Editorial Commentary

Text-mining in cancer research may help identify effective treatments

Yi-Wen Hsiao, Tzu-Pin Lu

Institute of Epidemiology and Preventive Medicine, Department of Public Health, College of Public Health, National Taiwan University, Taipei
Correspondence to: Tzu-Pin Lu, PhD. Room 518, No. 17, Xu-Zhou Road, 100, Taipei. Email: tplu@ntu.edu.tw.
Provenance: This is an invited article commissioned by the Editorial Office of *Translational Lung Cancer Research*.
Comment on: Lever J, Jones MR, Danos AM, *et al.* Text-mining clinically relevant cancer biomarkers for curation into the CIViC database. Genome Med 2019;11:78.

Submitted Nov 02, 2019. Accepted for publication Dec 12, 2019.
doi: 10.21037/tlcr.2019.12.20
View this article at: http://dx.doi.org/10.21037/tlcr.2019.12.20

Global cancer incidence has rapidly increased over the past few decades. In 2018, around 18.1 million people were diagnosed with cancer worldwide (1). With advanced cancer research and healthcare systems, five-year survival trends are generally increasing. Still, cancer is the second leading causes of death and counts for 9.6 million deaths globally (2). Systematical curations of cancer-related studies are beneficial to researchers as they seek to discover novel treatments for curing cancer.

In the biological aspect, cancer is known to be a type of genetic disease, that is, specific changes to genes may increase one's predisposition to cancer (3). Intensive clinical research has been devoted to identifying biomarkers, such as genes and proteins, that serve as the diagnostic, predictive, prognostic, or even therapeutic molecules for precision medicine (4-6). Both web-lab experiments and bioinformatical approaches have been developed to investigate potential biomarkers for specific cancer types (7), and the rapid accumulation of such scientific literature provides an enormous number of resources to clinical researchers. More than 950,000 publications dated from 2009 to 2019 are available in PubMed by searching abstracts (8). However, it is time-consuming to evaluate such a mountain of biomedical texts manually and challenging to unveil the underlying relationships among these studies or what knowledge is new.

Text-mining technology has emerged as a powerful approach to extract information and knowledge from a wealth of scientific literature. Typically, four main stages are involved in this technology: (I) information retrieval (IR), a process in which biomedical domains such as PubMed are searched for subjects of interest; (II) named entity recognition (NER), which evaluates the occurrence of keywords from collected texts; (III) information extraction (IE), which detects links among the recognized keywords in the text; and (IV) knowledge discovery (KD), which identifies concepts that are new, based on known facts derived from detected relationships (9). Finally, visualization of the extracted information, such as literature networks and word clouds, is also an important step for interpreting the results correctly and efficiently, guiding researchers to generate hypotheses and start follow-up studies (10).

To accurately decode the information from the text based on human Natural Language Processing (NLP), four different levels of language structure must be known: (I) words, which are the units of language; (II) syntax, which presents how to group words into phrases and ultimately sentences; (III) semantics, which captures the logical meaning of sentences; and (IV) pragmatics, which seeks to understand the relationships of sentences in larger context to determine meaning (11). Aside from that, the understanding of biological terminology is also essential to avoid ambiguity and redundancy, enabling effective IE. Many biomedical corpora, such as MeSH (Medical Subject Headings) (12), and databases, such as UniProt, RefSeq (NCBI Reference Sequences) and KEGG (the Kyoto Encyclopedia of Genes and Genomes) (13-15), are established and regularly updated to accelerate the recognition process of biomedical terms and their variations.

IE is an essential text-mining step, for the current data

structure is becoming multifaceted, and the extracted results can directly reflect the topic of the text. Hence, many algorithms have been proposed to make this process more efficient and the output more accurate for further analysis. TF-IDF (term frequency-inverse document frequency) is a simple and efficient approach for calculating a value that represents the word's frequency in both the short text and a given collection of documents in order to determine the importance of a keyword in a short text (16). The algorithm TextRank is a graph-based model, built on an iterated random walk strategy, to rank the importance of words in the text. Additionally, it provides unsupervised text summarization from a set of textual resources (17). For broader texts, topic models, including latent semantic analysis (LSA), latent semantic indexing (LSI) and latent Dirichlet allocation (LDA), are widely used (18-20). Generally, these algorithms rely on probabilistic generative models within the framework of Bayesian statistical inference. These developments greatly accelerate the process of the automatic organization and classification of digitized information.

Many limitations exist in the current algorithms, including in the three described above. First, the weighted relations from multiple text collections remains difficult to correctly interpret; this error may result from incorrect linking of texts due to different biological entities sharing the same name or abbreviation (11). Second, topic models like LDA suffer from some conceptual flaws in practical use; for example, lack of robustness for the number of input texts and free setting of parameters, which causes discrepancies in context. Furthermore, these Bayesian-based approaches ignore the justification for the use of prior knowledge (21). Lastly, cancer involves a complex molecular mechanism; that is, different genotypes or gene expression profiles in the same pathway or network may be triggered even within the same cancer phenotype (22). Analyzing different levels of studies such as molecules, motifs and pathways at the same time to discover new knowledge with a hierarchal view may still not feasible.

In the last section, we use real studies to illustrate how text-mining strategy plays an important role in cancer-related biomarker identification and cancer drug discovery or repositioning (23,24). Previous studies have applied the text-mining approach to systematically discover the relationship between mutations and cancer (25). In addition to discovering a mutation-cancer association, a literature-based mining website extracts sentences from the different sections of the paper to identify whether the gene is a novel driver, oncogene or tumor suppressor (26). This tool can also identify well-supported genes with a dual role as oncogenes or tumor suppressors in different types of cancer. For example, ATF3 is found as an oncogene in breast cancer, but it is a tumor suppressor in prostate cancer. For individual cancer phenotypes, text-mining has been applied to extract the relationship between breast cancer and candidate genes, and candidate association words were found to point to the relationship between breast cancer and related genes using pattern clustering (27). In another example, clinical concepts and genes associated with colorectal cancer were explored through literature and statistical analysis of clinical information and candidates (28). That study confirmed 51 genes from the mined results involving the colorectal cancer pathway. For example, among them, KRAS and CTNNB1 are important oncogenes in colorectal cancer.

Aside from identifying cancer biomarkers, text-mining technology also accelerates the process of cancer drug discovery and repurposing. DrugQuest is a website that applies text-mining technologies and the TextQuest algorithm to mine a publicly accessible database to identify biologically significant terms and group these words based on the textual content (29). Another cancer-related study demonstrated using text-mining and pathway analysis tools is a prospective method for exploring potential drugs targeting the genes or pathways related to cutaneous squamous cell carcinoma (cSCC) (30). In this study, 121 genes relevant to cSCC were identified using a text-mining approach, and 55 drugs were revealed that could target 10 pathways through gene enrichment analysis. Among these identified drugs, 49 have not been tested for this cancer type, suggesting these may be novel, targeted therapies for cSCC treatment. For the drug repositioning study, the importance of all words in cancer-related abstracts in PubMed were represented using word embedding technology. Further, the drug-disease associations were identified, and repurposed drugs were found by classifying candidate genes (31).

In conclusion, advanced text-mining technology may shed light on not only the identification of predictive biomarkers for multiple cancers but also protentional drugs, including novel drug discovery and existing drug repositioning, for effective treatments in the future.

Acknowledgments

Funding: This manuscript was partly supported by grants

from the Ministry of Science and Technology, Taiwan, (MOST-106-2314-B-002-134-MY2 and MOST-104-2314-B-002-107-MY2). The funders had no role in study design, data collection and analysis, the decision to publish, or preparation of the manuscript.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
2. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer 2019;144:1941-53.
3. Haber DA, Fearon ER. The promise of cancer genetics. Lancet 1998;351:SII1-8.
4. Nair M, Singh Sandhu S, K Sharma A. Prognostic and predictive biomarkers in cancer. Curr Cancer Drug Targets 2014;14:477-504.
5. Shen Z. Cancer biomarkers and targeted therapies. Cell Biosci 2013;3:6.
6. Wang Y. Development of cancer diagnostics—from biomarkers to clinical tests. Transl Cancer Res 2015;4:270-9.
7. Goossens N, Nakagawa S, Sun X, et al. Cancer biomarker discovery and validation. Transl Cancer Res 2015;4:256.
8. Goeckenjan G, Sitter H, Thomas M, et al. PubMed Results. Pneumologie 2011;65:e51-75.
9. Spasić I, Livsey J, Keane JA, et al. Text mining of cancer-related information: review of current status and future directions. Int J Med Inform 2014;83:605-23.
10. Fleuren WW, Alkema W. Application of text mining in the biomedical domain. Methods 2015;74:97-106.
11. Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol 2008;9 Suppl 2:S8.
12. Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc 2000;88:265.
13. Apweiler R, Bairoch A, Wu CH, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2004;32:D115-9.
14. Pruitt KD, Tatusova T, Brown GR, et al. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 2012;40:D130-5.
15. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27-30.
16. Ramos J. Using TF-IDF to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. Piscataway, NJ, 2003.
17. Mihalcea R, Tarau P. Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing; 2004.
18. Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse Processes 1998;25:259-84.
19. Papadimitriou CH, Raghavan P, Tamaki H, et al. Latent semantic indexing: A probabilistic analysis. J Comput Syst Sci 2000;61:217-35.
20. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3:993-1022.
21. Gerlach M, Peixoto TP, Altmann EG. A network approach to topic models. Sci Adv 2018;4:eaaq1360.
22. Zhu F, Patumcharoenpol P, Zhang C, et al. Biomedical text mining and its applications in cancer research. J Biomed Inform 2013;46:200-11.
23. Deyati A, Younesi E, Hofmann-Apitius M, et al. Challenges and opportunities for oncology biomarker discovery. Drug Discov Today 2013;18:614-24.
24. Vanhaelen Q, Mamoshina P, Aliper AM, et al. Design of efficient computational workflows for in silico drug repurposing. Drug Discov Today 2017;22:210-22.
25. Kim J, Kim JJ, Lee H. An analysis of disease-gene relationship from Medline abstracts by DigSee. Sci Rep 2017;7:40154.
26. Lever J, Zhao EY, Grewal J, et al. CancerMine: A literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat Methods 2019;16:505.
27. Kawashima K, Bai W, Quan C. editors. Text mining and pattern clustering for relation extraction of breast cancer and related genes. 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed

Computing (SNPD); 2017, IEEE.
28. Liu F, Feng Y, Li Z, et al. Clinic-genomic association mining for colorectal cancer using publicly available datasets. Biomed Res Int 2014;2014:170289.
29. Papanikolaou N, Pavlopoulos GA, Theodosiou T, et al. DrugQuest-a text mining workflow for drug association discovery. BMC Bioinformatics 2016;17:182.
30. Pan Y, Zhang Y, Liu J. Text mining-based drug discovery in cutaneous squamous cell carcinoma. Oncol Rep 2018;40:3830-42.
31. Ngo DL, Yamamoto N, Tran VA, et al. Application of word embedding to drug repositioning. J Biomed Sci Eng 2016;9:7-16.

Cite this article as: Hsiao YW, Lu TP. Text-mining in cancer research may help identify effective treatments. Transl Lung Cancer Res 2019;8(Suppl 4):S460-S463. doi: 10.21037/tlcr.2019.12.20

