Correspondence on "Comparison of literature mining tools for variant classification: Through the lens of 50 RYR1 variants" by Wermers et al.

Chih-Hsuan Wei, Lon Phan, Timothy Hefferon, Melissa Landrum, Heidi L. Rehm, Zhiyong Lu

PII: S1098-3600(24)00142-4

DOI: https://doi.org/10.1016/j.gim.2024.101208

Reference: GIM 101208

To appear in: *Genetics in Medicine*

Received Date: 12 June 2024

Revised Date: 1 July 2024

Accepted Date: 2 July 2024

Please cite this article as: Wei CH, Phan L, Hefferon T, Landrum M, Rehm HL, Lu Z, Correspondence on "Comparison of literature mining tools for variant classification: Through the lens of 50 RYR1 variants" by Wermers et al., *Genetics in Medicine* (2024), doi: https://doi.org/10.1016/j.gim.2024.101208.

Correspondence on "Comparison of literature mining tools for variant classification: Through the lens of 50 RYR1 variants" by Wermers et al.

Chih-Hsuan Wei[1], Lon Phan[1], Timothy Hefferon[1], Melissa Landrum[1], Heidi L. Rehm[2,3] & Zhiyong Lu[1]*

[1]National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD, USA. [2]Program in Medical and Population Genetics, Broad institute of MIT and Harvard, Cambridge, MA, USA. [3]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA.*Corresponding Author: Zhiyong Lu, National Center for Biotechnology Information, Bethesda, MD, USA. Email: zhiyong.lu@nih.gov

To the Editor: We read with interest the recent study by Zara Wermers and colleagues[1] that analyzed the utility of different databases and literature mining tools for the classification of variant pathogenicity according to the American College of Medical Genetics and Genomics and the Association for Molecular Pathology criteria. The authors' comparison offers critical insights over some of the existing tools. However, several aspects of their tool assessment, specifically for LitVar 2.0[2], warrants further discussion.

One important dimension that merits attention is the scope of content accessible by LitVar 2.0. Unlike the other three database/systems in the comparison, LitVar 2.0's processing of the full-text articles is restricted to those within the open-access subset in PubMed Central (PMC) that are made freely available under license terms for text and data mining. As of May 2024, articles in the PMC Open Access Subset only account for 60% of total articles in PMC. LitVar 2.0 does not mine dbSNP or ClinVar as stated in Wermer's study[1].

According to the updated table 1[3] in response to Kiel and Kozaric[4], there are a total of 508 relevant references, of which 163 were successfully retrieved by LitVar 2.0. However, for the 345 that were missed in LitVar 2.0, nearly 90% (304) are articles not available for text and data mining through open access. In other words, LitVar 2.0 retrieved 163 out of the 204 accessible articles, demonstrating a real sensitivity of 0.799. This analysis highlights LitVar 2.0's strong capability in identifying variants within the relevant literature when full text is accessible to its algorithm. This issue was briefly mentioned in the previous study: "LitVar 2.0 is restricted to publicly available literature, which likely limited its sensitivity." However, we believe it is critically important to emphasize the extent to which limited access to full text affects the LitVar system in the original paper.

Additionally, the evaluation conducted between December 22, 2022, and January 12, 2023 was using the beta version of LitVar 2.0, which was prior to its official release in July 2023 that includes the use of advanced tmVar 3.0[5] and GNorm2[6] algorithms for variant and gene recognition, respectively. As a result, the post-evaluation release of LitVar 2.0 offer substantially improved performance over what was reported. The

extracted variants in retrieved articles by official LitVar 2.0 rise by more than 100 compared to what was reported in the original article (for a fair comparison, we limit retrieval results to those papers published before January 2023). The results suggest an underestimation of our tool's full capabilities.

It is also important to note that LitVar 2.0 was designed to search the literature and return all relevant articles when searching for genetic variants, regardless their relevance to the ACMG/AMP/ClinGen classifications. For specifically assisting variant classifications, we utilized the expert-curated results (1,321 PubMed papers in total; 508 positive ones) by Wermers and colleagues to fine-tune a language model LinkBERT[7] and performed standard five-fold cross validation evaluation. Our results show that such an AI model can effectively rank relevant articles higher in the result set with an accuracy of 82.7% in F1 score. Given these promising results, we believe additional expert-curated data could further enhance the accuracy and robustness of such a language model that can be readily used in LitVar and other similar AI systems for assisting variant curation and classification. Hence, we call for more future evaluation studies like this one that provide invaluable open data beyond a single gene.

Looking ahead, the landscape of scientific publishing is also undergoing significant changes, with a growing emphasis on open access. For instance, more than 95% of new articles published in 2023 are available freely for text and data mining. Given this trend, LitVar 2.0 is poised to capture significantly more relevant papers containing genetic variants in the future.

In summary, we anticipate a continued evolution of AI-powered literature mining tools like LitVar 2.0, emphasizing the impact of open-access policies and the potential of utilizing computational techniques together with expert curation to help researchers stay abreast of the dynamic nature of genomic research and variant classification.

Conflict of Interest Statement: This research is supported by the NIH Intramural Research Program, National Library of Medicine. The authors declare no conflict of interest.

REFERENCE

1. Wermers, Z. *et al.* Comparison of literature mining tools for variant classification: Through the lens of 50 RYR1 variants. *Genetics in Medicine*, 101157 (2024).
2. Allot, A. *et al.* Tracking genetic variants in the biomedical literature using LitVar 2.0. *Nature Genetics* **55**, 901-903 (2023).
3. Wermers, Z. *et al.* Response to Kiel and Kozaric regarding “Comparison of Literature Mining Tools for Variant Classification: Through the Lens of Fifty RYR1 Variants” by Wermers et al. *Genetics in Medicin*, 101162 (2024).
4. Kiel, M.J. & Kozaric, A. Correspondence on "Comparison of literature mining tools for variant classification: Through the lens of 50 RYR1 variants" by Wermers et al. *Genetics in Medicine*, 101161 (2024).
5. Wei, C.-H., Allot, A., Riehle, K., Milosavljevic, A. & Lu, Z. tmVar 3.0: an improved variant concept recognition and normalization tool. *Bioinformatics* **38**, 4449–4451 (2022).

6. Wei, C.-H., Luo, L., Islamaj, R., Lai, P.-T. & Lu, Z. GNorm2: an improved gene name recognition and normalization system. *Bioinformatics* **39**, btad599 (2023).
7. Yasunaga, M., Leskovec, J. & Liang, P. LinkBERT: Pretraining Language Models with Document Links. *Association for Computational Linguistics*, 8003-8016 (2022).