Predicting ClinVar Reclassifications

Ani Khachatryan[1]
[1]**Department of Biomedical Informatics, Columbia University in the City of New York**

Abstract

ClinVar contains more than 600 thousand variant-condition pairs that have assertions on the clinical significance of variants for the specified condition. The assertions can be of varying granularity and quality and so an evaluation of ClinVar is necessary to assess the quality and trustworthiness of the knowledge contained in ClinVar. In this project I evaluated a subset of features that might be predictive of reclassification efforts of ClinVar variants. I highlighted some of the issues regarding analyzing ClinVar reclassifications by conducting an exploratory analysis which can serve as a pilot study for later projects studying ClinVar reclassifications. I also fit a logistic regression model to predict reclassification events and discussed some of its limitations and performance issues.

Introduction

ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) at the National Center for Biotechnology Information (NCBI) is a freely available archive for interpretations of clinical significance of variants for reported conditions.[1] As next-generation sequencing becomes more and more ubiquitous, the amount of novel variants keeps growing and the interpretation of those variants becomes increasingly challenging. Thus, it is highly advantageous for the community to share, centralize, and standardize the data available from a multitude of sources to combine efforts in identifying the relationships between genetic variation and human disease. The ClinVar database provides a platform for such efforts and accepts submissions relating to interpretations of clinical and/or functional significance of variants for specified conditions, with opportunities to provide the supporting evidence.

ClinVar supports submissions of differing levels of granularity. Therefore, the submissions can range from simple assertions about an allele and its interpretation to detailed descriptions of multiple types of structured observational or experimental evidence about the effect of the variation on phenotype.[2] Interpretations are aggregated by variant-condition combination and are assigned an accession number prefixed with RCV. Clinical significance is calculated for the aggregate record indicating consensus or conflict in the submitted interpretations. Variants that do have a conflict within the scale of pathogenicity are reported with a clinical significance of conflicting interpretations of pathogenicity.[1] It is of paramount importance to note that ClinVar neither curates nor modifies interpretations independent of an explicit submission.[2] Therefore, the quality of assertions in ClinVar are highly variable and largely dependent on the external submissions regarding the interpretations.

As the process of determining the clinical significance for variant-condition pairs is dynamic and dependent of multiple sources of assertions and evidence, the current interpretations of the variant-condition pairs are subject to change with accumulating evidence. In fact, a major goal of ClinVar is to support computational (re)evaluation of assertions and to support the evolution of and development of knowledge regarding variations and associated phenotypes.[2] Occasionally, reclassifications of variant-condition clinical significance do occur and it is of value to understand the patterns of such reclassifications. Understanding the core reasons for reclassifications and building a robust model that can predict the necessity of reclassifications will enable the evaluation of data quality in ClinVar. In this paper, I will discuss my efforts towards assessing different features for relevance to reclassification events and building a model that can predict need of reclassification for a given variant.

Literature Review

There has been some effort in analyzing discordance in ClinVar variants and studies have focused on analyzing the indicators of conflicts. Yang *et al.*[3] have showed that in germline variants concordance rates were different among different clinical areas and variant types. They also showed that clinical testing variants had much higher concordance rates than research and curation variants. Another determinant of concordance rate was the age of the variant, older variants having higher discordance rates. And last but not least, low-penetrance variants had higher discordance rates. Penetrance was studied also by Shah *et al.*[4] who sequenced more than 10,000 individuals and showed that number of starts (Review Status) is indicative of higher rates of misclassification through comparing observed genetic risk to the reported population prevalence for the conditions. Another study[5] used significance type, age of submission and submitter expertise to prioritize variants in need of reclassification. However, to my knowledge, there have been no studies trying to predict reclassification events.

Methods

Data acquisition and processing

ClinVar releases an xml file containing the complete variant-condition-level data every month. I downloaded all such releases starting from November 2012 from ClinVars Downloads/FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/xml/). Each unzipped file was around 8 GBs large. Since I did not have access to a CPU/GPU cluster, I was constrained to use a personal laptop for all analyses (Core i7 7500U - 16GB RAM) and therefore had to filter the more than 600,000 ClinVar RCV records (aggregate variant-condition record) for further analysis.

I chose to concentrate on conditions associated with cancer and after initial filtering was left with 134,870 RCV records in the last ClinVar release. I converted the data from an .xml to a .csv format and extracted the features selected for further analysis: Review Status, Clinincal Significance, Date Last Evaluated, Type, Method Type, and Origin. There were inconsistencies between releases and a lack of standard vocabulary for reporting clinical significance. For instance, inconsistencies ranged from simple and easy to solve issues such as ”pathogenic” vs ”Pathogenic” to more involved differences such as ”uncertain significance” vs ”variant of unknown significance” which required manual mapping and standardization.

After such data cleaning which also involved removing RCV records with missing data and outlier records with clinical significance outside of the five terms recommended by ACMG/AMP[6], I was left with 129,347 RCV records. To assess the degree of reclassification among the records I looked at the distribution of versions among the 129,347 records (Fig. 1). Over 40,000 records have versions higher than one. However, as it turned out, version change does not necessarily imply change in clinical significance. Therefore, there was a need to identify records with actual clinical significance modifications which was not a trivial task since ClinVar does not offer a straightforward way of accessing such information. Thus, I devised an algorithm that systematically searches all ClinVar releases and for each record identifies the release in which the record appeared the first time and compares the clinical significance in the first version to the version in the latest ClinVar release. With such comparison I was able to identify 5,211 records that have been reclassified compared to their first versions.

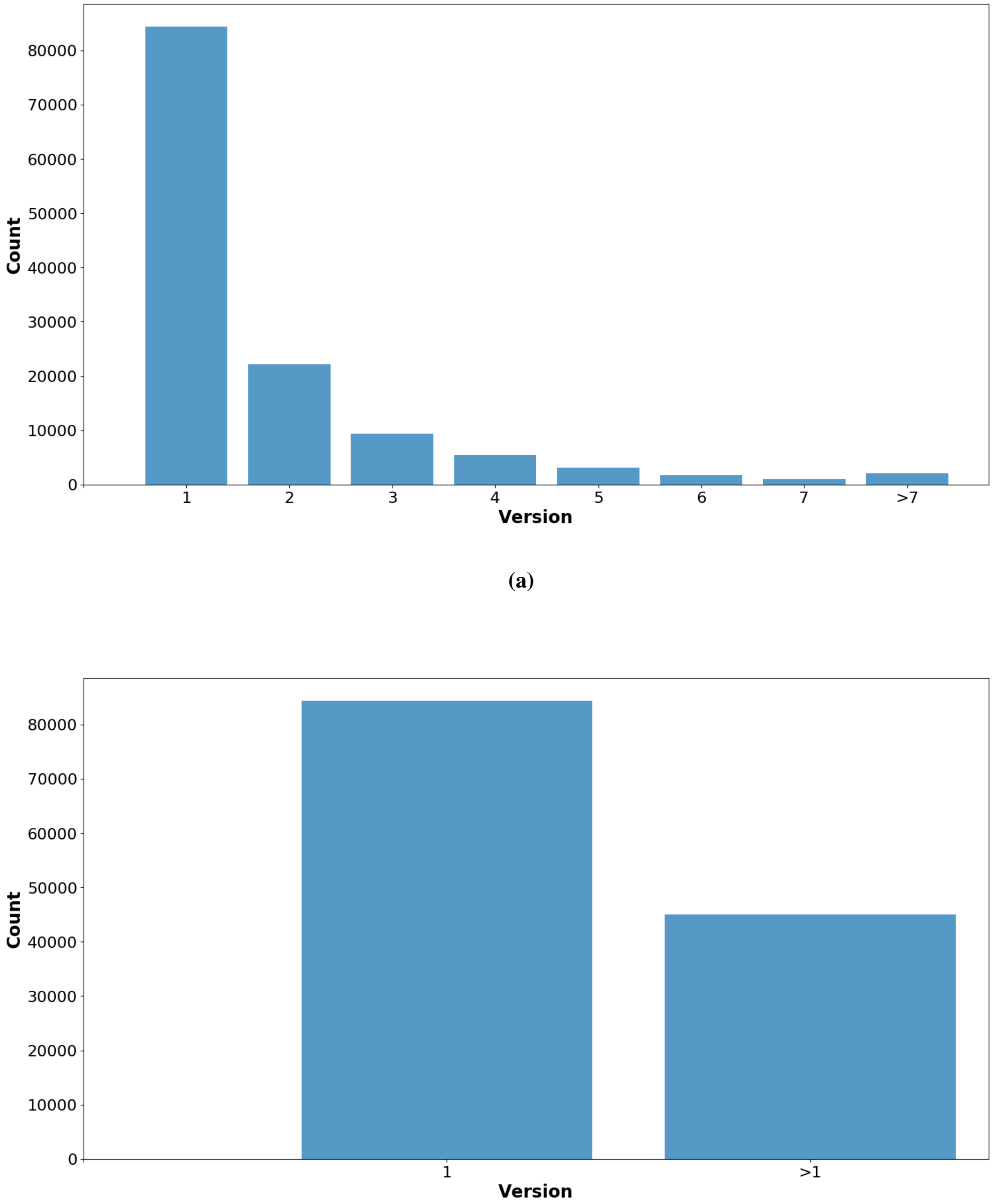
It is important to note that for such reclassified records I extracted the information from their respective first versions and not the latest ClinVar release because the latest release contains already reclassified records and I was interested in records that get reclassified. These records were labelled as “positive“ for further analysis and records that have not been reclassified were labeled as “negative.“

Exploratory Analysis

In order to assess the selected features for relevance to reclassification efforts, I conducted an exploratory analysis which will be discussed in more detail in the **Results** section. First, I looked at the distribution of clinical significance among reclassified variants before and after reclassification (Fig. 3a) and at the distribution of types of reclassifications (Fig. 3b).

I also looked at the distribution of variant age in reclassified and non-reclassified records. For reclassified records, age was defined as the number of years between variant addition and variant reclassification. For non-reclassified variants, it was defined as the number of years between variant addition and November 2018 (last ClinVar full xml release analyzed). Fig 4a shows the age distribution of reclassified variants and Fig 4b shows the age distribution of non-reclassified variants.

I then calculated the chi-square statistic for each of the features (Table 1). To get a more visual representation of the features in the dataset, I plotted for each feature its distribution in both reclassified and not reclassified records (Fig. 2). To test the hypothesis that reclassifications happen due to previous assertions being made on small and non-inclusive



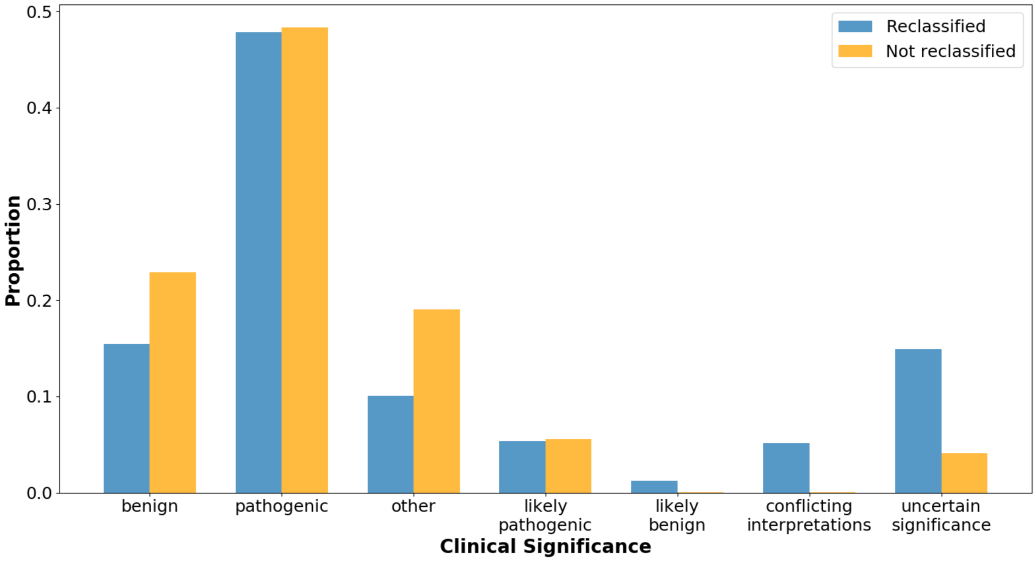
(a)

(b)

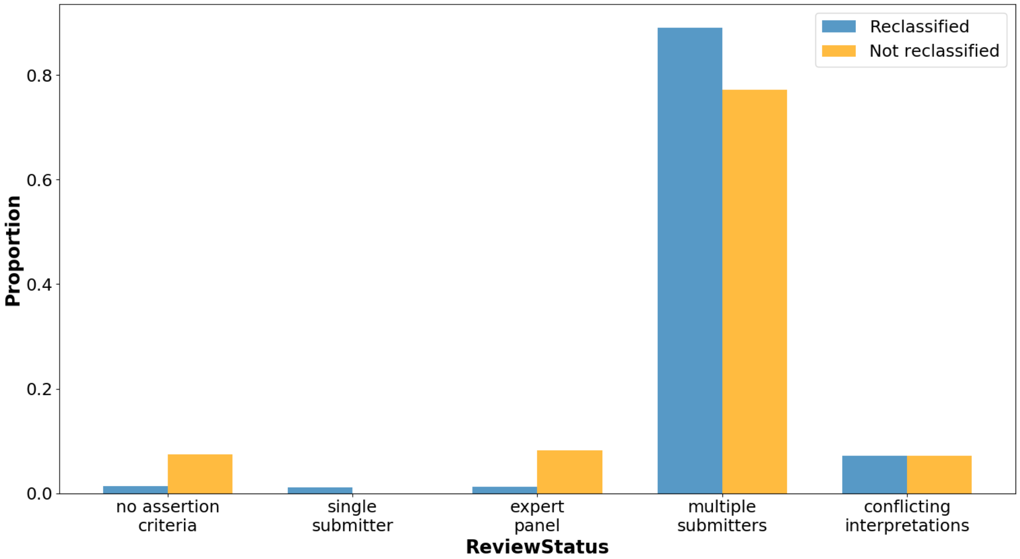
Figure 1: The distribution of versions among the cancer subset of ClinVar records.

samples, I also did a qualitative analysis and took a random sample of 30 reclassified variants that had sample ethnicity and/or sample size information and looked at the sample descriptions.

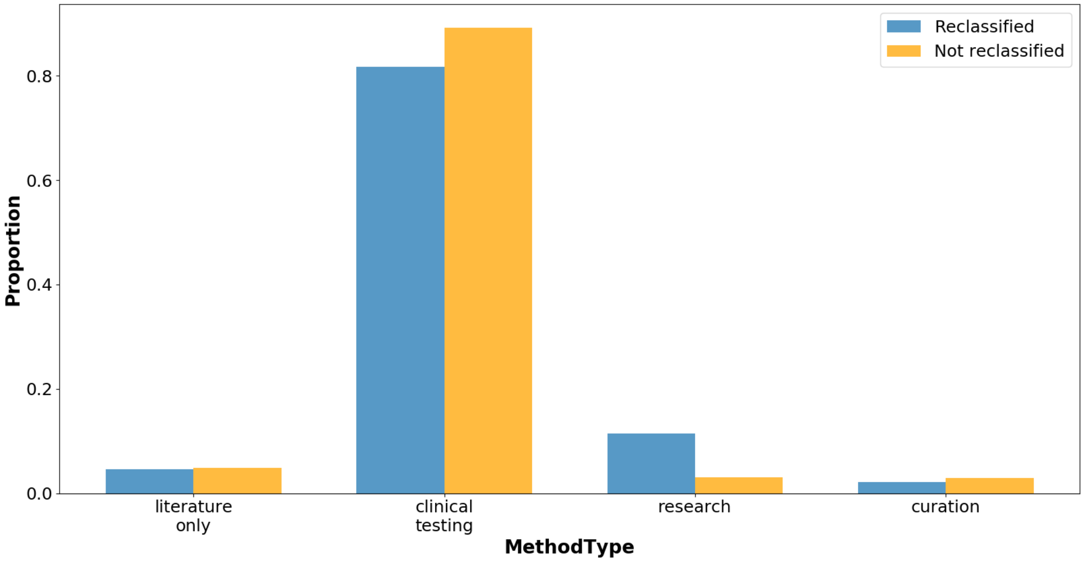
Then I utilized different feature extraction methods to find k best features (Table 2) and fit a logistic regression model using sci-kit learns LogisticRegression module[7] (Table 3). To find the optimal k, I looked at the various accuracy metrics dependent on k (Fig. 6).



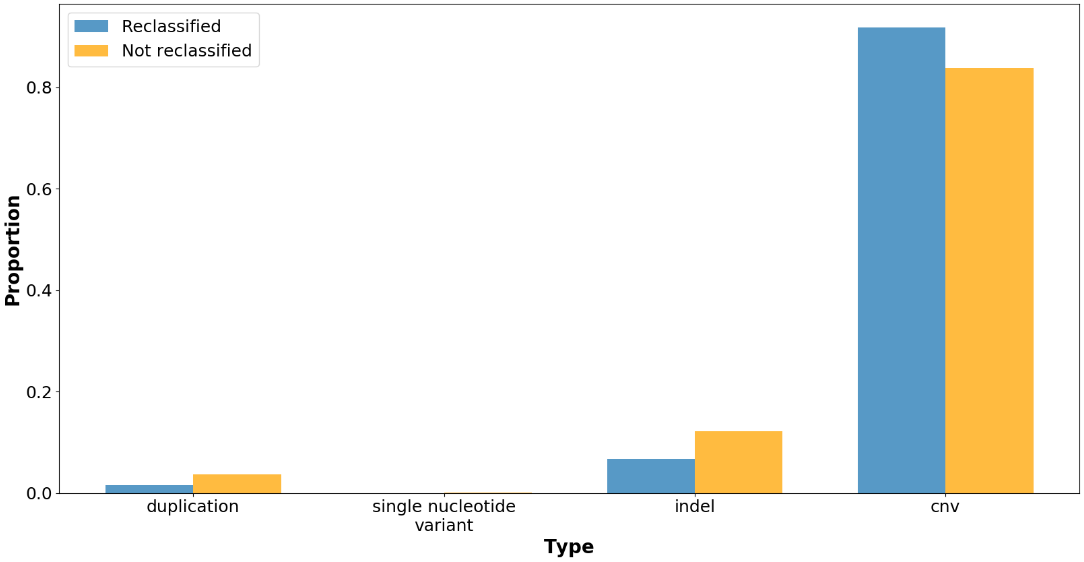
(a)



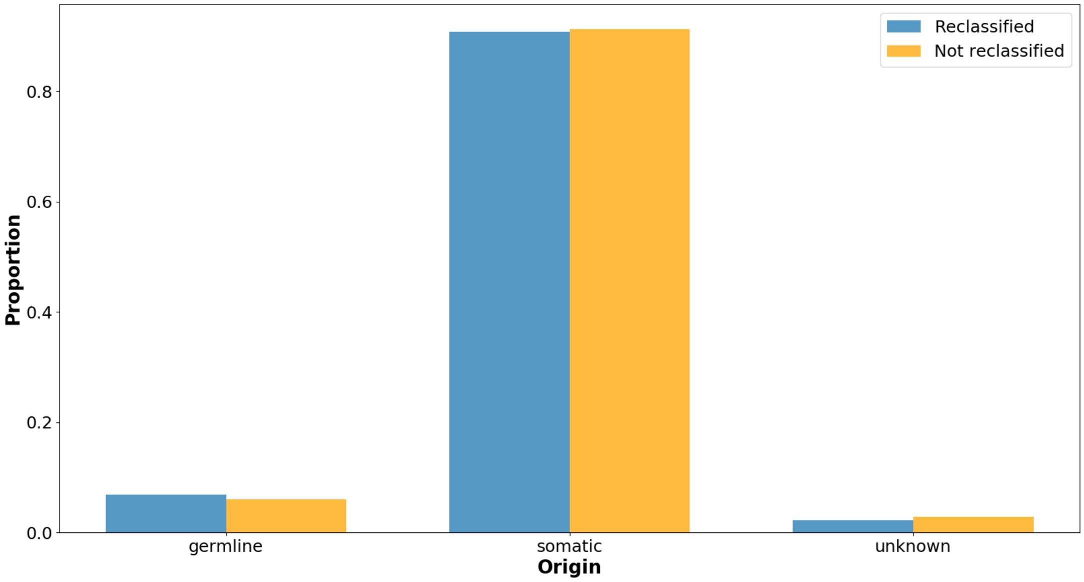
(b)


Reclassified
Not reclassified
Proportion
0.0
0.2
0.4
0.6
0.8
literature only
clinical testing
research
curation
MethodType

(c)


Reclassified
Not reclassified
Proportion
0.0
0.2
0.4
0.6
0.8
duplication
single nucleotide variant
indel
cnv
Type

(d)

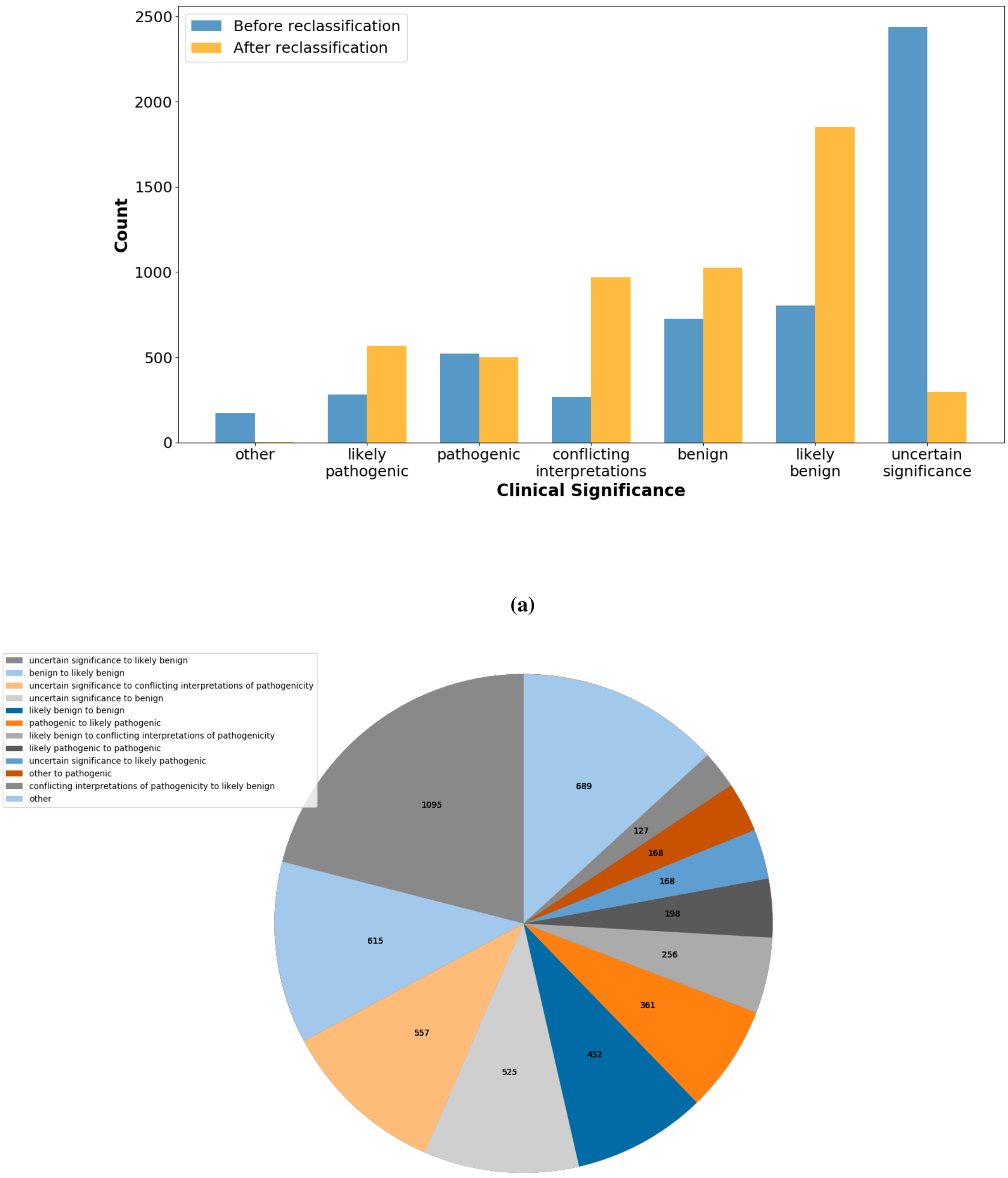


(e)

Figure 2: The distribution of features between reclassified and not reclassified records.

Results

As has already been discussed in **Methods**, I started the analysis with 129,347 RCV records 5,211 of which have been reclassified. Fig. 3a shows the distribution of clinical significance before and after reclassification in the 5,211 reclassified records. Fig 3b shows the most common types of changes in the reclassified dataset.



(a)

(b)

Figure 3: (a) Distribution of Clinical Significance before and after reclassification. (b) Pie chart showing the distribution of the types of reclassifications.

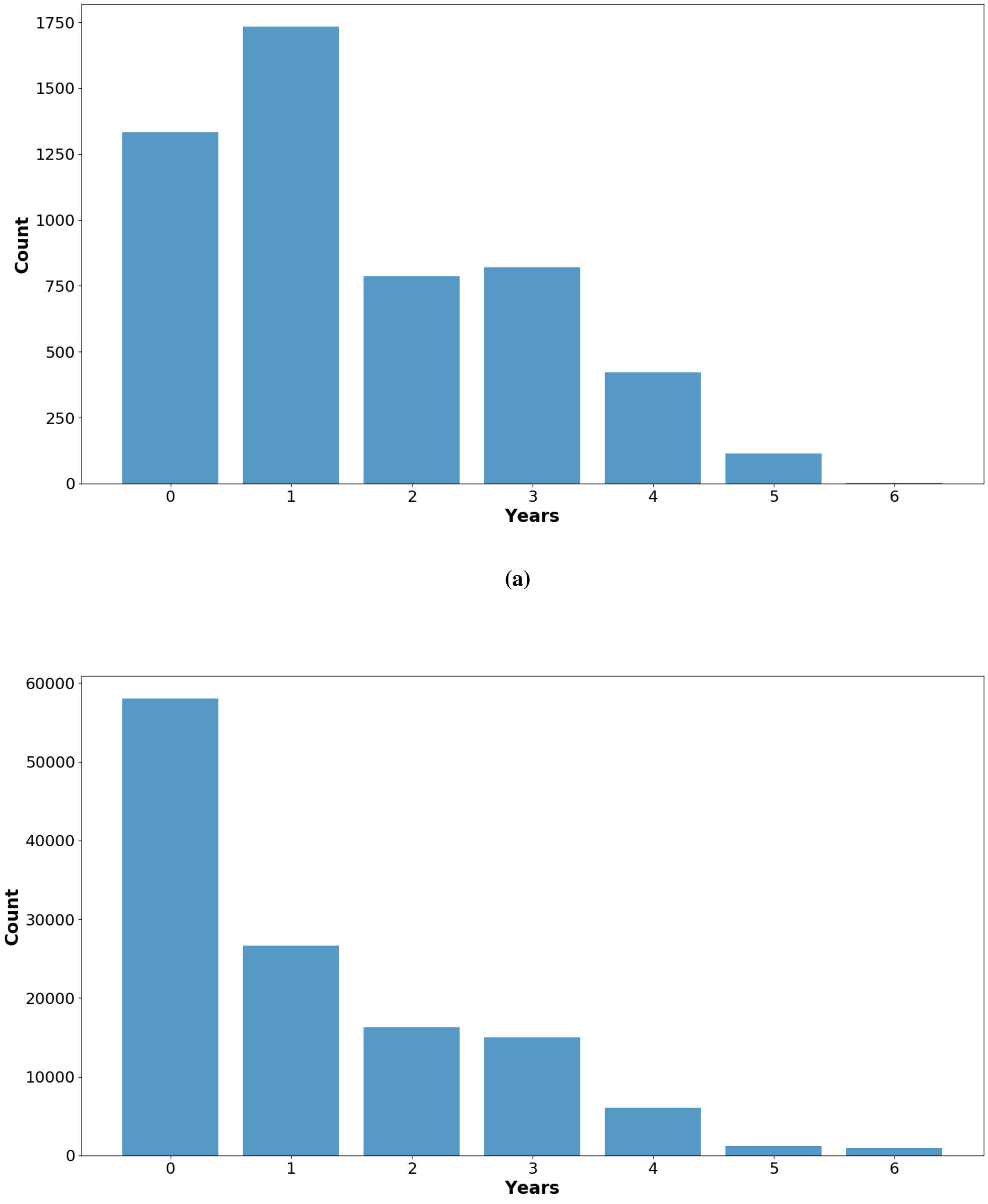
It can be seen from Fig. 3 that the most common variants to get reclassified are Variants of Uncertain Significance (VUS) which is expected since VUSs are the primary records in need of reclassification. Another thing to note is the difference between proportions of variants with conflicting interpretations of pathogenicity, more specifically, the increase in such variants after reclassification which is due to new evidence resulting in reclassification which often contradicts with previous assertions. In fact, the qualitative analysis of 30 random reclassified records showed that the initial assertions were made with a sample size of on average 1.17 individuals of predominantly European ethnicity. It can be hypothesized that such small and non-inclusive samples might be inadequate for drawing conclusions on variant clinical significance. However, more sophisticated methods are necessary to test the hypothesis.

I also looked at the differences between the positive and the negative dataset in terms of the age of the variants. As can be seen from Fig. 4, there is a substantial difference between the two distributions (2 sample Kolmogorov-Smirnov test, p-value = 1.33e-188). The mean age of reclassification is 1.58 years, whereas the mean age of records in the negative dataset is just 1.16 years. It can be concluded from these observations that the negative dataset contains records that have not been reclassified because they are too ”young” but nevertheless need reclassification. On top of that, the negatives might also contain records that have not been reclassified either due to a lack of effort or additional evidence. In fact, as can be seen from Fig. 5, the majority of records in the negative dataset have a clinical significance of ”uncertain significance” which indicates need of reclassification. Moving on with the analyses, it should be taken into account that the negative dataset does not fully comprise of true negatives, but might be heavily contaminated by false negatives and it is a non-trivial task to differentiate between those. Such false negatives effectively add noise in the chi-square statistic calculated for each feature and subsequently, any model built on these features.

I then used sci-kit learn’s SelectKBest module[7] which utilizes the chi-square statistic to select the most important features for fitting a logistic regression (Table 1). Figure 6 shows the dependency of model performance in terms of various accuracy metrics (accuracy, ROC-AUC, PR-AUC) from k. I chose 8 as the optimal k and used it for further analyses. Table 2 shows the comparison between the rankings of two other feature selection methods (Recursive Feature Elimination[7] and Feature Importance[7]). It can be noted that the different methods give different rankings for the features. Recursive Feature Elimination uses logistic regression while Feature Importance uses decision trees to select for best features. The discrepancy between the results might be explained by the noise in the negative dataset.

To see if the model will be able to differentiate between reclassified and not reclassified variants, I fit a logistic regression model using the 8 best features as predicted by SelectKBest with sci-kit learn’s LogisticRegression module[7]. The model classification report is shown in Table 3.

The model performance is in no way close to desirable and my analyses show that the reason is the considerable amount of false negatives in the dataset. It can be seen from the model parameters (Table 3) that the model has a low recall for reclassified variants. This is an issue and comes from the fact that the negative dataset is full of false negatives that bias the analyses. Thus, removing false negatives and re-running the model should improve the predictive power of the model. One such obvious example of false negatives are variants with “uncertain significance.“ Table 4 shows the classification report of a model trained on the same dataset, but variants with “uncertain significanc“ were removed from the negatives. As can be seen, the recall for the positive class after such rudimentary filtering went from 0.25 to 0.60. Thus, having a trustworthy negative dataset will enable to build a model with sufficient predictive power allowing to evaluate ClinVar variants and highlight reclassification needs.



(a)

(b)

Figure 4: Distribution of age of variants in the reclassified (b) and non-reclassified (a) records.

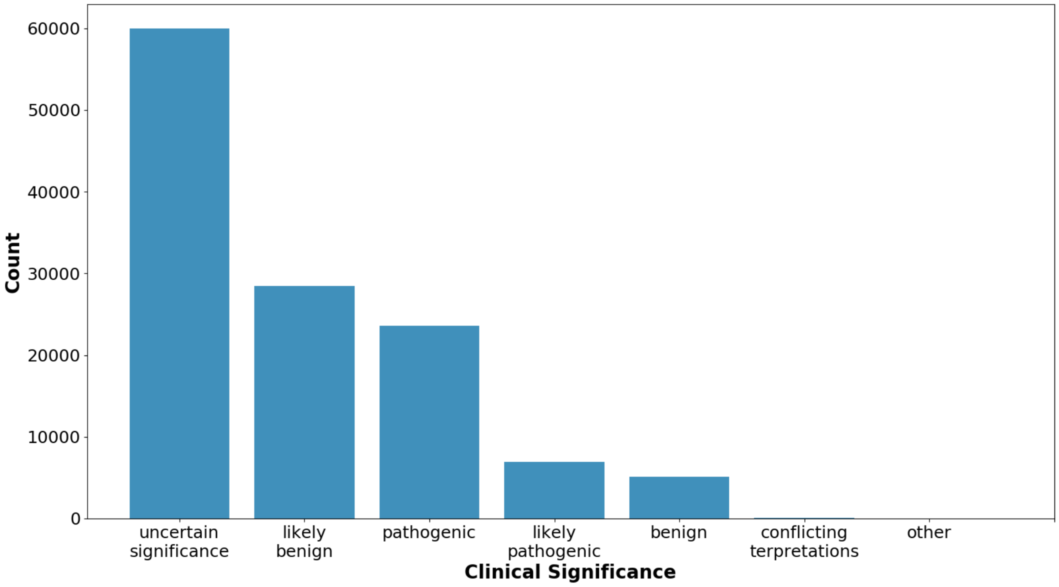


Figure 5: The distribution of clinical significance in the negative dataset.

Table 1: Chi-square statistics for each feature studied.

Feature	**Chi-square statistic**
Date Last Evaluated	0.0
Clinincal Significance: conflicting interpretations of pathogenicity	0.0
Clinincal Significance: benign	0.0
Clinincal Significance: other	0.0
Review Status: criteria provided, conflicting interpretations	0.0
Method Type: curation	0.0
Review Status: no assertion criteria provided	0.0
Review Status: reviewed by expert panel	0.0
Clinincal Significance: pathogenic	0.0
Type: indel	0.0
Clinincal Significance: likely benign	0.0
Review Status: criteria provided, single submitter	0.0
Type: duplication	0.0
Origin: unknown	0.0
Type: single nucleotide variant	0.0
Origin: somatic	6.07e-06
Method Type: clinical testing	0.00038692
Type: cnv	0.00727529
Method Type: research	0.01065522
Origin: germline	0.0182922
Clinincal Significance: likely pathogenic	0.48726769
Review Status: criteria provided, multiple submitters, no conflicts	0.57938503
Clinincal Significance: uncertain significance	0.62021063
Method Type: literature only	0.66055066

Table 2: Comparison of results of different feature selection methods. RFE: Recursive Feature Elimination. ETC: Extra Tree Classifier

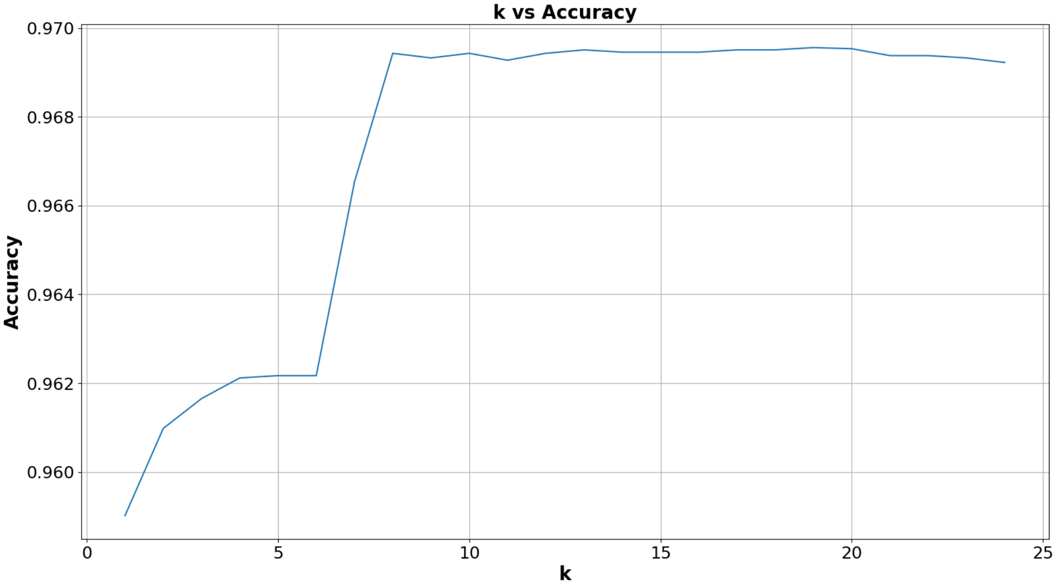
Feature	**Rank by RFE**	**Rank by ETC**
Date Last Evaluated:	19	1
Clinical Significance: conflicting interpretations of pathogenicity	1	3
Clinical Significance: benign	16	6
Clinical Significance: other	2	9
Review Status: criteria provided, conflicting interpretations	14	15
Method Type: curation	6	4
Review Status: no assertion criteria provided	13	7
Review Status: reviewed by expert panel	7	5
Clinical Significance: pathogenic	8	16
Type: indel	21	20
Clinical Significance: likely benign	9	10
Review Status: criteria provided, single submitter	3	2
Type: duplication	20	17
Origin: unknown	15	18
Type: single nucleotide variant	22	23
Origin: somatic	12	19
Method Type: clinical testing	5	12
Type: cnv	24	24
Method Type: research	23	21
Origin: germline	18	22
Clinical Significance: likely pathogenic	11	14
Review Status: criteria provided, multiple submitters, no conflicts	4	11
Clinical Significance: uncertain significance	10	13
Method Type: literature only	17	8

Table 3: Classification report for the logistic regression model (k=8). 0 refers to non-reclassified, 1 refers to reclassified.

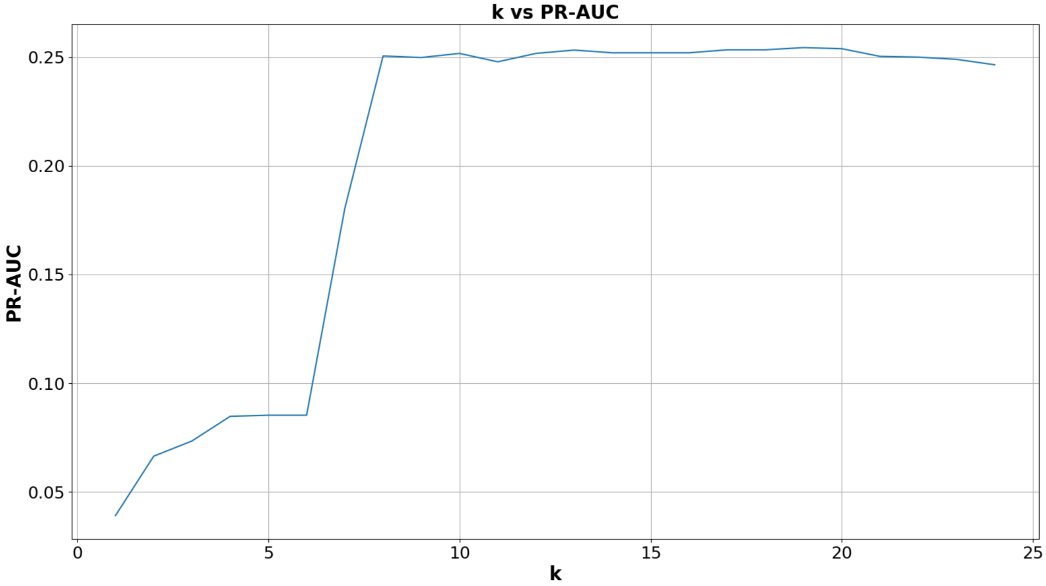
	precision	**recall**	**f1-score**	**support**
0	0.97	1.00	0.98	37,186
1	0.90	0.25	0.39	1,513
micro avg	0.97	0.97	0.97	38,699
macro avg	0.93	0.62	0.69	38,699
weighted avg	0.97	0.97	0.96	38,699

Table 4: Classification report for the logistic regression model (k=8) with variances of ”uncertain significance” removed from the negatives. 0 refers to non-reclassified, 1 refers to reclassified.

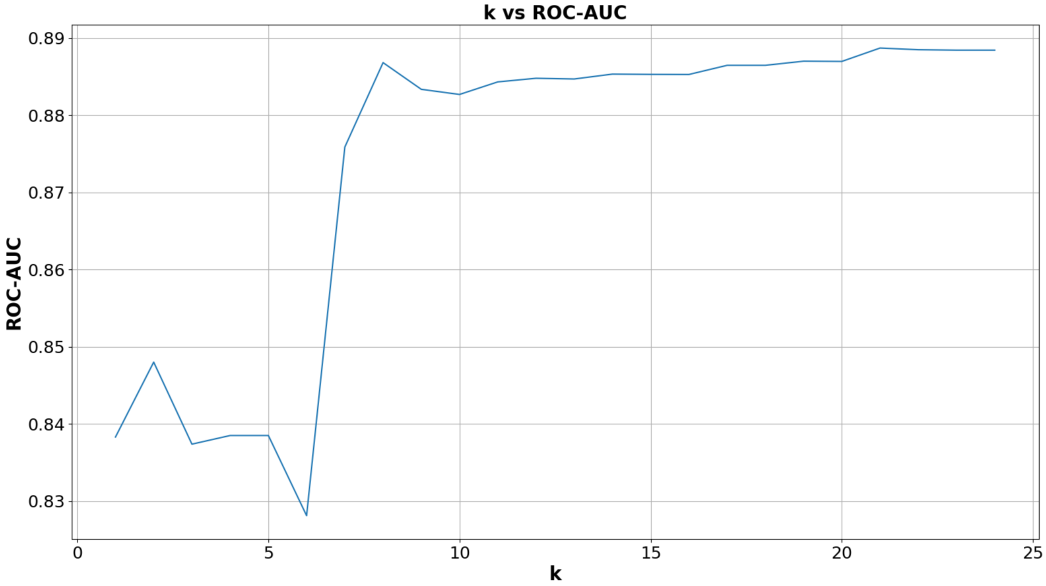
	precision	**recall**	**f1-score**	**support**
0	0.97	1.00	0.98	19,268
1	0.97	0.60	0.74	1,443
micro avg	0.97	0.97	0.97	20,711
macro avg	0.97	0.80	0.86	20,711
weighted avg	0.97	0.97	0.97	20,711


k vs Accuracy
Accuracy
0.970
0.968
0.966
0.964
0.962
0.960
k
0
5
10
15
20
25

(a)


k vs PR-AUC
PR-AUC
0.25
0.20
0.15
0.10
0.05
k
0
5
10
15
20
25

(b)



(c)

Figure 6: Dependency of various model performance metrics on k. (a) Accuracy, (b) PR-AUC, (c) ROC-AUC

Discussion

ClinVar is an example of many Biomedical databases and knowledge bases that incorporate data from various sources and acts as a platform for the community to share and use knowledge. However, as has already been noted, ClinVar does not curate any of the submitted information thus resulting in records of varying quality and granularity. Therefore, ClinVar is in need of assessment of the quality and trustworthiness of the information contained. There are several challenges involved in carrying out such an assessment which I stumbled across during my analyses.

Firstly, the lack of interoperability between disparate Biomedical sources is a known challenge that is being tackled by many researchers and organizations. Another challenge that I noticed while working with ClinVar was the lack of ”intraoperability”, that is the lack of consistency inside the database. Different releases had different ways of representing the clinical significance. The differences ranged from the simple-to-resolve case differences to more sophisticated cases when manual mapping was unavoidable.

The biggest challenge, however, is the noise in the negative dataset. If one aims to study reclassification efforts in ClinVar, there is strong need of a gold standard, a subset of records that are of sufficient quality and can serve as a negative dataset. This is a tricky challenge not only in this case but in a lot of classification problems and the choice of negatives can hugely impact the outcome of the analysis. This is an open research question and further efforts should be made to come up with a consensus on how to best select the negative dataset. It assumes differentiating between the records that have never been reclassified either because of sufficient quality or because of a lack of effort and/or evidence. It might be reasonable to use the records reviewed by an expert panel as the gold standard, however, in doing so one might introduce confounding factors that are unknown. The model might be unable to correctly label the records that are of sufficient quality but have not been review by an expert panel as such. This is a non-trivial question the answer to which I do not currently have and it would be interesting to try different ways of choosing the negatives and compare outcomes and estimates.

And last but not least, in building the model I have only utilized six features which were comparably straightforward to extract and represent. However, there are other types of information available for ClinVar records that can be utilized to obtain more accurate estimates. Some of that information is in an unstructured form, so NLP tools might be used to extract information from those free text descriptions.

Conclusion

To conclude, I extracted several features from ClinVar cancer-related variants and carried out exploratory analysis to find features that might be predictive for variant reclassifications and fit a model that can predict reclassification events. Exploratory analyses showed that the negative dataset is not reliable and contains a lot of false negatives which negatively affect the logistic model performance and pose a great challenge in trying to predict variants that are in need of reclassifications. I fit a logistic model using the 8 most predictive features and showed that the model performance is not desirable due to the noise in the negative data. Filtering some of that noise greatly improved model performance. Thus, a gold standard of ClinVar records not in need of reclassification is needed to build a model that can predict reclassification events.

Code

Code written for the analyses can be found on the following GitHub repository: https://github.com/AniKhachatryan/BINF-G4003.

References

1. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research. 2013 Nov 14;42(D1):D980-5.
2. Introduction - ClinVar - NCBI [Internet]. Ncbi.nlm.nih.gov. 2018 [cited 21 December 2018]. Available from: https://www.ncbi.nlm.nih.gov/clinvar/intro/
3. Yang S, Lincoln SE, Kobayashi Y, Nykamp K, Nussbaum RL, Topper S. Sources of discordance among germ-line

variant classifications in ClinVar. Genetics in Medicine. 2017 Oct;19(10):1118.

4. Shah N, Hou YC, Yu HC, Sainger R, Caskey CT, Venter JC, Telenti A. Identification of misclassified clinvar variants via disease population prevalence. The American Journal of Human Genetics. 2018 Apr 5;102(4):609-19.
5. Butler III RR, Gejman PV. Clinotator: analyzing ClinVar variation reports to prioritize reclassification efforts. F1000Research. 2018;7.
6. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K. 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. Genetics in medicine. 2015 May;17(5):405.
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825.