Comparison of Literature Mining Tools for Variant Classification: Through the Lens of Fifty *RYR1* Variants

Zara Wermers, Seeley Yoo, Bailey Radenbaugh, Amber Douglass, Leslie G. Biesecker, Jennifer J. Johnston

PII: S1098-3600(24)00016-9

DOI: https://doi.org/10.1016/j.gim.2024.101083

Reference: GIM 101083

To appear in: *Genetics in Medicine*

Received Date: 5 October 2023

Revised Date: 19 January 2024

Accepted Date: 22 January 2024

Please cite this article as: Wermers Z, Yoo S, Radenbaugh B, Douglass A, Biesecker LG, Johnston JJ, Comparison of Literature Mining Tools for Variant Classification: Through the Lens of Fifty *RYR1* Variants, *Genetics in Medicine* (2024), doi: https://doi.org/10.1016/j.gim.2024.101083.

Comparison of Literature Mining Tools for Variant Classification: Through the Lens of Fifty *RYR1* Variants

Authors: Zara Wermers, Seeley Yoo, Bailey Radenbaugh, Amber Douglass, Leslie G Biesecker, Jennifer J Johnston

Center for Precision Health Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

Corresponding author:

Jennifer Johnston

Building 50 Room 5139

Bethesda, MD 20892

Tel: 301-594-3981

Email: jjohnston@mail.nih.gov

Abstract

Purpose: The American College of Medical Genetics and Genomics and the Association for Molecular Pathology have outlined a schema that allows for systematic classification of variant pathogenicity. While gnomAD is generally accepted as a reliable source of population frequency data and ClinGen has provided guidance on the utility of specific bioinformatic predictors, there is not a consensus source for identifying publications relevant to a variant. Multiple tools are available to aid in the identification of relevant variant literature including manually curated databases and literature search engines. We set out to determine the utility of four literature mining tools used for ascertainment to inform the discussion of the use of these tools.

Methods: Four literature mining tools including the Human Gene Mutation Database, Mastermind®, ClinVar, and LitVar 2.0 were used to identify relevant variant literature for 50 *RYR1* variants. Sensitivity and precision were determined for each tool.

Results: Sensitivity among the four tools ranged from 0.332 to 0.687. Precision ranged from 0.389 to 0.906. No single tool retrieved all relevant publications.

Conclusion: At the current time, the use of multiple tools is necessary to completely identify the literature relevant to curate a variant.

Keywords: biocuration, literature mining, ACMG guidelines, variant classification, RYR1

Introduction

The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) have developed a framework for pathogenicity classification in Mendelian disorders [1] that combines disease, gene, and variant-specific evidence to classify the pathogenicity of a variant. Refinements to this framework have been proposed by the Sequence Variant Interpretation Group from within ClinGen [2-6]. Much of the relevant variant information, including functional and case data, can be collected from the literature.

When classifying variant pathogenicity, it is important to access relevant literature that informs the ACMG/AMP/ClinGen criteria. The goal is to fully inform the criteria with all relevant information while minimizing the number of manuscripts that need to be assessed. The ideal literature mining tool would identify all relevant primary literature without the return of secondary reporting of the same data. This is as critical as it is challenging, as it can be difficult to determine if one's literature ascertainment is complete. The variant analyst does not want to miss a publication as it may provide important or contradictory data regarding the pathogenicity of a variant.

To address this, we have compared four tools that can be used in the identification of relevant variant literature. The Human Gene Mutation Database (HGMD® Professional), Mastermind® Genomic Intelligence Platform, ClinVar, and LitVar 2.0. HGMD Professional [7] is a curated fee-for-service database that provides variant-centric information including references relevant to variant classification. Variants can be searched by gene or HGVS variant nomenclature. While a public version of HGMD is available (http://www.hgmd.org), it only

shows variants three years or older and is inadequate for complete ascertainment of the literature. Mastermind [8] has multiple options for search terms that are not limited to gene and HGVS nomenclature of a variant. Mastermind normalizes variant input and recognizes standard HGVS notations, reference SNP cluster ID numbers (RSID), cDNA, genes, abbreviated proteins, genomic positions and text excerpts as input. A Basic edition (free) and Professional edition (fee based) of Mastermind are available. The free version of Mastermind restricts the search input to genes and variants, and data found in supplemental tables are not accessible using the free edition. The Professional version of the tool provides additional search capabilities including phenotypes and CNVS, and functionalities including refining searches by ACMG criteria.

ClinVar [9] is a public database that aims to provide variant classifications provided by experts in the field including research laboratories, clinical laboratories and expert panels. Literature submitted with variant classifications is presented on the variant page, the literature in ClinVar is not centrally curated. Recently, ClinVar has added a literature mining tool, LitVar 2.0 [10]. LitVar 2.0 mines PubMed, PubMed Central (PMC) Open Access Subset, dbSNP, and ClinVar. LitVar 2.0 uses the text-mining application tmVar3.0 [11] to normalize variant queries at the level of cDNA or protein into dbSNP RSID numbers that are used to index abstracts from PubMed and full-text publications from PMC Open Access. LitVar 2.0 can access supplemental content.

As members of the ClinGen Malignant Hyperthermia Susceptibility (MHS) Variant Curation Expert Panel (VCEP), we have classified *RYR1* variants according to revised ACMG/AMP/ClinGen guidelines [12, 13] for *RYR1*-related malignant hyperthermia susceptibility.

RYR1 is one of 81 genes recommended for return of secondary findings by the ACMG (v3.2) [14, 15] and routinely assessed in clinical sequencing data. For this study, we assessed the performance of the HGMD (Professional edition), Mastermind (Professional edition), ClinVar, and LitVar 2.0 to assess their ability to identify relevant literature that informs the ACMG/AMP/ClinGen criteria for *RYR1* (HGNC:10483) variants while minimizing identification of uninformative literature or secondary reports of information.

Methods:

Variant Selection and Literature Mining:

Fifty *RYR1* variants were selected for this analysis including 12 variants of uncertain significance (VUS) and 38 pathogenic and likely pathogenic (P/LP) variants as assessed according to the ACMG/AMP/ClinGen criteria as defined by the ClinGen MHS VCEP [12]. Literature mining for each variant was performed using HGMD® Professional (https://my.qiagendigitalinsights.com/bbp/view/hgmd/pro/start.php, Qiagen, Germantown, MD), Mastermind Genomic Intelligence Platform (https://www.genomenon.com/mastermind, Genomenon®, Ann Arbor, MI), ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/), and LitVar 2.0 (https://www.ncbi.nlm.nih.gov/research/litvar2/) as detailed below. We then manually reviewed each identified publication to assess it for information that informed the ACMG/AMP/ClinGen criteria [1, 12, 13]. Database queries were performed between 12/22/2022 and 1/12/2023. All identified publications are presented in Table S1.

HGMD was searched using gene symbol (*RYR1*) and cDNA nomenclature based on reference NM_000540.3. Mastermind and LitVar 2.0 were searched using gene symbol (*RYR1*) and

protein nomenclature based on reference NP_000531.2. When two cDNA variants result in the same protein change papers relevant to both variants would be expected to be returned using protein nomenclature, this is relevant for variant c.12700G>C (c.12700G>T), p.(Val4234Leu) in this study. Variant pages in ClinVar were accessed through Alamut Visual 2.15 (Sophia Genetics, Boston, MA) based on genomic position (GRCh37). The ClinVar section "Citations for this variant" was used for literature identification. Importantly, while the MHS VCEP had submitted classifications for these fifty variants to ClinVar the VCEP did not submit detailed citations.

The content of each reference was analyzed to determine if the publication was a primary or secondary source. Sources that presented novel data specific to the variant in question were defined as primary. A paper that only referenced prior work related to the variant or where the variant was identified in a large genomics screen unrelated to malignant hyperthermia or myopathy was defined as secondary. All references were further reviewed for variant-specific information that informed the ACMG/AMP/ClinGen criteria (as noted in Table S1). Primary references that presented information that informed the ACMG/AMP/ClinGen criteria were defined as "relevant". For each tool, the total number of references and the number of relevant references returned was calculated.

Sensitivity and Precision:

We calculated sensitivity and precision for the four literature mining tools with reference to the number of papers we deemed relevant for ACMG/AMP/ClinGen classifications (Table S2). Sensitivity was calculated by dividing the number of relevant papers returned by a tool by the union of the relevant papers returned by all tools. Precision was calculated by

dividing the number of relevant references returned by a tool by the number of references (relevant and not relevant) returned by that same tool.

Novel Reference Analysis

References that were identified by a single tool were defined as "novel". We quantified the number of novel references that each tool returned and determined the number of novel references defined as relevant.

Results:

Primary and Relevant Reference Identification

HGMD: We had an overall return of 171 publications across all variants, 156 (91.2%) of these references were primary. A total of 15 references were novel. Of these novel references, 13 (86.7%) were relevant to variant classification.

Mastermind: We had an overall return of 608 publications across all variants, 265 (43.6%) of these references were primary. A total of 343 references were novel. Of these novel references, 82 (23.9%) were relevant to variant classification.

ClinVar: We had an overall return of 324 publications across all variants, 193 (59.6%) of these references were primary. A total of 135 references were novel. Of these novel references, 37 (27.4%) were relevant to variant classification.

LitVar: We had an overall return of 321 publications across all variants, 126 (39.3%) of these references were primary. Of these primary publications, 151 references were novel. Of these novel references, 19 (12.6%) were relevant to variant classification.

Sensitivity and Precision

Sensitivity demonstrated the ability of each tool to identify the complete set of papers relevant for variant pathogenicity classification (defined as the union of relevant papers from all four tools). Mastermind had the highest sensitivity, followed by ClinVar, HGMD, and LitVar 2.0 (Table S2). Precision measures the fraction of publications that were returned by a tool that were relevant for variant assessment, tools with a higher precision returning a smaller fraction of extraneous papers. HGMD had the highest precision at 90.6%. HGMD returned the fewest number of publications (Figure 1), but more of these papers were relevant to variant classification as compared to the other tools. Mastermind and LitVar 2.0 had the lowest precision with less than half of the identified references considered relevant for ACMG/AMP/ClinGen variant classification. Variant-specific metrics including precision and sensitivity are presented in Table S2.

Discussion

Literature curation is a critical step in identifying data to inform variant classification. For the purpose of informing ACMG/AMP/ClinGen pathogenicity classifications, optimal literature retrieval requires high sensitivity. However, identification of literature that is not a primary source of relevant information can waste the analyst's time. We determined the sensitivity and precision of four literature mining tools by evaluating the identification of literature than could inform the ACMG/AMP/ClinGen criteria for 50 *RYR1* variants.

Maximal retrieval of the literature depends on proper normalization of variant nomenclature [16]. Ideally, variant descriptors of cDNA, genomic position, and protein could all be used as input by search engines. For proteins with more than one cDNA, or proteins with

historical nomenclature that does not follow current guidelines (proteins with signal peptides, for example *APOB*), more complicated search parameters may be required for maximal sensitivity. As well, the search engine needs to access variant data whether it is in the main text, tables, figures, or supplemental information. Variant nomenclature normalization is an attribute of both Mastermind and LitVar 2.0. For the 50 *RYR1* variants evaluated here, Mastermind identified 608 references while LitVar 2.0 identified 321. Additionally, Mastermind identified references for 49/50 variants while LitVar 2.0 only identified references for 34/50 variants. LitVar 2.0 is restricted to publicly available literature which likely limited its sensitivity.

While HGMD has an automated search as part of its algorithm, it presents a curated set of papers related to the pathogenicity of variants rather than a complete set of papers and therefore would be expected to return fewer references (171 references for this set of variants) compared to a non-curated search engine. Likewise, ClinVar presents references submitted by clinicians and scientists as relevant to variant interpretation (324 references for this set of variants). Mastermind outperformed both tools in the absolute number of references returned. Another way to measure the success of a literature search, however, is to determine the relevance of the papers that are returned. HGMD and ClinVar are manually curated, and it might be expected that the additional curation would result in a higher percentage of publications relevant for the ACMG/AMP/ClinGen classification criteria (higher precision). As expected HGMD and ClinVar had higher precision (90.6% and 59.3% respectively) as compared to Mastermind and LitVar 2.0 (42.6% and 38.9%). For variants with ample literature, it is logical that HGMD and ClinVar can be first pass sources. When additional information is required to fully classify a variant Mastermind and LitVar 2.0 can allow for more complete ascertainment of

the literature with the tradeoff of identification of papers not relevant to pathogenicity classification.

Conclusion

We have provided a comparison of HGMD, Mastermind, ClinVar, and LitVar 2.0 specific to their utility for ACMG/AMP/ClinGen classification of genetic variants in *RYR1*. For this set of *RYR1* variants, Mastermind had the highest recall overall, followed by ClinVar, LitVar 2.0 and HGMD. HGMD and ClinVar had the highest precision, likely due to manual curation. No single tool retrieved all relevant publications. Unfortunately, at the current time, the use of multiple tools is necessary to completely identify the literature relevant to curate a variant. We are aware that each of these resources are dynamic. The automated ones (LitVar 2.0, and Mastermind) are continually updating and refining their algorithms to improve their performance. HGMD is constantly recurating variants to represent current knowledge more completely. However, we do not anticipate in the near future that there will be a single tool or resource that serves the need of the variant analyst for the complete and specific identification of data relevant to a variant. For the foreseeable future, analysts will need to rely on multiple tools for complete ascertainment and must reckon with the low precision, which increases the time required to classify a variant.

Data Availability

This paper does not contain primary research data; data developed during this study is included in supplemental tables.

Funding Statement

This research was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health, HG200388-10 and HG200387-10. This publication represents the views of the authors and does not necessarily represent the position of the National Institutes of Health or the National Human Genome Research Institute.

Author Contributions

BR and AD: Data curation. SY: Data curation, Formal analysis, Writing- Original draft preparation. ZW: Data curation, Formal analysis, Writing- Reviewing and Editing. JJJ and LGB: Conceptualization, Writing- Reviewing and Editing. All authors read and approved the final manuscript.

Ethics Declaration

This study did not involve human subjects or any individual-level data. All data used were publicly available.

Conflict of Interest

LGB is a member of the Illumina Medical Ethics Advisory committee and receives research support from Merck, Inc. ZW, SY, BR, AD and JJJ declare no conflicts of interest.

Supplemental Files

Table S1. Complete list of publications identified for 50 RYR1 variants. Literature mining tools that identified publication are noted as well as if the publication was considered primary, secondary and/or relevant to ACMG variant classification. ACMG category for relevant information identified in publication noted.

Table S2. Variant specific metrics for different literature mining tools.

References

1. Richards, S, Aziz, N, Bale, S, et al., Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*, 2015. **17**(5): p. 405-24. https://doi.org/10.1038/gim.2015.30
2. Abou Tayoun, AN, Pesaran, T, DiStefano, MT, et al., Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum Mutat*, 2018. **39**(11): p. 1517-1524. https://doi.org/10.1002/humu.23626
3. Brnich, SE, Abou Tayoun, AN, Couch, FJ, et al., Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med*, 2019. **12**(1): p. 3. https://doi.org/10.1186/s13073-019-0690-2
4. Ghosh, R, Harrison, SM, Rehm, HL, et al., Updated recommendation for the benign stand-alone ACMG/AMP criterion. *Hum Mutat*, 2018. **39**(11): p. 1525-1530. https://doi.org/10.1002/humu.23642
5. Pejaver, V, Byrne, AB, Feng, BJ, et al., Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet*, 2022. **109**(12): p. 2163-2177. https://doi.org/10.1016/j.ajhg.2022.10.013
6. Walker, LC, Hoya, M, Wiggins, GAR, et al., Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: Recommendations from the ClinGen SVI Splicing Subgroup. *Am J Hum Genet*, 2023. **110**(7): p. 1046-1067. https://doi.org/10.1016/j.ajhg.2023.06.002
7. Stenson, PD, Mort, M, Ball, EV, et al., The Human Gene Mutation Database: 2008 update. *Genome Med*, 2009. **1**(1): p. 13. https://doi.org/10.1186/gm13
8. Chunn, LM, Nefcy, DC, Scouten, RW, et al., Mastermind: A Comprehensive Genomic Association Search Engine for Empirical Evidence Curation and Genetic Variant Interpretation. *Front Genet*, 2020. **11**: p. 577152. https://doi.org/10.3389/fgene.2020.577152
9. Landrum, MJ, Lee, JM, Riley, GR, et al., ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*, 2014. **42**(Database issue): p. D980-5. https://doi.org/10.1093/nar/gkt1113
10. Allot, A, Peng, Y, Wei, CH, et al., LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. *Nucleic Acids Res*, 2018. **46**(W1): p. W530-W536. https://doi.org/10.1093/nar/gky355
11. Wei, CH, Allot, A, Riehle, K, et al., tmVar 3.0: an improved variant concept recognition and normalization tool. *Bioinformatics*, 2022. **38**(18): p. 4449-4451. https://doi.org/10.1093/bioinformatics/btac537
12. Johnston, JJ, Dirksen, RT, Girard, T, et al., Variant curation expert panel recommendations for RYR1 pathogenicity classifications in malignant hyperthermia susceptibility. *Genet Med*, 2021. **23**(7): p. 1288-1295. https://doi.org/10.1038/s41436-021-01125-w

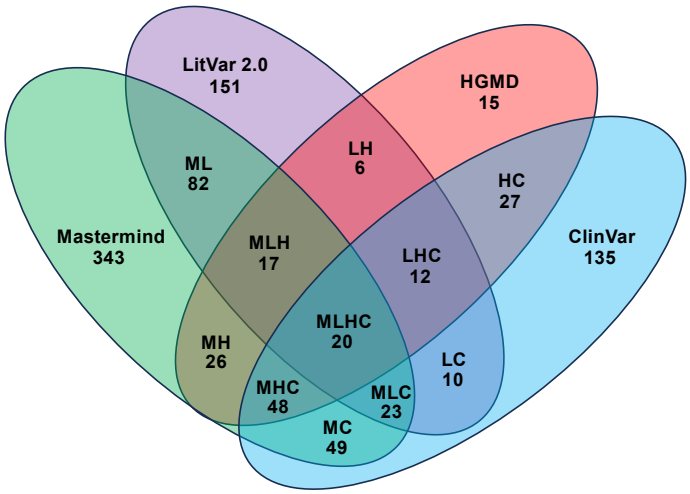
13. Johnston, JJ, Dirksen, RT, Girard, T, et al., Updated variant curation expert panel criteria and pathogenicity classifications for 251 variants for RYR1-related malignant hyperthermia susceptibility*. Hum Mol Genet*, 2022. **31**(23): p. 4087-4093. https://doi.org/10.1093/hmg/ddac145
14. Miller, DT, Lee, K, Abul-Husn, NS, et al., ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG)*. Genet Med*, 2023. **25**(8): p. 100866. https://doi.org/10.1016/j.gim.2023.100866
15. Green, RC, Berg, JS, Grody, WW, et al., ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing*. Genet Med*, 2013. **15**(7): p. 565-74. https://doi.org/10.1038/gim.2013.73
16. Lyons, EL, Watson, D, Alodadi, MS, et al., Rare disease variant curation from literature: assessing gaps with creatine transport deficiency in focus*. BMC Genomics*, 2023. **24**(1): p. 460. https://doi.org/10.1186/s12864-023-09561-5

Figure 1. Number of publications returned by HGMD® Professional (H), Mastermind® (M), ClinVar (C), and LitVar 2.0 (L) for all 50 variants. Publications returned by multiple tools are located within overlap regions, for overlap areas tools are denoted by codes as noted.

Table 1. Metrics for individual tools including number of papers returned, precision and sensitivity. A complete list of references can be found in Table S1.

Tool	Total Papers[a]	Primary Papers	Secondary Papers	Relevant Papers	Total Novel	Relevant Novel	Precision	Sensitivity
HGMD	171	156	14	155	15	13	0.906	0.411
Mastermind	608	265	334	259	343	82	0.426	0.687
ClinVar	324	193	99	192	135	37	0.593	0.509
LitVar 2.0	321	126	190	125	151	19	0.389	0.332

[a]Total number of papers includes papers that were returned by a tool that did not mention the variant, these were not counted for other categories.



Conflict of Interest

LGB is a member of the Illumina Medical Ethics Advisory committee and receives research support from Merck, Inc. ZW, SY, BR, AD and JJJ declare no conflicts of interest.