DAVI: a Dataset for Automatic Variant Interpretation

Francesca Longhin[1[0009-0007-2833-3543]], Alessandro Guazzo[1[0000-0001-5155-2567]], Enrico Longato[1[0000-0001-5940-645X]], Nicola Ferro[1[0000-0001-9219-6239]] and Barbara Di Camillo[1,2[0000-0001-8415-4688]]

[1] University of Padova, Department of Information Engineering, Padova, 35131, Italy
[2] University of Padova, Department of Comparative Biomedicine and Food Science, Legnaro (PD), 35020, Italy
barbara.dicamillo@unipd.it

Abstract. The analysis of an individual's genetic material may uncover genetic variants, which can be classified as disease-causing (pathogenic) or benign. Identifying pathogenic variants among millions of variants relies on the research of evidence in support of or against variant pathogenicity, a process regulated by the American College of Molecular Genetics (ACMG) guidelines, which leverages data from the scientific literature. Despite recent improvements towards automation, searching shreds of evidence for pathogenicity in the literature still requires manual curation, a time-consuming process, due to the ever-growing number of published papers.

In this work, we built DAVI (Dataset for Automatic Variant Interpretation), a reliable, manually curated dataset comprising articles both containing (positive) and not containing (negative) evidence activating two opposing ACGM criteria, namely PS3 and BS3, for a pool of 41 variants. Moreover, we demonstrated that DAVI can be used to train a predictive model that automatically identifies positive *(variant, article)* associations.

DAVI contains 311 *(variant, article)* pairs: 154 positive and 157 negative associations. We used three different text representation models combined with a logistic regression to efficiently identify positive associations, with an F1-score of 0.84. The model's performance constitutes a clear proof of concept for automatic PS3/BS3 evidence identification. DAVI represents a useful resource to train further models.

Keywords: Clinical Genetics, Variant Interpretation, Natural Language Processing.

1 Introduction

Deoxyribonucleic acid, more commonly known as DNA, is a complex molecule that stores the genetic information needed for the development and functioning of an organism. The DNA molecule is contained in each of an organism's cells, which are the basic biological building blocks that provide structure to its tissues. The DNA is composed of a series of four different smaller molecules, called nucleotides: adenine ("A"), thymine ("T"), guanine ("G"), and cytosine ("C"). In 2003, with the completion of the

Human Genome Project [1], the first human genome, i.e. the four-letter sequence encoding a person's DNA, was determined through a laboratory technique called sequencing. Thereafter, sequencing technologies have become more and more sophisticated and widely accessible, enabling the resolution of thousands of genomes and the detection of small differences among them, known as genetic *variants*. Variants can be inherited from a parent or occur during a person's lifetime. Identification of genetic variants, which consists in assessing the variants' positions in the genome and affected nucleotides, is crucial as variants are not only responsible for differences in appearance among individuals of the same species, but also associated to their health status. For example, some variants are located within genes, which are chunks of nucleotides in the genome carrying instructions for the synthesis of proteins, complex molecules that play many critical roles (signalling, structural support, nutrients storage) in the organism. Alterations in gene sequences can result in the production of inactive proteins, increasing an individual's susceptibility to a certain disease (*pathogenic variants*), or they can have no impact on the function of the gene/protein (*benign variants*).

Recently, sequencing technologies have been increasingly used for personalised healthcare, as the identification of a person's genetic variants and the assessment of their benignity/pathogenicity allow clinicians to provide suitable therapies to patients [2]. However, a correct variant benignity/pathogenicity assessment, a process also known as variant interpretation, does not rely only on information about variant position, affected gene, and affected protein, but it requires the clinician to perform a complete *variant annotation*, gathering all relevant evidence about the nature and the effect of the variant from biological databases and the scientific literature [3]. To be meaningful, variant annotation should follow the recommended guidelines defined by the American College of Medical Genetics and Genomics (ACMG) in 2015 [4]. These guidelines contain 28 criteria, each identified by an *evidence code* representing evidence in support of variant benignity or pathogenicity. Some criteria are applied to a specific variant based on the evidence contained in databases (variant frequencies in healthy reference populations, prediction scores based on the probability of damaging protein structure, etc.), while others require information contained in the literature (results of experimental tests carried out on the variant, disease-association studies, etc.). An example of two opposing ACMG criteria which are applied based on information contained in the literature are PS3 and BS3. These criteria are alternatively assigned to a specific variant when an experimental test, in which the variant is injected in the DNA of an animal (in vivo) or cell culture (in vitro), proved that the variant has a damaging or null effect on protein function, respectively.

The task of mining for evidence of a variant's benignity/pathogenicity from the literature, a process known as *manual curation*, is extremely complex and time-consuming. It requires highly qualified curators who scan the continuously growing biomedical literature in the quest for the required evidence. This typically happens by querying a literature search engine with "`Variant_Name` AND `Gene_Name`", where `Variant_Name` is a variant's identification code and `Gene_Name` is the symbol of the gene where the variant is located. Then, curators have to proceed by reading all retrieved articles, looking for relevant information in figures, tables, sentences that contain the variant's identification code, and those nearby (e.g. typically only the previous

and next sentences) [5]. When curators find a relevant article, they assign the specific ACMG criterion to the *(variant, article)* pair.

Considering that the number of biomedical publications that contain genetic variants grows day by day and that the research community uses multiple forms to refer to genetic variants (variant synonyms), it is increasingly difficult to have enough expert curators to read all available publications and to find all relevant information about each discovered variant [6]. As a result, currently, there is the lack of a complete and constantly updated database containing *(variant, article)* associations curated following the ACMG guidelines for each variant and for each criterion. The only resource of this kind is ClinGen [7]: it contains expert-curated assertions regarding variant pathogenicity, as well as supporting evidence summaries, and it is used for consultation in clinical decision-making. However, the number of variants annotated in ClinGen is very limited and the information used for variant interpretation is partial: indeed, most of the time, ClinGen curators make their statements on variant pathogenicity when they think they have collected enough evidence from a restricted number of analysed publications, possibly missing lots of useful information contained in the remaining overlooked papers.

Given the abovementioned considerations, there is a need for a tool able to automatically identify the evidence needed for an ACMG-compliant variant interpretation, which could be easily applied to any variant at any time. Indeed, several tools have been proposed to automatically collect variant annotations [8], [9], but none of them performs a comprehensive screening of the extensive and ever-growing literature, nor automatically extract the information needed for the activation of ACMG criteria.

The first aim of this work is to build a high quality manually curated dataset of articles that either activate, for a specific variant, one of two opposing ACMG criteria, namely PS3 and BS3, or that activate neither. This dataset, named DAVI (Dataset for Automatic Variant Interpretation), will be available on Zenodo. Besides being a useful resource by itself, will be the basis for developing automatic methods for variant annotation. To the best of our knowledge, this is the first type of such dataset available for research. The second aim is to perform a preliminary exploratory analysis of DAVI via the development of an automatic, machine-learning-based predictive model to identify *(variant, article)* pairs where either PS3 or BS3 are activated. This can be thought as a first step before a second classification step to distinguish between articles that activate PS3 vs. those which activate BS3.

The paper is organized as follows: Section 2 describes the methodology used to create DAVI; Section 3 describes the implementation of a predictive model, trained on DAVI, that automatically performs identification of *(variant, article)* associations where either PS3 or BS3 are activated; finally, Section 4 draws some conclusions and outlooks for future work.

2 Dataset Construction

In the following, we call *positive articles* those articles which activate either the PS3 or the BS3 criterion, while we call *negative articles* those activating neither of them. Typically, in positive articles, the result of the experimental test is summarised in one or

more sentences (*positive sentences*), which trigger the activation of the PS3 or BS3 criteria. Positive sentences contain the functional comparison between two analysed models, in vivo or in vitro, one carrying the variant (mutant) and the other carrying the non-mutated sequence of DNA (wild-type). All other sentences activate neither PS3 nor BS3 (*negative sentences*).

2.1 Article Retrieval

To build DAVI, we started from downloading the ClinGen Evidence Repository, whose rows contain information about 4980 curated genetic variants, distributed across 88 genes of interest. Each curated variant is reported with its Human Genome Variation Society (HGVS) [10] standard nomenclature. According to HGVS, variants (e.g., NM_000277.2(PAH): c.472C>T (p.Arg158Trp)) are unambiguously described at the DNA level through an accepted reference DNA sequence (e.g., NM_000277.2), also called transcript, which is located in a gene (e.g., PAH); the position of the variant (e.g., 472), calculated with respect to that specific transcript; and replaced and replacing nucleotides (e.g., 472C>T means cytosine becomes thymine in position 472). In addition, variants can be described at the protein level, specifying the position of the variant in the amino acids sequence (e.g., 158), replaced and replacing amino acids (e.g., Arg158Trp means arginine becomes tryptophan in position 158).

Each variant is associated to a list of ACMG evidence codes assigned by ClinGen manual curators on the basis of information contained in databases (e.g., ACMG criterion applied: PM2, source of evidence: ExAC [11]) or one or more scientific papers, identified by PubMed [12] identification codes (PMIDs) (e.g., ACMG criterion applied: PS3, source of evidence: PMID:24401910).

We focused only on variant curations where either PS3 or BS3 evidence codes were assigned, given that the evidence needed for these assignments is often contained in articles' plain texts (most of times, the manual curator does not need to study tables and figures, but only textual information). Therefore, we filtered the ClinGen Evidence Repository for variant curations where either PS3 or BS3 evidence codes were assigned (*ClinGen variants*) and we extracted their corresponding articles' PMIDs (*ClinGen articles*). As we wanted to analyse the articles' full-text, we converted the PMIDs, which only refer to articles' abstracts, to PubMed Central [13] identification codes (PMCIDs). *ClinGen articles* with no PMCID were ignored. In this way, we obtained a set of variants for which ClinGen experts' manual curation produced at least one positive *(variant, ClinGen article)* association. Then, we applied our own manual curation to *ClinGen variants* in order to assess ClinGen's completeness in reporting positive *(variant, article)* associations; and to find negative *(variant, article)* associations for training a classifier to perform automatic positive evidence identification. We chose EuropePMC[1] as our reference literature search engine, for it has a very convenient R interface, provided by the package europepmc [14]. Specifically, the user can define a query through the function `epmc_search` and obtain PMCIDs of retrieved articles. As we

[1] https://europepmc.org/

want to mimic the same procedure followed by manual curators, our queries were structured as "`Variant_Name` AND `Gene_Name` AND `Keywords`".

`Variant_Name` is the variant identifier. Given the variety of formats commonly used in publications to refer to genetic variants [6], we used, for each *ClinGen variant*, five different queries in which `Variant_Name` was respectively represented by:

i) the nucleotide change in HGVS format (e.g., 1A>G);
ii) the nucleotide change in a non-HGVS format (e.g., A1G);
iii) the amino acid change in an HGVS format (e.g., Met1Val);
iv) the amino acid change in a non-HGVS format (e.g., M1V);
v) the RefSeq [15] Identification (rsID) code (e.g., rs786204467).

While i), ii), iii) and iv) can be derived from the *ClinGen variant*'s HGVS nomenclature reported in the ClinGen Evidence Repository, v) was obtained using the VEP [8] REST API. `Gene_name` is the gene identifier. It is needed together with the `Variant_Name` to ensure we are referring to the correct variant: two distinct articles might contain information about variants with the same variant identifier, but found on two different genes. `Keywords` is a set of 113 words extracted from two resources: Mastermind [16], which is a commercial search engine that allows paid users to rank retrieved articles according to ACMG relevance through criteria-specific keywords, and *ClinGen articles*. In particular, Mastermind contained 68 keywords for PS3/BS3, while the other 45 keywords were words recurrently found in *ClinGen articles*, which are known to be positive for PS3/BS3.

Furthermore, we refined the query syntax by adding the following flags:

- BODY: query terms were searched within the body of full-text articles. Sections such as "References" and "Acknowledgements" were not considered.
- OPEN_ACCESS: search results were limited to articles that are Open Access in EuropePMC. This was needed to access their full text.
- PUB_TYPE: filter by publication type. Only journal articles were considered.

These customised queries produced five lists of retrieved PMCIDs for each *ClinGen variant*, a list for each `Variant_Name` synonym. As some *ClinGen articles* are not Open Access in EuropePMC, some *ClinGen variant* queries did not retrieve any *ClinGen article* and thus they were discarded in the current analysis.

2.2 Manual Curation

Manual curation, i.e. manual variant annotation, was needed to distinguish between positive *(variant, article)* associations (assigned to the label 1), i.e., articles that contain at least one positive sentence activating either PS3 or BS3 evidence codes for a certain variant, and negative *(variant, article)* associations (assigned to the label 0), which do not contain any positive sentence.

For each article, we selected for manual curation only *target sentences*, i.e., sentences containing `Variant_Name` as used in all the five queries related to the same *ClinGen variant*, concatenated to the ones immediately adjacent (the previous and next

sentences). In this way, we considered only textual information, easily interpretable for an automatic algorithm (tables and figures are excluded, as they would require additional, specialised modules). Using the R package tidypmc [17], we downloaded the articles' XML code given their PMCIDs and then we performed *target sentences* extraction. For each *ClinGen variant*, we read all *target sentences* extracted from its *ClinGen articles* and, given the burden of human curation in terms of time, from a random subset of articles not included in *ClinGen articles* but retrieved by our queries. The number of articles to be curated *R* for each *ClinGen variant* was chosen considering the total number *T* of articles retrieved with all of its five queries as follows.

- If $T \leq 30$, then $R = T$.
- If $30 < T < 50$, then $R = 30$.
- If $T \geq 50$, then $R = 50$.

In this way, we included in DAVI a number of curated articles for each *ClinGen variant* that was representative of its presence in the EuropePMC database. We applied the same reasoning to choose the number r_i of articles to be curated for each of the five queries related to the same *ClinGen variant*. Considering the number t_i of articles retrieved with each of the five queries related to the same *ClinGen variant*, r_i was calculated as follows for i=1, …, 5:

$$r_i = \frac{t_i}{T} \times R \tag{1}$$

We performed manual curation considering the following rules and ensuring consistencies with the Genomic Variant Analysis & Clinical Interpretation [6] procedure. We assigned to each *(variant, article)* association the negative label (0) if none of *target sentences*, extracted from the considered article, contained sufficient information for assigning PS3 or BS3 (i.e., all *target sentences* were negative), regardless of the content of tables or figures (which might have contained information for assigning PS3 or BS3, but whose automated analysis was out-of-scope for this work). Instead, the positive label (1) was assigned to *(variant, article)* associations for which at least one *target sentence* contained information for assigning PS3 or BS3 (i.e., at least one *target sentence* was positive).

3 Automatic Variant Annotation

3.1 Pre-processing

We trained the automatic classification model on DAVI according to a by-sentence perspective, where we considered each *target sentence* as an independent entry. For performance evaluation only, we considered a by-article perspective, distinguishing between positive and negative *(variant, article)* associations according to the classification of each of their extracted *target sentences* (association is positive if the article contains at least one positive *target sentence*). We pre-processed the *target sentences* included in DAVI according to the following typical steps [18].

- English stop word removal, using the stop list provided by the package nltk [19]. We excluded the word "not", which is a relevant word in the context of PS3 or BS3 assignment, and we handled negation by concatenating it to the following word.
- Stemming using the snowball stemming algorithm implemented by the package nltk.
- Removal of words with an absolute frequency less than the 90th percentile of the absolute frequency distribution of words in the vocabulary.
- Exclusion of sentences consisting of less than 3 words

We split the pre-processed dataset into a training set, a test set and a validation set (70%, 15%, 15%), making sure that proportions of positive and negative *(variant, article)* associations and *target sentences* were similar (within a tolerance of $\varepsilon = 0.01$) in the 3 subsets. Finally, given that we were considering the classification of *(variant, article)* pairs, but we needed to construct a single dataset comprising all *target sentences*, we had to deal with the presence of duplicated *target sentences.* As duplicated *target sentences* could cause over-fitting (identical *target sentences* with concordant labels) or bias (identical *target sentences* with discordant labels), we removed one copy, if concordant, or both, if discordant, of such sentences from the training and validation sets. This reasoning was not applied to the test-set, as it was used for performance evaluation only: predicted labels were correctly computed considering *target sentences* extracted from articles in *(variant, article)* pairs.

3.2 Model Construction

We applied three different text representation schemes, implemented through the python package scikit-learn [20], to transform *target sentences* in the pre-processed DAVI into sequences of numbers.

- Binary Bag Of Words (BBOW) [21], in which each word was represented by 1, if the word is present in the *target sentence*, and 0 otherwise.
- Bag Of Words (BOW), in which each word was represented by its frequency in the *target sentence*.
- Term-frequency Inverse Document-Frequency (TF-IDF), in which each word was represented by its frequency in the *target sentence* weighted by how often it appeared in all *target sentences*.

We performed a preliminary exploratory analysis on automatic PS3/BS3 evidence identification using a logistic regressor (LR) trained on the three versions of DAVI. For this model, we considered a L2 regularisation loss-function with a single hyperparameter, the inverse of the regularisation strength C. For each version of the dataset (BBOW, BOW, TF-IDF), we performed hyperparameter optimisation considering only the training set, using a 5-fold cross validation [22] and a random search approach [23] accounting for 10000 values of C, randomly sampled from a log uniform distribution ranging from 10^{-4} to 10^{2}. We selected the best hyperparameter as the one that led to the minimum average binary cross-entropy across the 5 folds.

To transform the model from a ranker into a classifier, useable in practice for automatic PS3/BS3 evidence identification, we implemented a thresholding approach by

identifying one probability threshold (th) to discriminate between positive (1, if predicted probability $p \geq th$) and negative (0, if $p < th$) predictions on *target sentences*. We selected the optimal threshold by using each probability value predicted for *target sentences* in the validation set as a threshold and choosing the one associated to the maximum geometric mean between true positive and true negative rate in the validation set itself.

3.3 Performance Measures

In the by-sentence perspective, we evaluated the discrimination performance of the model via five measures: area under the receiver operating characteristic (AUROC) and area under the precision-recall curve (AUPRC) [24] for the continuous probability output; as well as precision, recall, and F1-score after applying the aforementioned thresholding approach.

In the by-article perspective, we did not consider AUROC and AUPRC as predicted labels were assigned as the logical OR of by-sentence outputs after thresholding and, hence, were Boolean in nature.

4 Results

4.1 Manual curation results

DAVI is organised into 6 columns, containing, for each *(variant, article)* pair, the variant's HGVS standard nomenclature, variant's HGVS nomenclature used in query, the article's PMCID, the label assigned to the article, a *target sentence* extracted from the article and, the label assigned to that *target sentence*. Table 1 shows an example of a DAVI entry.

Table 1. Example of an entry in DAVI

HGVS standard nomenclature	HGVS nomenclature used in query	Article PMCID	Article Label	*Target Sentence*	*Target Sentence* Label
NM_021133.4 (RNASEL): c.793G>T (p. Glu265Ter)	G793T	PMC2361943	0	All sequence variations […]. […], we discovered one protein-truncating variant, nt g793t, […]. This point mutation […].	0

Overall, DAVI contains the results of manual curation for 41 *ClinGen variants*, yielding 1239 *target sentences* extracted from 311 *(variant, article)* pairs, namely 44 *(variant, ClinGen article)* pairs and 267 *(variant, non-ClinGen article)* pairs. Table 2 provides a comparison of the labels assigned to *target sentences* and *(variant, article)* pairs in *ClinGen articles* and non-*ClinGen articles*.

Table 2. Comparison of the assigned labels in *ClinGen articles* and non-*ClinGen articles*

		Total	Positive	Negative
ClinGen articles	*Target sentences*	388	219	169
	(Variant, article) pairs	44	37	7
Non-*ClinGen articles*	*Target sentences*	851	378	473
	(Variant, article) pairs	267	117	150

DAVI contained almost the same amount of positive and negative *target sentences*, i.e., respectively 597 and 642, and positive and negative *(variant, article)* pairs, i.e., respectively 154 and 157. The number of *target sentences* extracted per *ClinGen article* was three times greater than the number of *target sentences* extracted per non-*ClinGen article*. Moreover, even though we assumed *(variant, ClinGen article)* associations to be positive, 7 *(variant, ClinGen article)* pairs were re-classified as negative after manual curation (see Section 2.2).

4.2 Pre-processing results

Initially, the vocabulary of *target sentences* contained 7745 words. Following the approach described in Section 3.1, we removed 733 stop words and 6309 words as they had an absolute frequency below the 90th percentile of the absolute frequency distribution of words in the vocabulary. We removed 2 negative *target sentences*, as they comprised less than 3 words. Lastly, we removed 66 and 22 duplicated concordant *target sentences* from training-set and validation-set, respectively, whereas no duplicated discordant *target sentences* were found. Thus, the pre-processed DAVI finally contained 1149 *target sentences*. Table 3 provides a comparison of the assigned labels of *target sentences* and *(variant, article)* pairs in the training, test, and validation sets after pre-processing.

Table 3. Comparison of assigned labels in the pre-processed training, test, and validation sets

		Total	Positive	Negative
Training-set	*Target sentences*	644	320	324
	(Variant, article) pairs	196	99	97
Test-set	*Target sentences*	302	144	158
	(Variant, article) pairs	52	25	27
Validation-set	*Target sentences*	203	99	104
	(Variant, article) pairs	48	24	24

4.3 Classification results

This section reports the performance of the models constructed following the approach described in Section 3.2 and using the measures introduced in Section 3.3. Results of hyperparameter C optimization on the three versions of the training set (BBOW, BOW,

TF-IDF), minimizing the score (binary cross entropy) across the 5-folds are reported in Table 4.

Table 4. Hyperparameter C optimization on the BBOW, BOW and TF-IDF versions of DAVI

Text representation scheme	Hyperparameter (C)	Best Score
BBOW	0.103	-0.541
BOW	0.055	-0.549
TF-IDF	2.236	-0.557

The combination of TF-IDF text representation model and hyperparameter C = 2.236 led to the lowest value of binary cross entropy. The performance metrics obtained in the by-sentence and by-article perspectives, using BBOW+LR, BOW+LR, and TF-IDF+LR, are shown in Table 5.

Table 5. Classification results according to by-sentence and by-article perspectives, using BBOW+LR, BOW+LR and TF-IDF+LR

Perspective	Model	AUROC	AUPRC	TP	TN	FP	FN	Precision	Recall	F1-score
By-sentence	BBOW+LR	0.805	0.767	132	81	77	12	0.631	0.917	0.748
	BOW+LR	0.815	0.771	124	108	50	20	0.713	0.861	0.780
	TF-IDF+LR	0.819	0.796	121	99	59	23	0.672	0.840	0.747
By-article	BBOW+LR	-	-	24	15	12	1	0.667	0.960	0.787
	BOW+LR	-	-	23	19	8	2	0.742	0.920	0.821
	TF-IDF+LR	-	-	24	19	8	1	0.750	0.960	0.842

In the by-sentence perspective, the TF-IDF+LR model performed better, yielding an AUROC and a AUPRC of 0.819 and 0.796, respectively. However, the BBOW+LR model showed a higher recall and, overall, the BOW+LR model had a higher F1-score. In the by-article perspective, the best performing model was TF-IDF+LR. As the number of false negatives was lower than the one of false positives, recall was higher than precision. This result suggests that correctly identifying positive sentences and articles was slightly more challenging than correctly identifying negative cases. While not directly comparable, performance was overall better in the by-article setting than in the by-sentence one, which was expected as it is easier to obtain a correct classification looking at multiple *target sentences* for each *(variant, article)* pair rather than classifying each *target sentence* independently.

5 Discussion and Future Work

The main aim of this work was to build a high quality and manually-curated dataset that associates each variant to its PS3/BS3-activating articles (positive associations) and non PS3/BS3-activating articles (negative associations), as such resource is critical for clinical decision-making and it is currently missing. The second aim was to use such dataset to train a predictive model that efficiently performs automatic positive associations identification.

We built DAVI, a manually-curated dataset comprising 1239 sentences related to 311 *(variant, article)* associations. In order to guarantee a sufficient number of positive associations, we included in DAVI 44 *(variant, ClinGen article)* pairs and, to consider a more representative sample of the entire corpus of articles retrieved when querying for a specific variant, 267 *(variant, non-ClinGen article)* pairs. As expected, most *(variant, ClinGen articles)* pairs were positive, but 7 were reclassified as negative, on the basis of textual information only. Overall, about half of the extracted sentences contained sufficient evidence for activating PS3 or BS3 evidence codes, and same for the *(variant, article)* pairs. A positivity offset is given by the fact that we forcedly included in DAVI an elevated number of positive sentences extracted from few *ClinGen articles*, but, generally, positive sentences and *(variant, article)* associations are respectively fewer than negative ones (378 vs. 473 sentences extracted from 117 vs. *150 (variant, non-ClinGen articles)*). However, we found a significant number of positive examples in *(variant, non-ClinGen articles)* pairs (117 out of 267), suggesting that the manually curated information contained in ClinGen is incomplete and/or not updated frequently enough. As ClinGen has been recognised by the Food and Drug Administration as a source of valid scientific evidence for support in clinical decisions, it should be always up-to-date, containing all new evidence about all discovered variants.

ACMG criteria and, specifically, PS3 and BS3, can activate for a *(variant, article)* in relation to multiple specific diseases or through the use of different types of experimental texts, sometimes even at the same time. Therefore, it is crucial to provide to the clinician all available positive evidence, even if this implies higher costs for manual curation. In order to reduce these costs, automatic models could be integrated in the curation pipeline. As an exploratory analysis on the feasibility of this approach, we tested the discrimination performances of three predictive models, trained on DAVI, for the automatic identification of positive *(variant, article)* associations. Performance was good both in the by-sentence and by-article perspective, with F1-scores well above 0.70 and 0.80 respectively. This result suggests that reliable tools could be developed in support of manual curation, efficiently enriching biological databases with all the information needed for a complete and correct variant interpretation.

Future developments include the further distinction of (PS3 or BS3)-positive examples into PS3-positive vs. BS3-positive examples. Moreover, the solid manual curation procedure described in this work may be applied to variants which are not included in ClinGen and expanded to the evaluation of other evidence codes among the 28 covered by ACMG guidelines. Lastly, we may focus on the development of more complex architectures for text representation and classification, including deep learning approaches.

References

1. Collins, F. S., Fink, L.: The Human Genome Project. Alcohol Health Res World 19(3), 190–195 (1995).
2. Morash, M., Mitchell, H., Beltran, H., Elemento, O., Pathak, J.: The Role of Next-Generation Sequencing in Precision Medicine: A Review of Outcomes in Oncology. Journal of Personalized Medicine 8(3), (2018).
3. Amendola, L. M., et al.: Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. Am J Hum Genet 98(6), 1067–1076 (2016).
4. Richards, 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 17(5), 405–424 (2015).
5. GVACI Course 2022. https://gvaci.genomes.in/home, last accessed 2022/12/29.
6. Lee, K., Wei, C.-H., Lu, Z.: Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature. Brief Bioinform 22(3), (2020).
7. Welcome to ClinGen. https://www.clinicalgenome.org/, last accessed 2022/12/29.
8. McLaren, W., et al.: The Ensembl Variant Effect Predictor. Genome Biol 17, (2016).
9. Wang, K., Li, M., Hakonarson, H.: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16), (2010).
10. Den Dunnen, J. T., et al.: HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Human Mutation 37(6), 564–569 (2016).
11. Karczewski, K. J., et al.: The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res 45, (2017).
12. PubMed, https://pubmed.ncbi.nlm.nih.gov/, last accessed 2023/01/03.
13. Home - PMC – NCBI, https://www.ncbi.nlm.nih.gov/pmc/, last accessed 2023/01/03.
14. Levchenko, M., et al.: Europe PMC in 2017. Nucleic Acids Research, (2017).
15. RefSeq: NCBI Reference Sequence Database, https://www.ncbi.nlm.nih.gov/refseq/, last accessed 2023/05/03.
16. Chunn, L. M., et al.: Mastermind: A Comprehensive Genomic Association Search Engine for Empirical Evidence Curation and Genetic Variant Interpretation. Front Genet 11, (2020).
17. Stubben, C.: tidypmc: Parse Full Text XML Documents from PubMed Central. (2019).
18. Kathuria, A., Gupta, A., Singla, R. K.: A Review of Tools and Techniques for Preprocessing of Textual Data. In: Singh, V., Asari, V. K., Kumar, S. Advances in Intelligent Systems and Computing 2021, pp. 407–422. Springer, Singapore (2021).
19. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., (2009).
20. Pedregosa, F., et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011).
21. Qader, W. A., Ameen, M. M., Ahmed, B. I.: An Overview of Bag of Words; Importance, Implementation, Applications, and Challenges. In: International Engineering Conference (IEC) 2019, pp. 200–204, (2019).
22. Berrar, D.: Cross-Validation (2018).
23. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res 13, 281–305 (2012).
24. Keilwagen, I. G., Grau, J.: Area under Precision-Recall Curves for Weighted and Unweighted Data. PLoS One 9(3), (2014).