://www.nsf.gov/statistics/2018/nsb20181/ (accessed September 2019).
2. NIH. Secondary. https://www.ncbi.nlm.nih.gov/pubmed (accessed July 2020).
3. Hirschman L, Burns G, Krallinger M, *et al*. Text mining for the biocuration workflow. *Database* 2012; 2012 (0): bas020.
4. Xu J, Yang P, Xue S, *et al*. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. *Hum Genet* 2019; 138 (2): 109–24.
5. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics* 2007; 8 (1): 423.

6. Bao Y, Deng Z, Wang Y, *et al*. Using machine learning and natural language processing to review and classify the medical literature on cancer susceptibility genes. *JCO Clin Cancer Inform* 2019; 3 (3): 1–9.
7. Deng Z, Yin K, Bao Y, *et al*. Validation of a semiautomated natural language processing-based procedure for meta-analysis of cancer susceptibility gene penetrance. *JCO Clin Cancer Inform* 2019; 3 (3): 1–9.
8. Lee K, Kim B, Choi Y, *et al*. Deep learning of mutation-gene-drug relations from the literature. *BMC Bioinformatics* 2018; 19 (1): 21.
9. Madhavan S, Subramaniam S, Brown TD, *et al*. Art and challenges of precision medicine: interpreting and integrating genomic data into clinical practice. *Am Soc Clin Oncol Educ Book* 2018; 38 (38): 546–53.
10. McGraw SA, Garber J, Jänne PA, *et al*. The fuzzy world of precision medicine: deliberations of a precision medicine tumor board. *Per Med* 2017; 14 (1): 37–50.
11. Doerstling S, Winski D, Hintze B, *et al*. Association of mutational profile and human papillomavirus status in patients with head and neck squamous cell carcinoma [abstract]. *J Mol Diagn* 2019; 21 (6): 1204.
12. Frank MO, Koyama T, Rhrissorrakrai K, *et al*. Sequencing and curation strategies for identifying candidate glioblastoma treatments. *BMC Med Genomics* 2019; 12 (1): 56.
13. Itahashi K, Kondo S, Kubo T, *et al*. Evaluating clinical genome sequence analysis by Watson for genomics. *Front Med (Lausanne)* 2018; 5: 305.
14. Kim M, Snowdon J, Weeraratne SD, *et al*. Clinical insights for hematological malignancies from an artificial intelligence decision-support tool. *J Clin Oncol* 2019; 37 (15_suppl): e13023–e23.
15. Patel NM, Michelini VV, Snell JM, *et al*. Enhancing next-generation sequencing-guided cancer care through cognitive computing. *The Oncol* 2018; 23 (2): 179–85.
16. Rhrissorrakrai K, Koyama T, Parida L. Watson for genomics: moving personalized medicine forward. *Trends Cancer* 2016; 2 (8): 392–95.
17. Wrzeszczynski KO, Frank MO, Koyama T, *et al*. Comparing sequencing assays and human-machine analyses in actionable genomics for glioblastoma. *Neurol Genet* 2017; 3 (4): e164.
18. Caporaso JG, Baumgartner WA, Jr., Randolph DA, *et al*. MutationFinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics (Oxf, Engl)* 2007; 23 (14): 1862–65.
19. Doughty E, Kertesz-Farkas A, Bodenreider O, *et al*. Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. *Bioinformatics (Oxf, Engl)* 2011; 27 (3): 408–15.
20. Piantadosi ST. Zipf's word frequency law in natural language: a critical review and future directions. *Psychon Bull Rev* 2014; 21 (5): 1112–30.
21. Salton G, McGill MJ. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, Inc.; 1986.
22. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. *Commun ACM* 1975; 18 (11): 613–20.
23. Chen RL, Zhao J, Zhang XC, *et al*. Crizotinib in advanced non-small-cell lung cancer with concomitant ALK rearrangement and c-Met overexpression. *BMC Cancer* 2018; 18 (1): 1171.
24. Ma F, Li Q, Chen S, *et al*. Phase I study and biomarker analysis of pyrotinib, a novel irreversible Pan-ErbB receptor tyrosine kinase inhibitor, in patients with human epidermal growth factor receptor 2-positive metastatic breast cancer. *J Clin Oncol* 2017; 35 (27): 3105–12.
25. Li Q, Guan X, Chen S, *et al*. Safety, efficacy, and biomarker analysis of pyrotinib in combination with capecitabine in HER2-positive metastatic breast cancer patients: a phase I clinical trial. *Clin Cancer Res* 2019; 25 (17): 5212–20.
26. Ma F, Ouyang Q, Li W, *et al*. Pyrotinib or lapatinib combined with capecitabine in HER2-positive metastatic breast cancer with prior taxanes, anthracyclines, and/or trastuzumab: a randomized, phase II study. *J Clin Oncol* 2019; 37 (29): 2610–19.
27. Wang T, Liu Y, Zhou B, *et al*. Effects of icotinib on early-stage non-small-cell lung cancer as neoadjuvant treatment with different epidermal growth factor receptor phenotypes. *Onco Targets Ther* 2016; 9: 1735–41.
28. Guan Y, Song Z, Li Y, *et al*. Effectiveness of EGFR-TKIs in a patient with lung adenocarcinoma harboring an EGFR-RAD51 fusion. *The Oncol* 2019; 24 (8): 1027–30.
29. Ryan MR, Sohl CD, Luo B, *et al*. The FGFR1 V561M gatekeeper mutation drives AZD4547 resistance through STAT3 activation and EMT. *Mol Cancer Res* 2019; 17 (2): 532–43.
30. Paik PK, Shen R, Berger MF, *et al*. A phase Ib open-label multicenter study of AZD4547 in patients with advanced squamous cell lung cancers. *Clin Cancer Res* 2017; 23 (18): 5366–73.
31. Cohen AM, Hersh WR, Peterson K, *et al*. Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc* 2006; 13 (2): 206–19.
32. Ji X, Ritter A, Yen PY. Using ontology-based semantic similarity to facilitate the article screening process for systematic reviews. *J Biomed Inform* 2017; 69: 33–42..
33. Jonnalagadda S, Petitti D. A new iterative method to reduce workload in systematic review process. *Int J Comput Biol Drug Des* 2013; 6 (1/2): 5–17.
34. Matwin S, Kouznetsov A, Inkpen D, *et al*. A new algorithm for reducing the workload of experts in performing systematic reviews. *J Am Med Inform Assoc* 2010; 17 (4): 446–53.
35. Borah R, Brown AW, Capers PL, *et al*. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* 2017; 7 (2): e012545.