# Analyzing and Predicting the Phenotype of Aspirin Hypersensitivity

Brendan J. Lee

Johns Hopkins University | Whiting School of Engineering | Baltimore, MD

**Foundations of Computational Biology and Bioinformatics Spring 2021**

## Summary

This project aims to understand the genetic information related to the phenotype of aspirin hypersensitivity, and build machine learning models to predict this phenotype based on a person's genetics. The genetic variant data from 44 individuals (22 aspirin hypersensitive, 22 aspirin tolerant) were annotated in OpenCRAVAT and filtering was performed to find relevant variants. 265 significant variants were found using Fisher's exact test, many of which were linked to biological processes relevant to the phenotype. Then, variants from genes that are known to be linked to aspirin hypersensitivity were used to build machine learning models from the annotated data. Using 5-fold and 10-fold cross validation, the 6-nearest neighbors, SVM, and Naïve Bayes models performed the best, using 30 variants as features and dimensionality reduction by PCA to 6 features; the average AUC was ranged from 0.80 to 0.87 for all three classifiers. Finally, these three classifiers were used predict the phenotype of aspirin hypersensitivity for 23 external samples from Harvard PGP, with phenotypes directly linked to aspirin hypersensitivity (asthma). Given the accepted estimated percentages of people with the phenotype of aspirin hypersensitivity, all three classifiers predicted a reasonable number of individuals with this phenotype.

## Background

Aspirin is a very common non-steroidal anti-inflammatory drug (NSAID), used for treating inflammation-related conditions and as a blood thinner. About 1-3% of people have an allergy to this drug, although this estimated proportion has increased over time as many people are unaware that they have this allergy. This often involves the development of persistent asthma symptoms after ingesting this drug, or aspirin-induced respiratory disease (AERD); it is estimated that about 9% of asthmatic patients have AERD. Other common reactions to this drug include nasal polyps and skin hives. In this study, the phenotype of aspirin hypersensitivity is predicted from individuals' genetic data; this way, people will be able to know if they are at risk for developing adverse reactions to aspirin, and adverse events will be prevented if at-risk individuals do not take the drug as a result of knowing about this prediction. Several variants, especially those related to the inflammation-producing pathway affected by aspirin's inhibition of the COX enzyme, are known to affect the likelihood of this phenotype; this study identifies these variants and discovers new ones.

## Methods

- 44 23andMe samples, 22 with the phenotype and 22 without the phenotype, were downloaded from OpenSNP
- The samples were annotated in OpenCRAVAT, using the Gene Ontology, PharmGKB, and GTEx eQTLs annotators to filter for relevant variants, related to the immune, respiratory, or integumentary systems or are linked to reactions to aspirin
- R's fisher.test function was used to find variants with sample counts that differ significantly between the two sample groups; an analysis of the biological processes linked to these variants was performed
- Variants from literature and from genes linked to this phenotype were used as features to train binary phenotype classifiers with these samples in Python, SVM, Naïve Bayes, Random Forest, K-Nearest Neighbors with optimized K of K = 6, and a risk-score thresholding used by Chang et. al. The number of variants used to build these models was varied from 5 to 30, and R's Exponential Family PCA (logsvd function) was used before training each model except the model from Chang et. al. The performance of these models was evaluated via 5-fold and 10-fold cross validation, in which average AUC was calculated.
- 23 samples from individuals with asthma were downloaded from HarvardPGP, since this database does not include the phenotype of aspirin allergy. Using the best performing classifiers as well as the PCA model from the previous step, each sample was predicted as having the phenotype or not having it. This way, the proportion of people with asthma who have AERD could be estimated.

## Results

### Finding Significant Variants with Fisher's Exact Test

- For variants that were present in at least 8 samples (20% of total samples), **265 variants** with a Fisher's Exact Test p-value less than 0.05 were found
- Only a 5 of these significant variants were found to be associated with drug interactions (PharmGKB); the drugs that were found were not the same among variants or related to each other, none associated with aspirin or other NSAIDs
- More information on the significant variants are given in Tables 1-4

**Table1—Most Common Gene Ontology Biological Processes and Significant Variant Counts**

| | |
|---|---|
| immune response | 45 |
| antigen processing and presentation | 22 |
| endocytosis | 13 |
| innate immune response | 12 |
| adaptive immune response | 12 |

**Table 2—Most Common Gene Ontology Molecular Functions and Significant Variant Counts**

| | |
|---|---|
| protein binding | 23 |
| ATP binding | 16 |
| metal ion binding | 14 |
| identical protein binding | 9 |
| nucleic acid binding | 8 |

**Table 3—Most Common GTEX eQTLs Tissue Types and Variant Counts**

| | |
|---|---|
| Nerve_Tibial | 106 |
| Esophagus_Mucosa | 103 |
| Lung | 102 |
| Thyroid | 100 |
| Adipose_Subcutaneous | 95 |

**Table 4– Most Common Genes and Variant Counts**

| | |
|---|---|
| HLA-B | 19 |
| ENPP2 | 7 |
| APP | 6 |
| BTK | 4 |
| IFI16 | 4 |

### Training and Testing Phenotype Predictors

- The only variant from literature that was present in the samples, and therefore used to build these models, was the well-known LTC4S -444A/C polymorphism. All other variants used to build these models were selected by taking variants that were present in many of the samples that were on genes known to play a role in aspirin hypersensitivity (ALOX5, HLA-DPB1, TBXA2R, and others)
- For Exponential Family PCA, a type of PCA for binary data, **the lowest number of components that resulted in an explained variance greater than 95% was 3 when 5 variants were used, 4 for 10 variants, 5 for 15 variants, and 6 for 20, 25, and 30 variants**

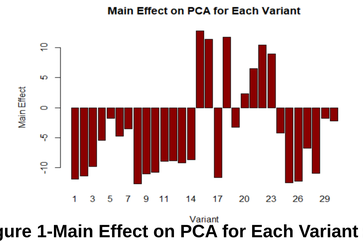


**Figure 1-Main Effect on PCA for Each Variant (all 30 included)**


**Figure 2 (top), Figure 3 (bottom)—Average AUC vs. Number of Variants for all 5 phenotype predictors; Figure 2 is for 5-fold cross validation, while Figure 3 is for 10-fold**

### Predicting the Phenotype of Asthmatic Individuals from Harvard PGP

- Using 6-Nearest Neighbors trained with the data from the original samples from OpenSNP, **5 out of 23 of the PGP samples** were predicted to have the phenotype of aspirin allergy. On the other hand, both SVM and Naive Bayes predicted **4 out of 23 PGP samples** to have the phenotype, where the 4 samples were the same for both models and were also the same ones predicted by 6-Nearest Neighbors.

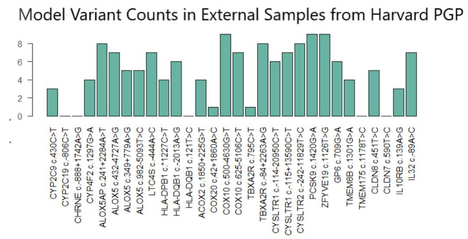


**Figure 4-Incidence of the 30 model variants in the Harvard PGP Samples**

## Discussion and Implications

- Because a large number of significant variants were found, the results demonstrate the polygenic nature of the phenotype of aspirin hypersensitivity. However, many significant variants could have been falsely determined because of the small sample size. This experiment should be repeated with a much larger sample size, such as 100 to 200 samples, to improve the credibility of the results.
- The absense of variants related to aspirin reactions via PharmGKB was likely because the list of variants in PharmGKB is incomplete, since literature variants linked to the phenotype were not found in this list
- Table 1: Verifies the immune system's role in causing an adverse reaction to aspirin.
- Table 2: Supports the widely-accepted idea that adverse reactions to aspirin are brought about by variants that encode for enzymes and receptors, since they involve binding of molecules
- Table 3: Respiratory tissues are affected, as expected, but the other tissues detected should be further investigated
- Table 4: The low variant counts per gene suggest the complexity of the phenotype

- The lack of presence of many literature variants in the samples suggests that the 23andMe files might be incomplete
- The success of PCA and the high main effect values for the individual variants suggests that variants are significantly correlated with each other in determining whether or not a person has the phenotype
- 6-nearest neighbors was overall the best classifier, and SVM and Naïve Bayes also performed well for all numbers of variants. This is because the 2 classes of data are well-separated from each other via decision boundary in the PCA space. The high AUC values mean that the predictors perform well.
- Risk score thresholding classification performed poorly compared to all other classes at small numbers of variants (5 and 10 variants). This further suggests that variants are significantly correlated to one another, since PCA improves model accuracy for small numbers of features.
- This experiment should be repeated with a larger sample size and training the models with the same variants to ensure the effectiveness of these 30 variants in predicting the phenotype

- Presence of model variants in external samples suggests that the phenotype of interest can be determined using the predictors
- 6-Nearest Neighbors, SVM, and Naïve Bayes predicted 21.7%, 17.4%, and 17.4%, respectively, of asthmatic individuals from PGP to be aspirin hypersensitive. These results are reasonable, given that the known proportion of asthmatic patients with AERD is 9%, and is likely higher.
- Therefore, the performance of these predictors is promising, but I cannot be sure that this will be an accurate predictor for all samples, due to the small sample size of both the OpenSNP and PGP samples.

## References


- Chang, H., Shin, S., Lee, T. *et al.* Development of a genetic marker set to diagnose aspirin-exacerbated respiratory disease in a genome-wide association study. *Pharmacogenomics J* **15,** 316–321 (2015). https://doi.org/10.1038/tpj.2014.78
- Palikhe et. al. "Genetic Mechanisms in Aspirin-Exacerbated Respiratory Disease". Journal of Allergy, 2012. Hindawi Publishing Corporation. https://doi.org/10.1155/2012/794890
- Kim, Seung-Hyun, and Hae-Sim Park. "Genetic markers for differentiating aspirin-hypersensitivity." *Yonsei medical journal* vol. 47,1 (2006): 15-21. doi:10.3349/ymj.2006.47.1.15