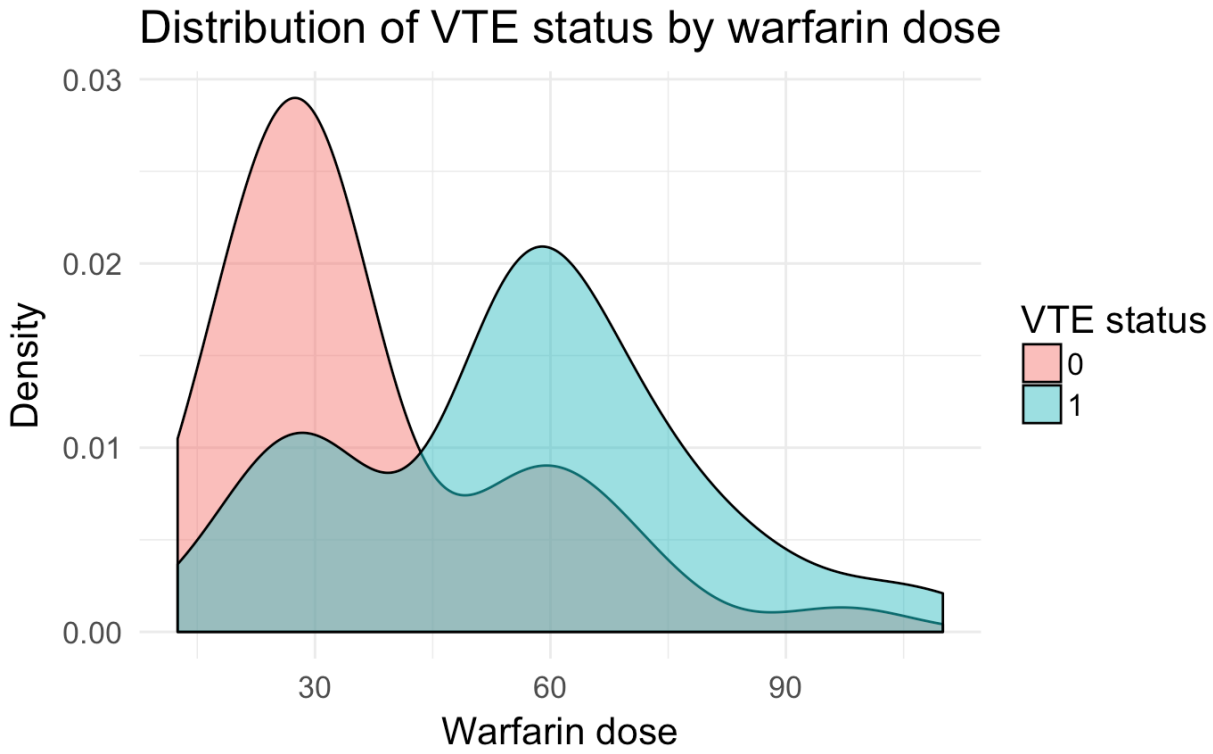
Supplementary information for “Predicting venous thromboembolism risk from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges”



Supplementary Figure S1: Distribution of warfarin dose by VTE status. In this dataset warfarin dosage is significantly associated with VTE status, as discussed in Daneshjou et al. 2016. Patients with VTE are more likely to be on a higher dose of warfarin than those taking warfarin for a different indication. The distribution of warfarin dosage for patients with VTE is shown in blue, and the distribution of warfarin dosage for patients taking warfarin for a different indication are shown in red.

Participant summaries

The remainder of the text in this document is from the participants describing their methodology for approaching the prediction challenge. Each summary is followed by a figure showing different visualizations of the predictions submitted by the preceding group.

Group 1

Panagiotis Katsonis[1], Olivier Lichtarge[1-4]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine.
[2]Department of Biochemistry & Molecular Biology, Baylor College of Medicine.
[3]Department of Pharmacology, Baylor College of Medicine.

[4]Computational and Integrative Biomedical Research Center, Baylor College of Medicine.

Evolutionary Action (EA) burden on known disease-associated genes

In order to separate individuals with venous thromboembolism (VTE) from those with atrial fibrillation (AF), we predicted the fitness effect of the genetic variants in genes associated with VTE and with AF.

We used the ***Evolutionary Action (EA)*** method to predict the fitness effect of the genetic variants (Katsonis and Lichtarge 2014). EA does not involve any training, because it relies on a formal equation of the genotype-phenotype relationship. The terms of this equation were m calculated using protein homology data. Briefly, the EA equation states that the fitness effect of a mutation equals the product of the sensitivity of the mutated position with the magnitude of the change. The sensitivity of the position is calculated by quantifying the correlation of the residue variations with phylogenetic branching within an alignment of homologous sequences (Lichtarge et al. 1996; Mihalek et al. 2004; Lichtarge and Wilkins 2010). The magnitude of the change is calculated from substitution likelihood according to numerous sequence alignments for the given context (strata of sensitivity of the position, and optionally additional stratification based on structural features). The calculated product is then normalized to represent the percentile rank of each variant within the protein in the scale of 0 (benign) to 100 (pathogenic). The EA scores are available for all human variants at: http://mammoth.bcm.tmc.edu/EvolutionaryAction

We used the ***DisGeNET*** platform to identify genes associated with VTE and genes associated with AF (Pinero et al. 2017). For each disease, DisGeNET provides a list of genes scored with an index that represents the confidence of association. High scores correspond to more reliable associations, therefore, we only used scores of 0.1 or above to avoid false positives. For VTE we found 8 genes scored above 0.1 (*F5, F2, FGA, PROC, PLAT, SERPINC1, TNF,* and *SERPIND1*), while for AF we found 38 genes (*SCN5A, KCNE2, HCN4, NKX2-5, ACE, GJA5, KCNQ1, NOS3, KCNA5, LMNA, NPPA, ZFHX3, KCNN3, VWF, NPPB, PRKAG2, NUP155, SELE, CAV1, SCN10A, MYH7, ANK2, SOX5, HTR4, SYNE2, PLN, C9orf3, PRRX1, CAV2, CACNA1C, WNT8A, EDN1, CACNB2, SMAD3, TNNI3K, TAB2, DTNA, DES*).

To calculate the ***likelihood of each individual to have VTE***, we used the ratio of the fitness effect on VTE genes over the fitness effect of VTE and AF genes. For one individual, the fitness effect on one gene (EA_{gene}) was defined to be equal to: 0 if there was no mutation, EA/100 if there was one mutation, or $1-\prod(1-EA_i/100)$ if there were multiple mutations ($\prod$ indicates the product for all mutations i). For synonymous variants EA was 0 and for nonsense variants EA was 100. To account for the strength of the association of a gene to disease we calculated a weighting factor for each gene based on the DisGeNET score ($S_{DisGeNET}$), as: $w_{gene}=w_{GI}\cdot(S_{DisGeNET}-0.1)$, where w_{GI} represents the ability of the genes to tolerate variations. For submission 1 we calculated w_{GI} as a percentile rank of the genes based on the average EA score of the variants seen in gnomAD (Lek et al. 2016), while in submission 2 we used $w_{GI}=1$ for all genes, since the DisGeNET scores may already account for this effect. Then, we calculated the fitness effect on VTE genes (EA_{VTE}) and on AF genes (EA_{AF}) as the sum of $w_{gene}\cdot EA_{gene}$, respectively. To normalize the two fitness effect scores, we calculated the factor: r= average(EA_{AF})/average(EA_{VTE}), based on the average values for all individuals (this normalization assumes that the number of VTE patients is about equal to the number of AF patients). The probability (p) of an individual to have VTE was finally calculated as: $p=r\cdot EA_{VTE}/(r\cdot EA_{VTE}+EA_{AF})$.

Katsonis P, Lichtarge O. 2014. A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome research* **24**: 2050-2058.

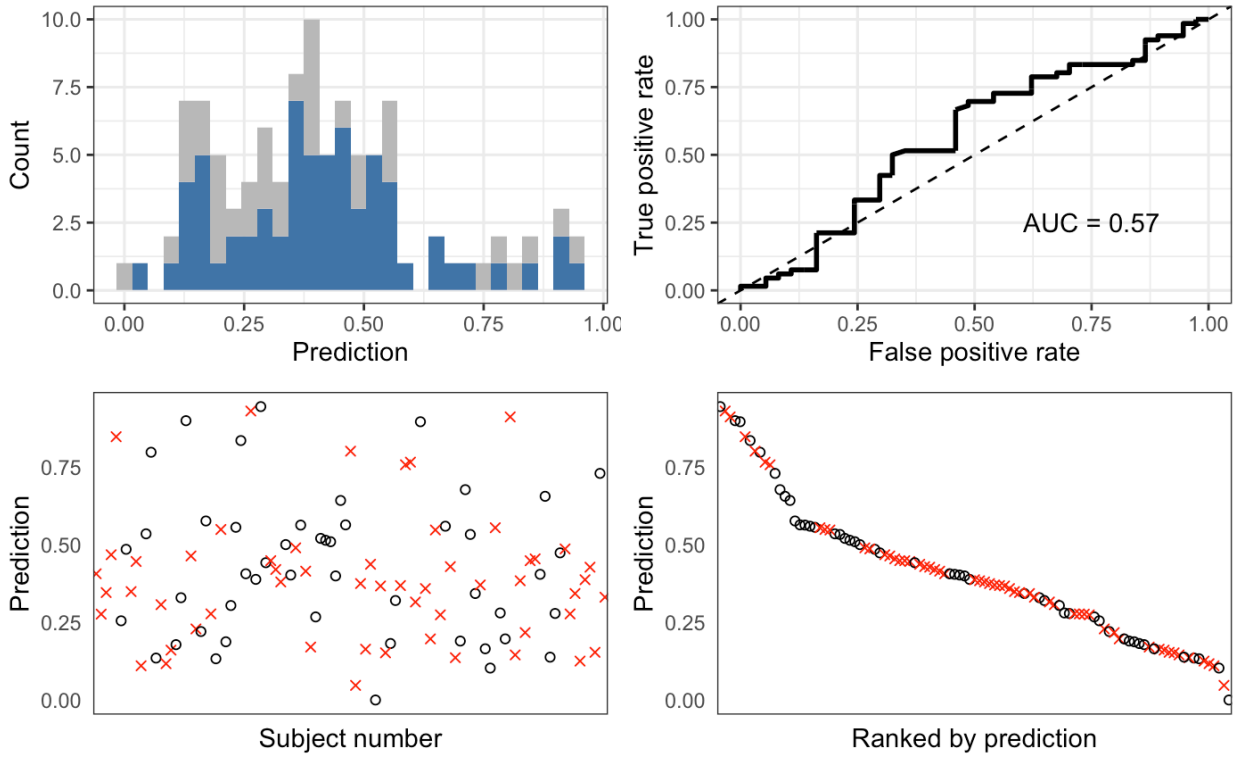
Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285-291.

Lichtarge O, Bourne H, Cohen F. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* **257**: 342 - 358.

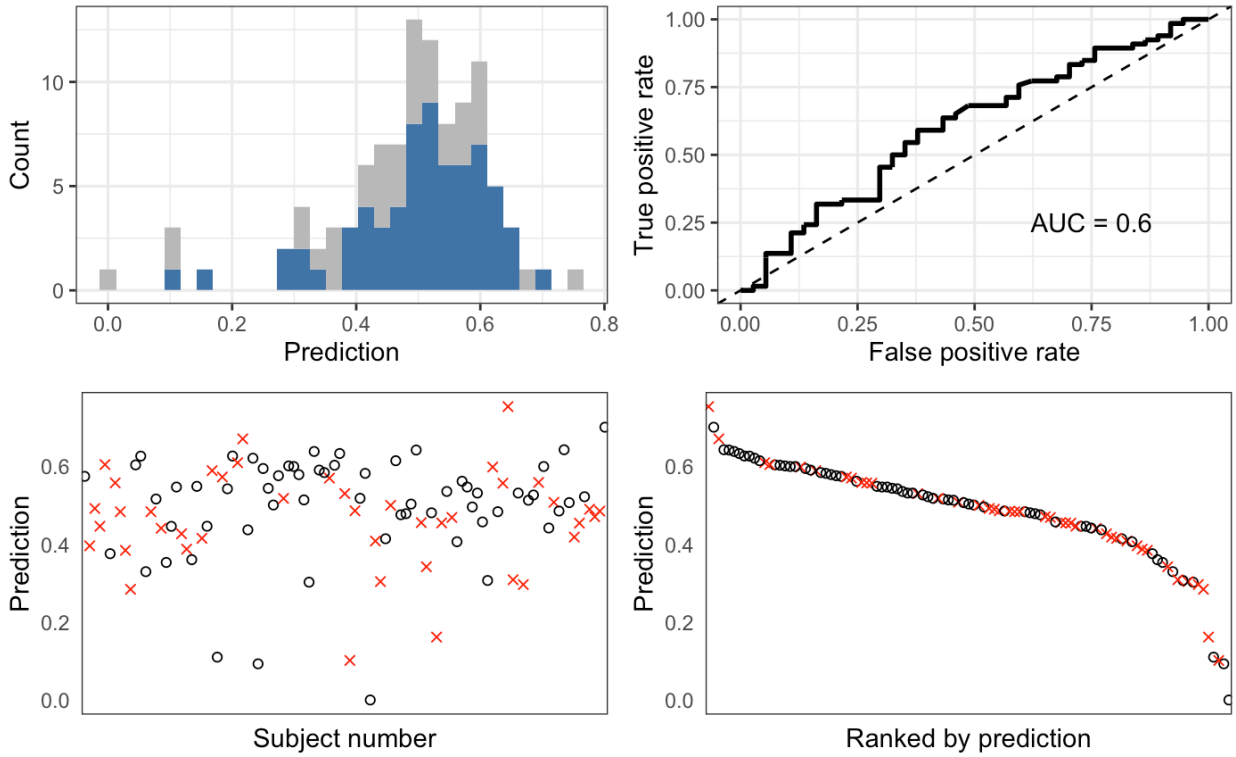
Lichtarge O, Wilkins A. 2010. Evolution: a guide to perturb protein function and networks. *Curr Opin Struct Biol* **20**: 351-359.

Mihalek I, Res I, Lichtarge O. 2004. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol* **336**: 1265 - 1282.

Pinero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E, Garcia-Garcia J, Sanz F, Furlong LI. 2017. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* **45**: D833-D839.



Supplementary Figure S2: Evaluation plots for the first submission from group 1. The top left plot shows a histogram of the predicted scores ranging from 0 to 1, with a 1 indicating the highest probability of VTE. Subjects with VTE are shown in blue, and those without are shown in gray. The top right plot shows a receiver operator characteristic (ROC) curve with the area under the ROC curve (AUC) of the method indicated on the plot, the bottom left plot shows the predicted scores of each subject in order of subject id, the bottom right plot shows the predicted scores for each subject sorted by the predicted score in descending order. On both the bottom plots the color and shape indicate correct and incorrect predictions, with correct predictions represented as black circles, and incorrect predictions as red X's. Supplementary figures 2-15 all show the same plots with the values changing based on the submission.



Supplementary Figure S3: Evaluation plots for the second submission from group 1. Refer to supplementary figure 2 for a description of the plots.

Group 2

Raj Gopal Srinivasan[1], Sadhna Rana[1]

[1]Innovations Labs, Tata Consultancy Services, Hyderabad, India

Variant annotation and scoring is done using Varant [1] program and an in-house variant prioritization program (unpublished) respectively. The variant prioritization tool integrates predictions from around twenty functional predictors. Each variant score varies from 0-1 (1 being the highest likelihood of pathogenicity). We did not apply any filter based on genomic regions to select variants, thus our analysis included intergenic, intronic, exonic and canonical splice variants. Intergenic, intronic and UTR variants are scored based on CADD [2], DANN [3] and regulome annotation [4]. Coding exonic variants are scored based on CADD, DANN, Polyphen2 [5], SIFT [6], FATHMM [1], MutationAssessor [7], MutationTaster [8], Provean [9], LRT [10], and Meta_SVM [11]. Conservation annotations used to score the variants are as follows: GERP++ [12], PhastCons [13] and SiPhy [14]. We also used MAF filter to select population specific common variants in African Americans. Thus common variants (MAF > 0.03) from African American population and all rare variants (MAF <=0.02) from all populations were selected for further analysis. Only variants with genotype quality greater than 20 are taken up for the analysis. Further variants are filtered based on list of genes known to be

involved in venous thrombosis. Genes are compiled from online Mendelian inheritance in Man (OMIM) [15], Human phenotype ontology (HPO) [16], genome wide association studies (GWAS) studies and review articles on clotting disorders. Natural inhibitors of clotting proteins i.e anticoagulant proteins are scored highest as normal functioning of these genes is necessary to prevent thromboembolism [17]. Proteins in the coagulation pathways associated with venous thrombosis and well established genes from literature [17] are given next best score of 0.67. These high scoring genes are also part of databases like OMIM, Clinvar [18] and HPO, and are annotated as, “deep venous thromboembolism”, “venous thrombosis” or “thrombophilia” causing genes. Besides the well-established genes in venous thrombosis, other genes associated with venous thrombosis from HPO database or literature are given a score of 0.33. The genes associated with venous thrombosis from GWAS studies and associated with thrombosis risk as given in Genetics Home Reference[15] are given a score of 0.17. A variant score cutoff of 0.4 is used to select likely pathogenic variants in the genes associated with thrombosis (Table 1). For each variant, the variant score is multiplied by the gene relevance score to get a final variant score. For each sample, total sample score is calculated as sum of all the variant scores. The sample score is normalized by dividing it with the total number of variants in the sample. There are certain gene combinations that increase the risk of venous thrombosis manifold [19]. We observed that *F2* and *F5* gene combination occurred very frequently in samples hence an additional gene combination score of 0.25 is added to such samples. It is presumed that *F2* and *F5* gene combination could potentially increase the likelihood of venous thrombosis in such samples. The variant scores range from 0-0.611. Samples with score greater than 0.200 are more likely to be venous thrombosis patients. This cutoff was based on number of high scoring genes in the sample. Samples with sample score less than 0.2 had genes with weak association (low gene scores) with venous thrombosis.

Table 1: Genes scored based on their relevance to venous thrombosis.

Gene	Database (Disease)	Gene score
PROC	OMIM (Thrombophilia), HPO (Deep Venous Thrombosis)	1
PROS1	OMIM (Thrombophilia), HPO (Deep Venous Thrombosis)	1
SERPINA10	OMIM (Venous Thrombosis)	1
SERPINC1	OMIM, HPO (Deep Venous Thrombosis)	1
SERPIND1	OMIM (Thrombophilia)	1
THBD	OMIM, HPO (Deep Venous Thrombosis)	1
ABO	PubMed [17, 20] (Well established venous thrombosis associated gene)	0.67
AKT1	OMIM, HPO (Deep Venous Thrombosis)	0.67
F13A1	OMIM (protection Venous Thrombosis)	0.67
F13B	Clinvar (Deep Venous Thrombosis)	0.67
F2	OMIM (Thrombophilia), HPO (Deep Venous Thrombosis)	0.67
F5	OMIM, HPO (Deep Venous Thrombosis)	0.67
F7	OMIM VT: Genes of Coagulation pathways"	0.67
F8	OMIM VT Genes of Coagulation pathways"	0.67
F9	OMIM, HPO (Deep Venous Thrombosis)	0.67
FCH2	OMIM (thrombophilia)	0.67
FGA	HPO (Venous Thrombosis)	0.67

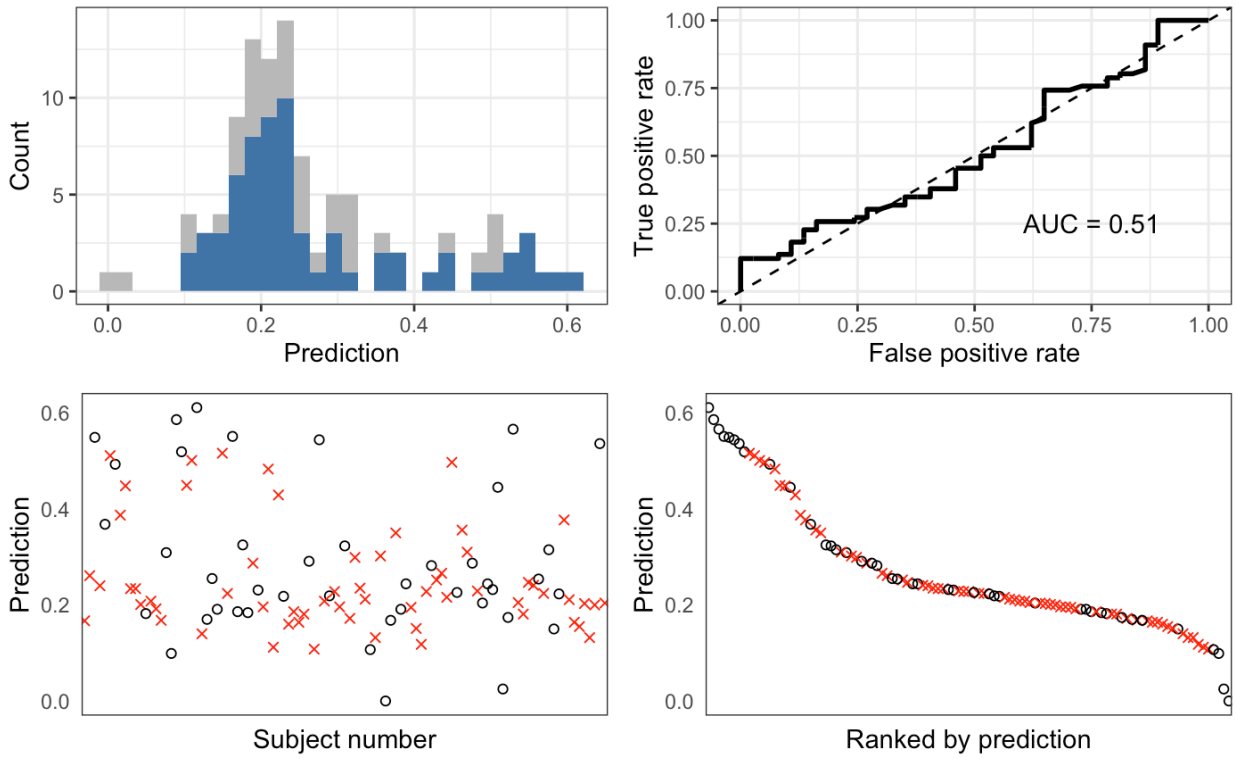
Gene	Database (Disease)	Gene score
FGG	HPO (Venous Thrombosis)	0.67
GP1BA	OMIM (Thrombosis)	0.67
HABP2	OMIM (VT susceptibility)	0.67
HRG	OMIM (Thrombophilia), HPO (Abnormal Thrombosis)	0.67
MTHFR	OMIM (Thromboembolism)	0.67
PIEZO1	(Calcium signaling)	0.67
PLAT	OMIM (Thrombophilia)	0.67
ACVRL1	HPO (Venous thrombosis)	0.33
ADAMTS13	PubMed [21]	0.33
AGGF1	HPO (Venous thrombosis)	0.33
C4A	HPO (Venous thrombosis)	0.33
CALR	HPO (Venous thrombosis)	0.33
CBS	HPO (Venous thrombosis), metabolic protein with coagulation phenotypes"	0.33
CCR1	HPO (Venous thrombosis)	0.33
CPB2	PubMed [22]	0.33
CTLA4	HPO (Venous thrombosis)	0.33
ENG	HPO (Venous thrombosis)	0.33
EPOR	HPO (Venous thrombosis)	0.33
ERAP1	HPO (Venous thrombosis)	0.33
F12	GWAS, coagulation pathway	0.33
FAS	HPO (Venous thrombosis)	0.33
GDF2	HPO (Venous thrombosis)	0.33
GNAQ	HPO (Venous thrombosis)	0.33
HBB	HPO (Venous thrombosis)	0.33
HLA-B	HPO (Venous thrombosis)	0.33
HLA-DPB1	HPO (Venous thrombosis)	0.33
IDH1	HPO (Venous thrombosis)	0.33
IDH2	HPO (Venous thrombosis)	0.33
IL10	HPO (Venous thrombosis)	0.33
IL12A	HPO (Venous thrombosis)	0.33
IL12A-AS1	HPO (Venous thrombosis)	0.33
IL23R	HPO (Venous thrombosis)	0.33
ITGB3	PUBMED [23]	0.33
JAK2	HPO (Venous thrombosis)	0.33
KLRC4	HPO (Venous thrombosis)	0.33
KNG1	GWAS and fibrinolysis and kallikrein pathways	0.33
MEFV	HPO (Venous thrombosis)	0.33
MPL	HPO (Venous thrombosis)	0.33
PDGFRA	HPO (Venous thrombosis)	0.33
PROCR	PubMed [24]	0.33
PROZ	PubMed [25]	0.33
PRTN3	HPO (Venous thrombosis)	0.33
PTH1R	HPO (Venous thrombosis)	0.33
PTPN22	HPO (Venous thrombosis)	0.33
SERPINA1	PubMed [26]	0.33
SERPINA5	PubMed [26]	0.33
SERPINE1	PubMed [26]	0.33
SERPINF2	PubMed [26]	0.33
SERPING1	PubMed [26]	0.33
SH2B3	HPO (Venous thrombosis)	0.33
SMAD4	HPO (Venous thrombosis)	0.33
SPINT2	UniProt (inhibits tissue kallikrein, and factor F11)	0.33
STAT4	HPO (Venous thrombosis)	0.33
THBS1	PubMed [27]	0.33
TLR4	HPO (Venous thrombosis)	0.33
TP53	HPO (Venous thrombosis)	0.33
UBAC2	HPO (Venous thrombosis)	0.33
USP8	HPO (Venous thrombosis)	0.33
VTN	PubMed [26]	0.33
BMPR2	Abnormal thrombosis	0.17
C4BPA	GWAS	0.17

Gene	Database (Disease)	Gene score
C4BPB	GWAS	0.17
CD59	Genetics Home Reference (increased risk thrombosis)	0.17
CYP4V2	GWAS	0.17
F11	GWAS	0.17
FGB	Genetics Home Reference (increased risk)	0.17
GP6	GWAS	0.17
HIVEP1	GWAS	0.17
KLKB1	Genetics Home Reference (increased risk)	0.17
LEMD3	GWAS	0.17
LY86	GWAS	0.17
MYH9	Abnormal thrombosis	0.17
PIGA	Genetics Home Reference (increased risk)	0.17
PIK3CA	Genetics Home Reference (increased risk)	0.17
PLA2G7	metabolic protein with coagulation phenotype (platelet activation)	0.17
PRSS1	HPO (Abnormal thrombosis)	0.17
RFT1	HPO (Abnormal thrombosis)	0.17
SLC2A10	HPO (Abnormal thrombosis)	0.17
SPINK1	HPO (Abnormal thrombosis)	0.17
STX2	GWAS	0.17
STXBP5	GWAS "	0.17
TC2N	GWAS	0.17
TET2	Genetics Home Reference (increased risk)	0.17
THPO	Genetic Home Reference (increased risk thrombosis)	0.17
VWF	GWAS	0.17

References:

1. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR, Campbell C: **An integrative approach to predicting the functional effects of non-coding and coding sequence variation**. *Bioinformatics* 2015, **31**(10):1536-1543.
2. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: **A general framework for estimating the relative pathogenicity of human genetic variants**. *Nat Genet* 2014, **46**(3):310-315.
3. Quang D, Chen Y, Xie X: **DANN: a deep learning approach for annotating the pathogenicity of genetic variants**. *Bioinformatics* 2015, **31**(5):761-763.
4. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S *et al*: **Annotation of functional variation in personal genomes using RegulomeDB**. *Genome Res* 2012, **22**(9):1790-1797.
5. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**(4):248-249.
6. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions**. *Genome Res* 2001, **11**(5):863-874.
7. Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics**. *Nucleic Acids Res* 2011, **39**(17):e118.
8. Schwarz JM, Cooper DN, Schuelke M, Seelow D: **MutationTaster2: mutation prediction for the deep-sequencing age**. *Nat Methods* 2014, **11**(4):361-362.
9. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP: **Predicting the functional effect of amino acid substitutions and indels**. *PLoS One* 2012, **7**(10):e46688.
10. Chun S, Fay JC: **Identification of deleterious mutations within three human genomes**. *Genome Res* 2009, **19**(9):1553-1561.

11. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X: **Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies**. *Hum Mol Genet* 2015, **24**(8):2125-2137.
12. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S: **Identifying a high fraction of the human genome to be under selective constraint using GERP++**. *PLoS Comput Biol* 2010, **6**(12):e1001025.
13. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al*: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res* 2005, **15**(8):1034-1050.
14. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X: **Identifying novel constrained elements by exploiting biased substitution patterns**. *Bioinformatics* 2009, **25**(12):i54-62.
15. **https://ghr.nlm.nih.gov/**. In.
16. Kohler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J *et al*: **The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data**. *Nucleic Acids Res* 2014, **42**(Database issue):D966-974.
17. Rosendaal FR, Reitsma PH: **Genetics of venous thrombosis**. *J Thromb Haemost* 2009, **7 Suppl 1**:301-304.
18. 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 Res* 2014, **42**(Database issue):D980-985.
19. Hasstedt SJ, Bovill EG, Callas PW, Long GL: **An unknown genetic defect increases venous thrombosis risk, through interaction with protein C deficiency**. *Am J Hum Genet* 1998, **63**(2):569-576.
20. Huang SS, Liu Y, Jing ZC, Wang XJ, Mao YM: **Common genetic risk factors of venous thromboembolism in Western and Asian populations**. *Genet Mol Res* 2016, **15**(1):15017644.
21. Zheng XL: **ADAMTS13 and von Willebrand factor in thrombotic thrombocytopenic purpura**. *Annu Rev Med* 2015, **66**:211-225.
22. Reiner AP, Lange LA, Smith NL, Zakai NA, Cushman M, Folsom AR: **Common hemostasis and inflammation gene variants and venous thrombosis in older adults from the Cardiovascular Health Study**. *J Thromb Haemost* 2009, **7**(9):1499-1505.
23. Bianconi D, Schuler A, Pausz C, Geroldinger A, Kaider A, Lenz HJ, Kornek G, Scheithauer W, Zielinski CC, Pabinger I *et al*: **Integrin beta-3 genetic variants and risk of venous thromboembolism in colorectal cancer patients**. *Thromb Res* 2015, **136**(5):865-869.
24. Dennis J, Johnson CY, Adediran AS, de Andrade M, Heit JA, Morange PE, Tregouet DA, Gagnon F: **The endothelial protein C receptor (PROCR) Ser219Gly variant and risk of common thrombotic disorders: a HuGE review and meta-analysis of evidence from observational studies**. *Blood* 2012, **119**(10):2392-2400.
25. Bafunno V, Santacroce R, Margaglione M: **The risk of occurrence of venous thrombosis: focus on protein Z**. *Thromb Res* 2011, **128**(6):508-515.
26. Rau JC, Beaulieu LM, Huntington JA, Church FC: **Serpins in thrombosis, hemostasis and fibrinolysis**. *J Thromb Haemost* 2007, **5 Suppl 1**:102-115.
27. Hansen GA, Vorum H, Jacobsen C, Honore B: **Calumenin but not reticulocalbin forms a Ca2+-dependent complex with thrombospondin-1. A potential role in haemostasis and thrombosis**. *Mol Cell Biochem* 2009, **320**(1-2):25-33.



Supplementary Figure S4: Evaluation plots for the submission from group 2. Refer to supplementary figure 2 for a description of the plots.

Group 3

Predrag Radivojac[1], Sean D Mooney[2], Kymberleigh A Pagel[3], Moses Stamboulian[3], Yuxiang Jiang[3]

[1]Northeastern University, Boston, MA, USA

[2]University of Washington, Seattle, WA, USA

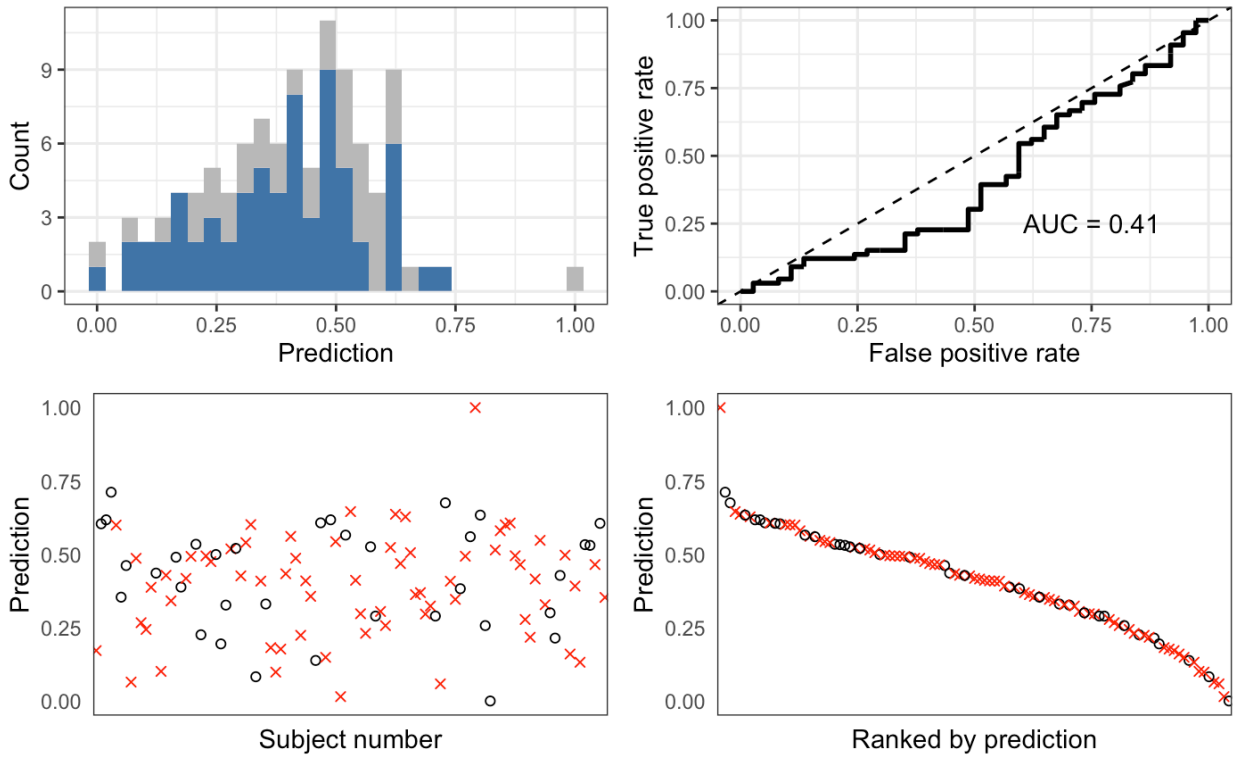
[3]Indiana University, Bloomington, IN, USA

The submitted predictions are derived from a combination of variant pathogenicity scores in relevant genes and the relationship of each clinical covariate to the phenotype. Annotation of the protein coding variation in the raw VCF files was performed using ANNOVAR. We assign pathogenicity prediction scores to missense and stop gain variants with MutPred2 and MutPred-LOF, respectively. Per exome, we include only the variant with the highest pathogenicity prediction score within each gene in further analyses. Confirmed risk genes are used as “seed” genes on the human protein-protein interaction network for running a network propagation algorithm. The propagation algorithm are performed in a 5-fold cross validation manner to get an initial score between [0, 1] for all the genes. We then use the AlphaMax algorithm to estimate the positive proportion of the risk genes and calibrate those initial scores to be proper probability scores measuring the likelihood of a gene being associated with the disease.

We generate a beta distribution based upon the MutPred scores of variants within the top one hundred highest scored genes for each phenotype. For each exome, we utilize the distribution to determine the p-value for the highest MutPred-scored variant within each gene. Next, we sought to incorporate the clinical covariates within a similar framework. For each clinical covariate, we search the published literature to find the mean and standard deviation values of the trait described in case/control studies. We utilize these variables from the literature to derive value distributions (binomial for gender and aspirin, Gaussian otherwise) that were used to derive p-values for each individual based upon their particular value for that covariate. The unnormalized score is the product of all gene and covariate scores, where each individual has scores for both VTE and atrial fibrillation. We then combine the VTE and atrial fibrillation score rankings using geometric mean, then transform with min-max normalization so that the values range between zero and one. The procedure is repeated one hundred times with differing amounts of seed genes (from 200 to 300), where the score for an individual is the mean score of the one hundred iterations.

Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam H, Mort M, Cooper DN, Sebat J, Iakoucheva LM, Mooney SD, Radivojac P. MutPred2: inferring the molecular and phenotypic impact of amino acid variants. *bioRxiv* 134981; doi: https://doi.org/10.1101/134981.

Kymberleigh A. Pagel, Vikas Pejaver, Guan Ning Lin, Hyun-Jun Nam, Matthew Mort, David N. Cooper, Jonathan Sebat, Lilia M. Iakoucheva, Sean D. Mooney, Predrag Radivojac; When loss-of-function is loss of function: assessing mutational signatures and impact of loss-of-function genetic variants. Bioinformatics 2017; 33 (14): i389-i398. doi: 10.1093/bioinformatics/btx272



Supplementary Figure S5: Evaluation plots for the submission from group 3. Refer to supplementary figure 2 for a description of the plots.

Group 4

BioFolD predictions of Clotting Disease from exome data.

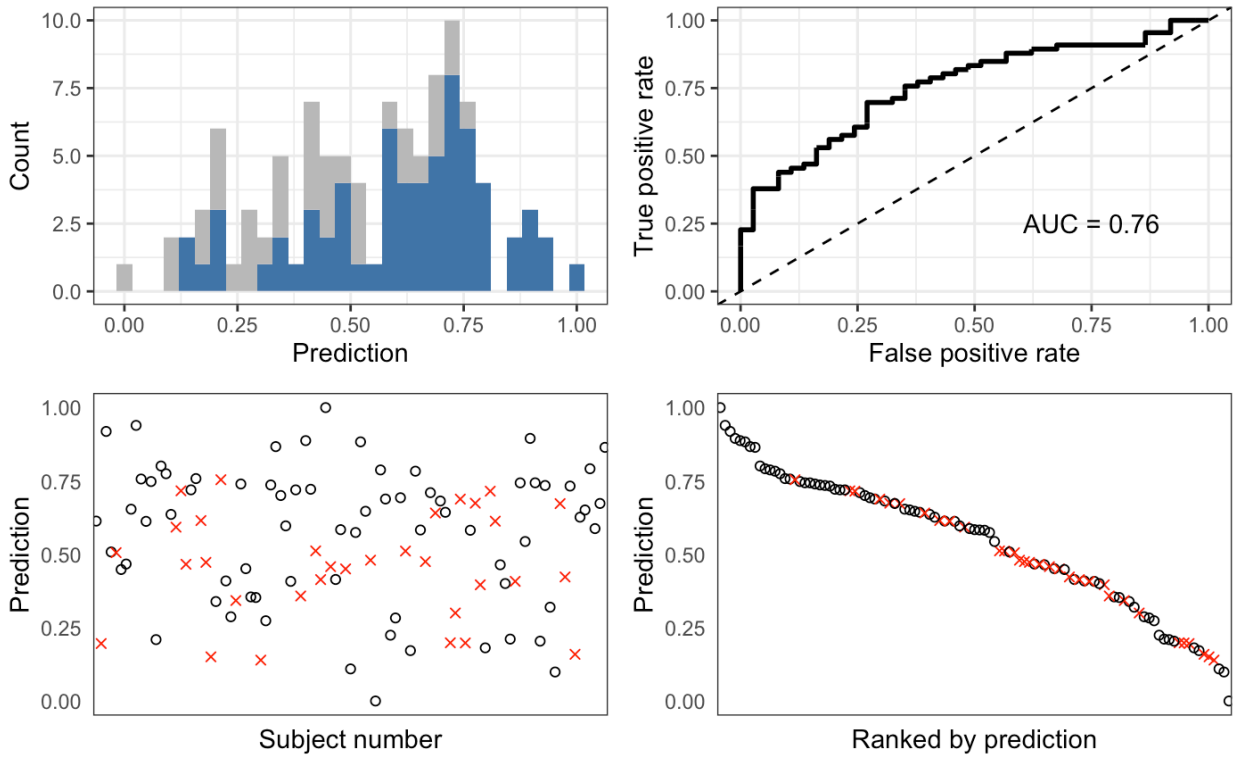
Emidio Capriotti
BioFolD Unit, Department of Pharmacy and Biotecnology (FaBiT), University of Bologna, Via Selmi 3, 40126 Bologna, Italy.

Our main hypothesis behind the predictions for the Clotting Disease challenge is that patients with higher Warfarin dose have higher incidence of venous thromboembolisms (VTE). Thus, we developed a linear regression approach that predicts the Warfarin dose taking as input the gene damaging (GD) scores and the clinical covariates collected by the data providers. In particular the input variables consist of the GD scores associated to 12 genes potentially involved in the disease (Coagulation factors 1 to 10, Vitamin K-dependent proteins C and S) and 6 clinical covariates data (gender, age, height, weight, aspirin and amiodarone).
The GD score is calculated summing the probabilistic output of the *PhD-SNP^g* algorithm (Capriotti and Fariselli 2017, PMID: 28482034) for each variant falling in the selected gene of an individual. *PhD-SNP^g* is a machine learning method that predicts deleterious variants using conservation scores made available by the UCSC genome browser. For the input the gender is represented by a binary variable.
The output of the linear regression method is rescaled between 0 and 1.
After the comparison with the clinical data and using a threshold of 0.5, our approach resulted in an overall accuracy of 0.70 a Matthews correlation coefficient of 0.40 and an AUC of 0.76. In terms of AUC, this result is comparable with that achieved by a naïve method based on the prescribed Warfarin dose rescaled by a factor of 0.01. The naïve approach reaches overall accuracy 0.73, Matthews correlation coefficient 0.44 and AUC 0.74. The score of the performance of both methods are reported in the following table.

Method	Accuracy	TNR	NPV	TPR	PPV	MCC	F1	AUC
Naïve	0.73	0.73	0.60	0.73	0.83	0.44	0.77	0.74
Linear	0.70	0.73	0.56	0.68	0.82	0.40	0.74	0.76

References

Capriotti E, Fariselli P. (2017). PhD-SNP^g: A webserver and lightweight tool for scoring single nucleotide variants. Nucleic Acids Research. DOI: 10.1093/nar/gkx369.



Supplementary Figure S6: Evaluation plots for the submission from group 4. Refer to supplementary figure 2 for a description of the plots.

Group 5

Yanran Wang[1], Yana Bromberg[1]

[1]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, New Jersey

The Bromberglab submitted four predictions CAGI 2018 Clotting disease African American (AA) exomes challenge. These are summarized below and detailed in our special issue manuscript[1].

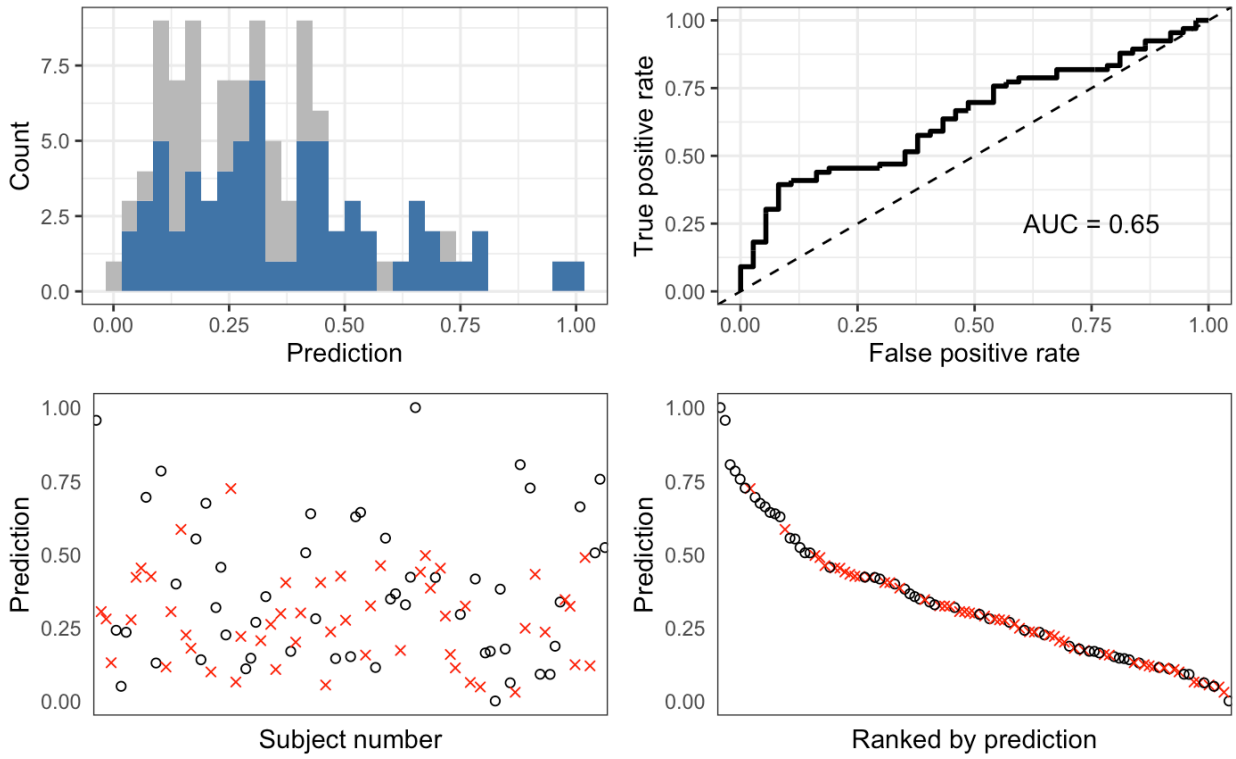
1. Our best result (ranked first of all submissions that used solely the genomic info) used the SNAP[2] predicted functional impact of variants within the known VTE genes extracted from DisGeNET database[3]. Each person/exome was represented as a vector of SNAP scores of the variants that were predicted to be non-neutral. We then clustered these vectors using K-means[4] into two groups.
2. We further aggregated the variant-level function changes to represent gene-level functionality. The resulting gene function vectors of each person/exome were K-means clustered into two groups for our second-best result (ranked third of all the methods that used solely the genomic info).

3. Unlike the first two methods that used function change as a feature, our other method which used the genotypes of variants within VTE genes directly. This approach had a nearly random performance. Note that K-modes[5] clustering was used here as genotypes are categorical (not continuous) features.
4. We also applied genetic risk scoring (GRS) using variants identified by the VTE genome-wide association study (GWAS) in AA population. All risk variants were extracted from Heit *et al.*[6] and VTE risk was calculated as the sum of all risk allele counts with log odds ratio weights. The GRS method (ranked fourth overall) performed worse than the two clustering methods: it had high precision (90%) at the cost of a low recall (26%). This result suggests that the GWAS VTE signal is not sufficient to explain a large portion of the disease pathogenesis.

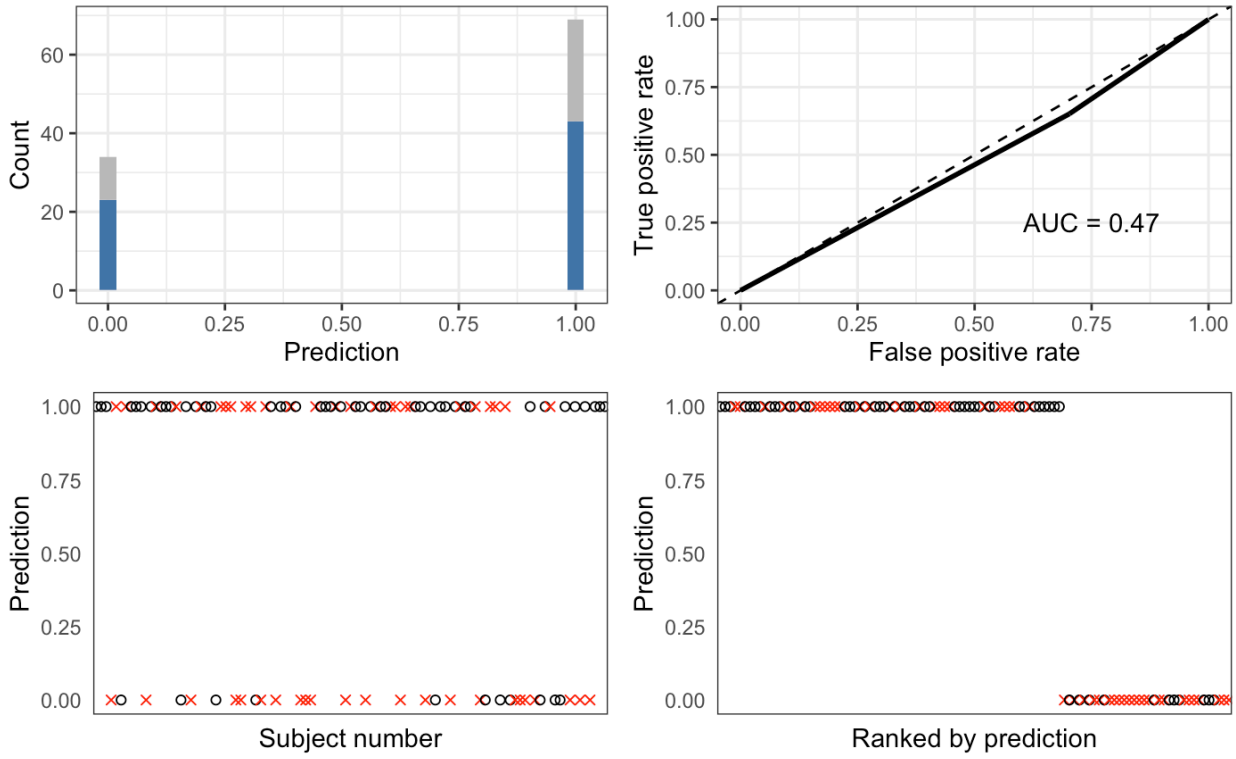
Our results, thus, suggest that variant functional impacts are critical to understanding VTE pathogenicity and for improving predictor performance.

References

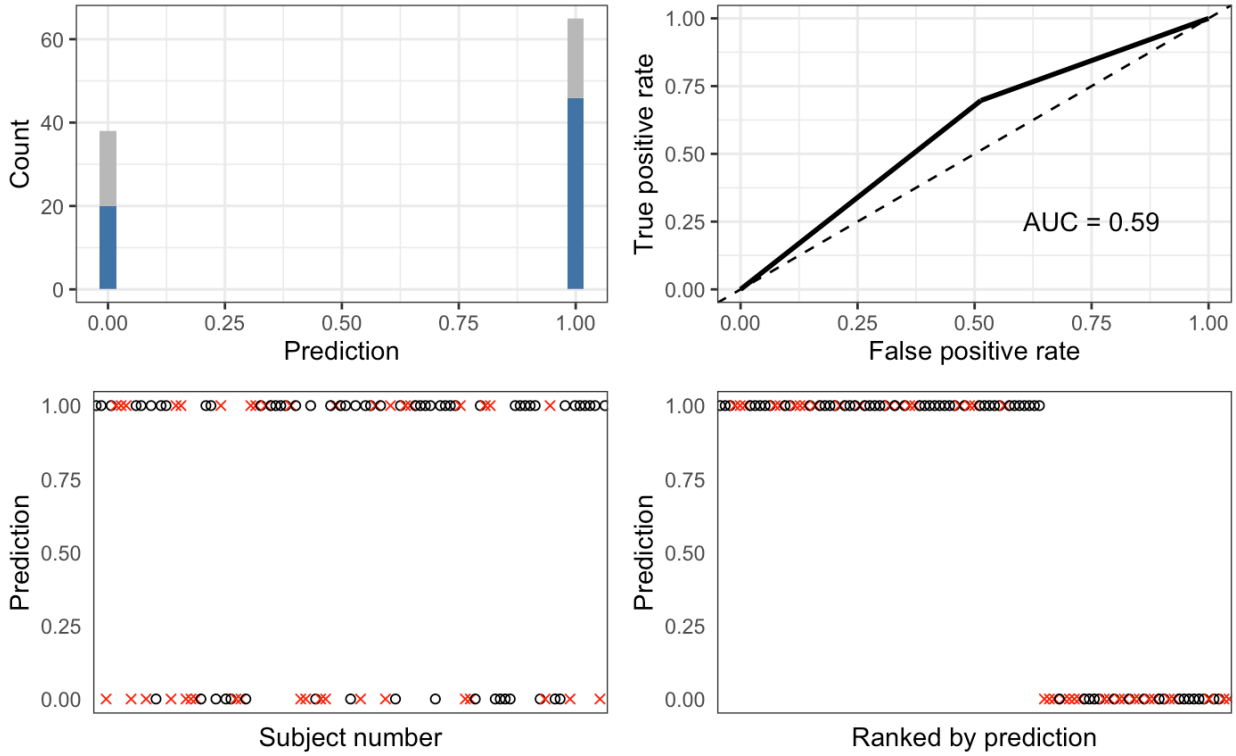
1. Wang, Yanran, and Yana Bromberg. "Identifying mutation-driven changes in gene functionality that lead to venous thromboembolism." Human Mutation (In Press)
2. Bromberg, Yana, and Burkhard Rost. "SNAP: predict effect of non-synonymous polymorphisms on function." Nucleic acids research 35.11 (2007): 3823-3835.
3. Piñero, Janet, et al. "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes." Database 2015 (2015).
4. MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. No. 14. 1967.
5. Huang, Zhexue. "A fast clustering algorithm to cluster very large categorical data sets in data mining." DMKD 3.8 (1997): 34-39.
6. Heit, John A., et al. "Identification of unique venous thromboembolism-susceptibility variants in African-Americans." Thrombosis and haemostasis 117.4 (2017): 758.



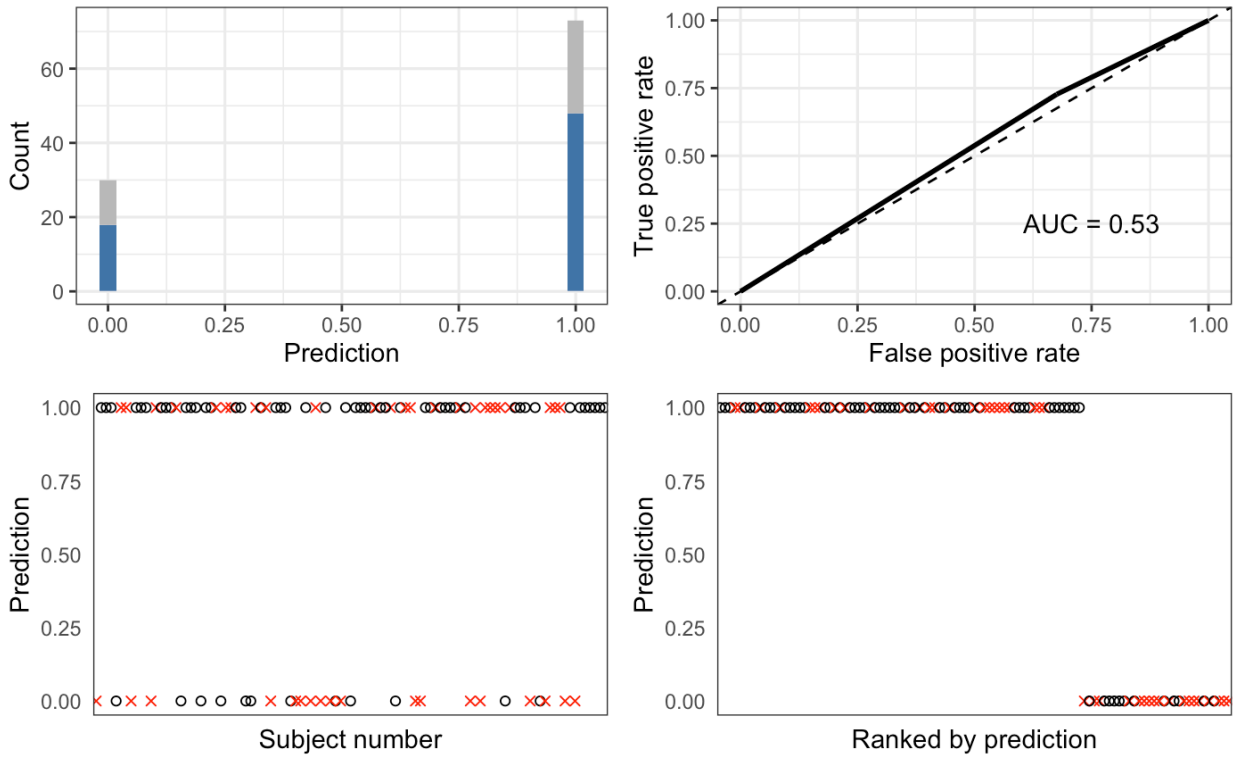
Supplementary Figure S7: Evaluation plots for the first submission from group 5 (Group 5a in Table 1, method 4 above). Refer to supplementary figure 2 for a description of the plots.



Supplementary Figure S8: Evaluation plots for the second submission from group 5 (Group 5b in Table 1, method 3 above). Refer to supplementary figure 2 for a description of the plots.



Supplementary Figure S9: Evaluation plots for the third submission from group 5 (Group 5c in Table 1, method 1 above). Refer to supplementary figure 2 for a description of the plots.



Supplementary Figure S10: Evaluation plots for the fourth submission from group 5 (Group 5d in Table 1, method 2 above). Refer to supplementary figure 2 for a description of the plots.

Group 6
Samuele Bovo[1], Castrense Savojardo[1], Pier Luigi Martelli[1], Rita Casadio[1,2]

[1] Bologna Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Italy

[2] CNR, Institute of Biomembrane and Bioenergetics (IBIOM), Via Giovanni Amendola 165/A - 70126 Bari Italy

Methods

The Bologna Biocomputing group (group 6) provided 4 different submissions. The first one (submission 6a) only analyses the provided metadata as grouped following aspirin and non-aspirin prescription. The other three submissions (6b, 6c, 6d) consider exome data and analyze the individual variations in a panel of 49 candidate genes derived from GWAS experiments and described in literature as being associated with venous thromboembolism (VTE), with differentiation of arterial and venous endothelia and with warfarin pharmacogenomics.

For calibrating our predictions, we took into consideration the statistics on the dataset reported in Table 1 of the paper describing the sampled population (Daneshjou et al., 2014). In particular, it reports that, out of 103 patients, 58 and 45 assume high and low doses of warfarin, respectively. Moreover, the VTE patients account for the 82.75% of patients assuming high dose of warfarin (48 individuals) and the 40.91% of patients assuming a low dose of warfarin (18

individuals). We then expect that the dataset includes 66 VTE affected patients. The remaining 37 patients are atrial fibrillation (AF) affected, as reported (Daneshjou et al., 2014).

Submission 6a

The provided metadata indicate that 35 individuals assume aspirin and 68 do not. These numbers are close to the expected numbers of AF (37) and VTE (66) patients. Moreover, warfarin-aspirin combination is often prescribed to AF patients (Turan et al., 2016). Considering this knowledge, we classified all the patients assuming aspirin to be AF affected.

Submissions 6b, 6c, 6d

The discrimination between VTE and AF cases exploits the analysis of variations in genes involved in blood coagulation or related to the arterial and venous endothelia.

To collect a set of possibly relevant genes, we retrieved: i) 15 genes involved in VTE from the GWAS Catalog (https://www.ebi.ac.uk/gwas/); ii) 11 genes influencing the anticoagulant warfarin activity from Daneshjou et al. (2014); iii) 23 genes reviewed in Dela Paz and D'Amore (2009), associated with the differentiation of venous and arterial endothelia and possibly related to VTE and AF, respectively.

The list of 49 candidate genes is available in the Appendix.

We started from the provided VCF files, filtering out 430,766 variations with a minor allele frequency (MAF) < 0.05 or with missing genotypes in at least one individual. Restricting the analysis to the 49 candidate genes, we found a total of 485 SNPs present in at least one of the 103 exomes, in either homo- or heterozygous form.

We then built a binary matrix reporting the presence of each SNP in each exome (485 SNPs x 103 exomes). We adopted Principal Component Analysis (PCA) to represent data in a low-dimensional space and to highlight possible splits in agreement with the expected number of VTE and AF cases (66 and 37, respectively). We considered different combinations of principal components (PC). In the PC1- PC3 plane, the 103 exomes well separate into two groups of 63 and 40 individuals (submission 6b). When considering the PC2 -PC3 plane we obtained groups of 59 and 44 exomes (submission 6c), while in the PC4-PC5 plane clusters 64 and 39 individuals (submission 6d). To better highlight the groups we performed, for each PC combination, a clustering analysis with the K-means algorithm (with K=2).

Appendix

Genes involved in VTE – ABO, COX7A2L, KCNG3, EPHA3, F11, F2, F5, F8, FGG, KNG1, PROCR, SLC44A2, TMEM170B, ADTRP, TSPAN15

Genes associated with warfarin dosage – ZFHX3, CYP2C9, VKORC1, FPGS, CD177, ZNF229, LBR, ALKBH5, LOC441601, HIVEP3, SEMA3G

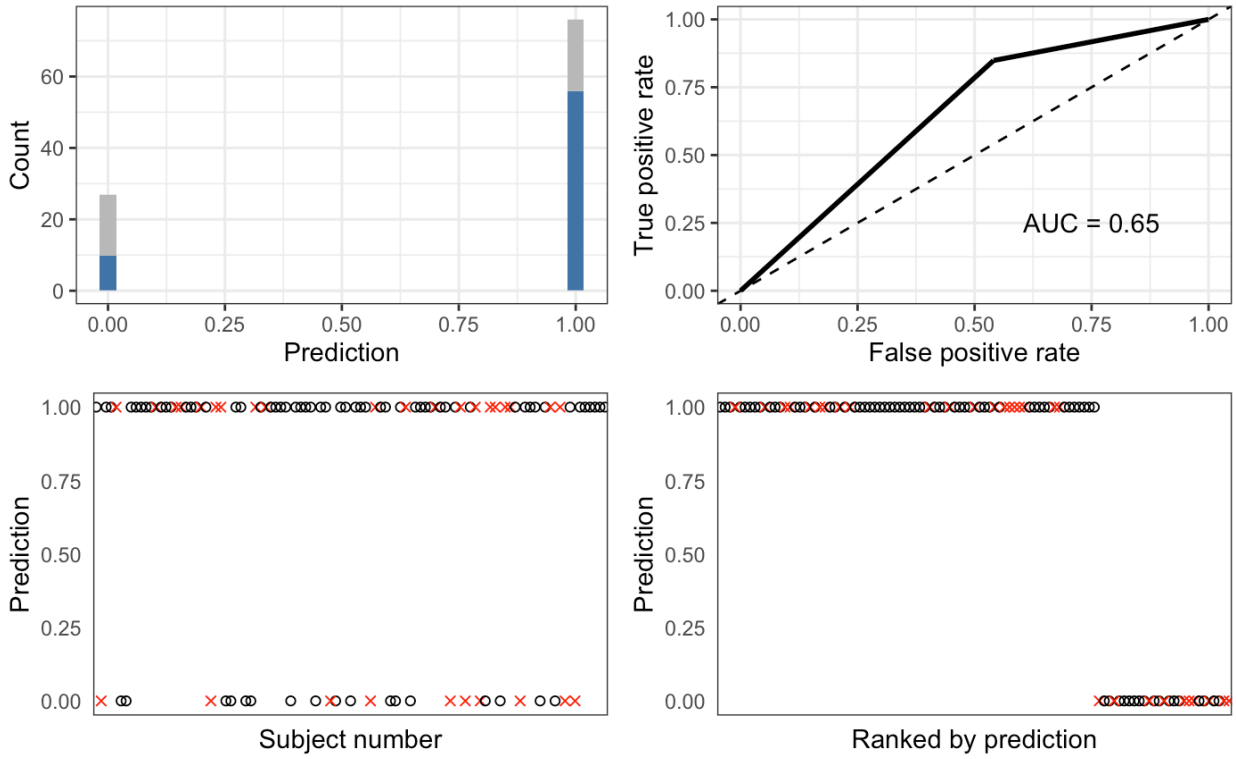
Genes involved in the development of arterial and venous endothelia – EFNB2, NRP1, GJA5, BMX, NOTCH1, NOTCH4, DLL4, JAG1, JAG2, HEY2, KDR, TBX20, ACVRL1, EPAS1, DEPP1, VEGFA, EPHB4, LEFTY1, LEFTY2, FLT4, NRP2, TEK, NR2F2, EMC

References

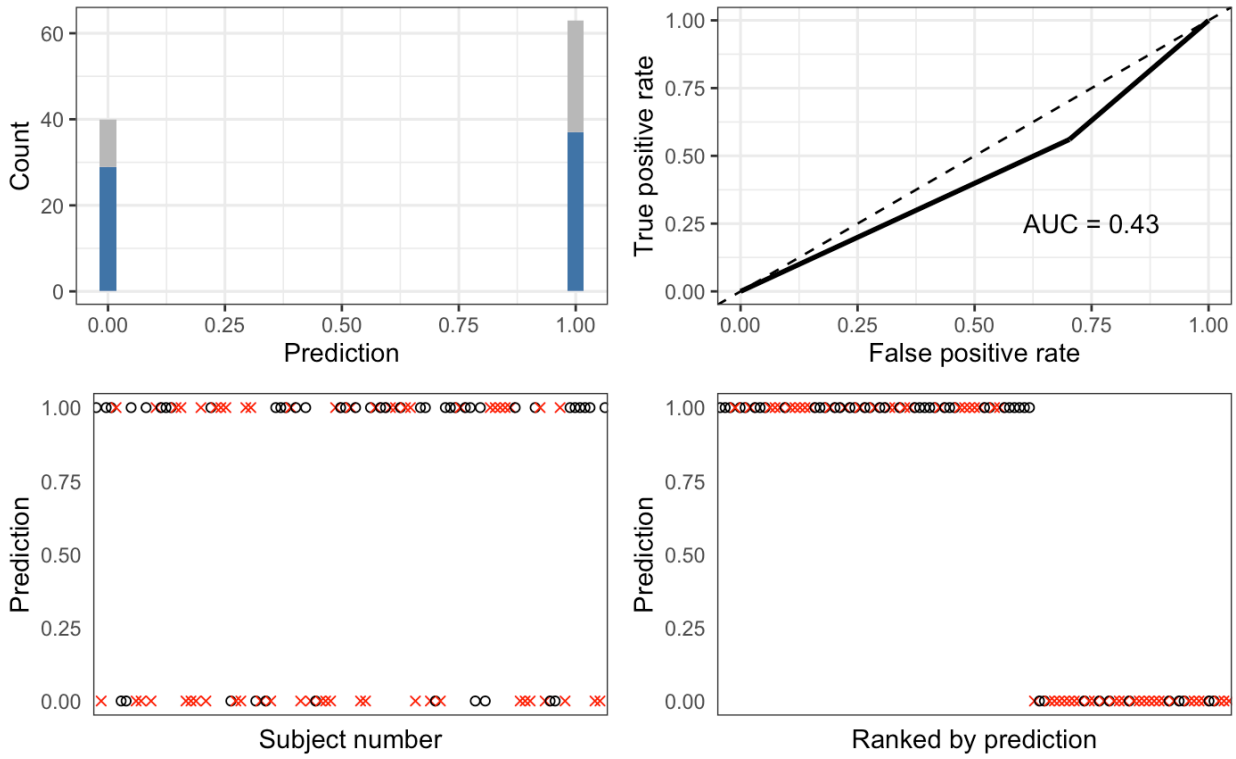
Daneshjou R, Gamazon ER, Burkley B, Cavallari LH, Johnson JA, Klein TE, Limdi N, Hillenmeyer S, Percha B, Karczewski KJ, Langaee T, Patel SR, Bustamante CD, Altman RB, Perera MA. Genetic variant in folate homeostasis is associated with lower warfarin dose in African Americans. Blood. 2014;124:2298-2305. doi: 10.1182/blood-2014-04-568436

Dela Paz NG, D'Amore PA. Arterial versus venous endothelial cells. Cell Tissue Res. 2009; 335:5-16. doi: 10.1007/s00441-008-0706-5.

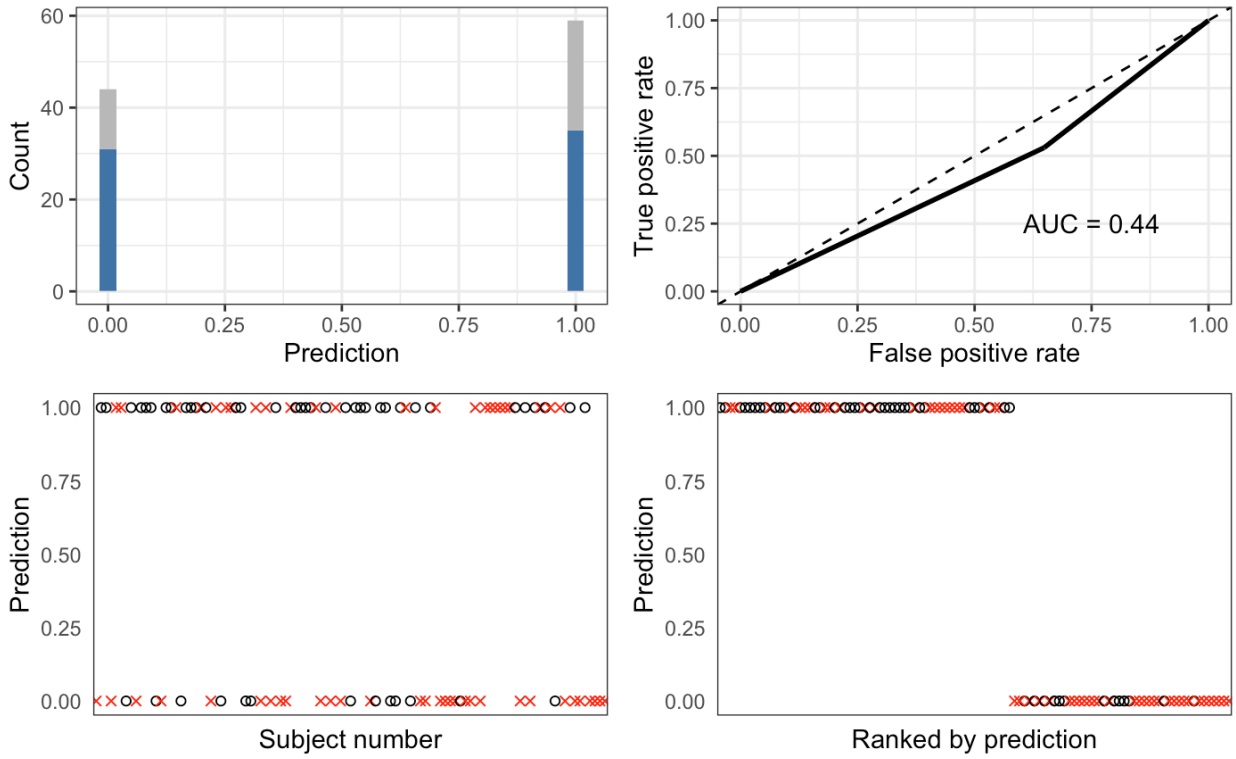
Turan B, Demir H, Mutlu A, Daşlı T, Erkol A, Erden İ. Inappropriate combination of warfarin and aspirin. Anatol J Cardiol. 2016;16:189-96. doi: 10.5152/akd.2015.6050



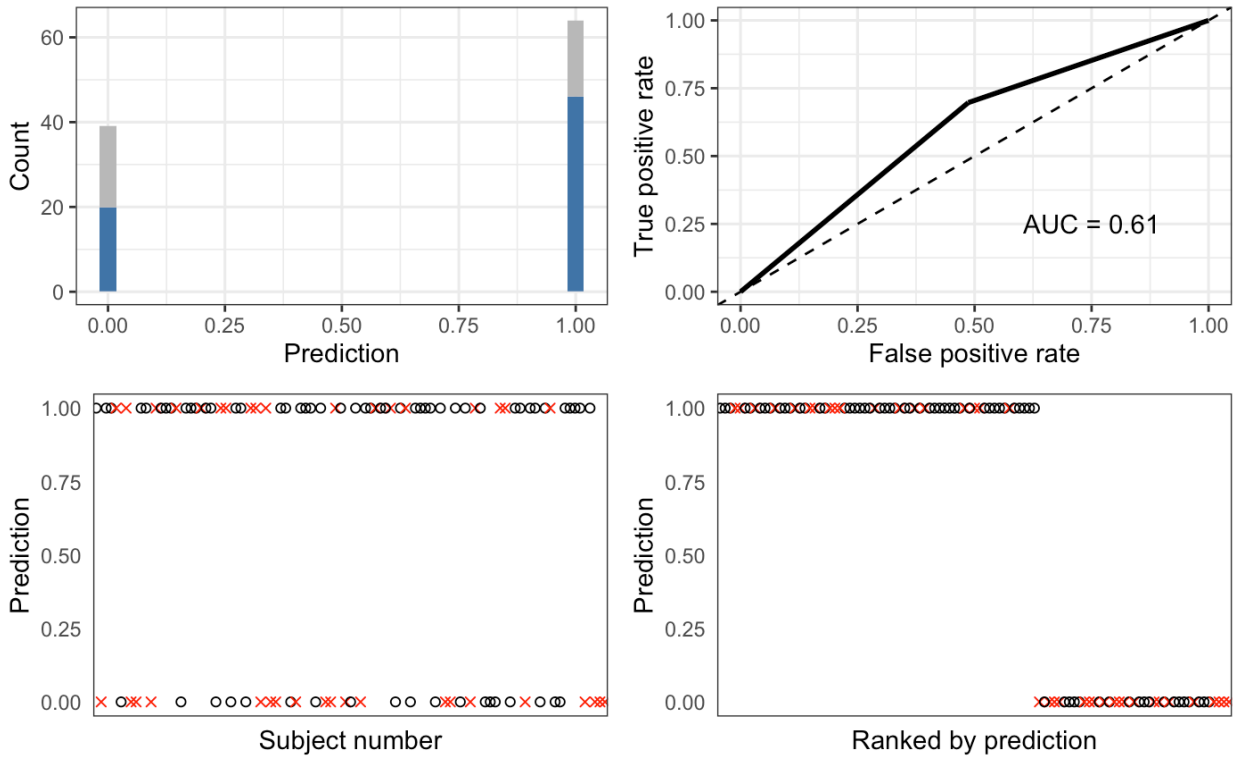
Supplementary Figure S11: Evaluation plots for the first submission from group 6. Refer to supplementary figure 2 for a description of the plots.



Supplementary Figure S12: Evaluation plots for the second submission from group 6. Refer to supplementary figure 2 for a description of the plots.



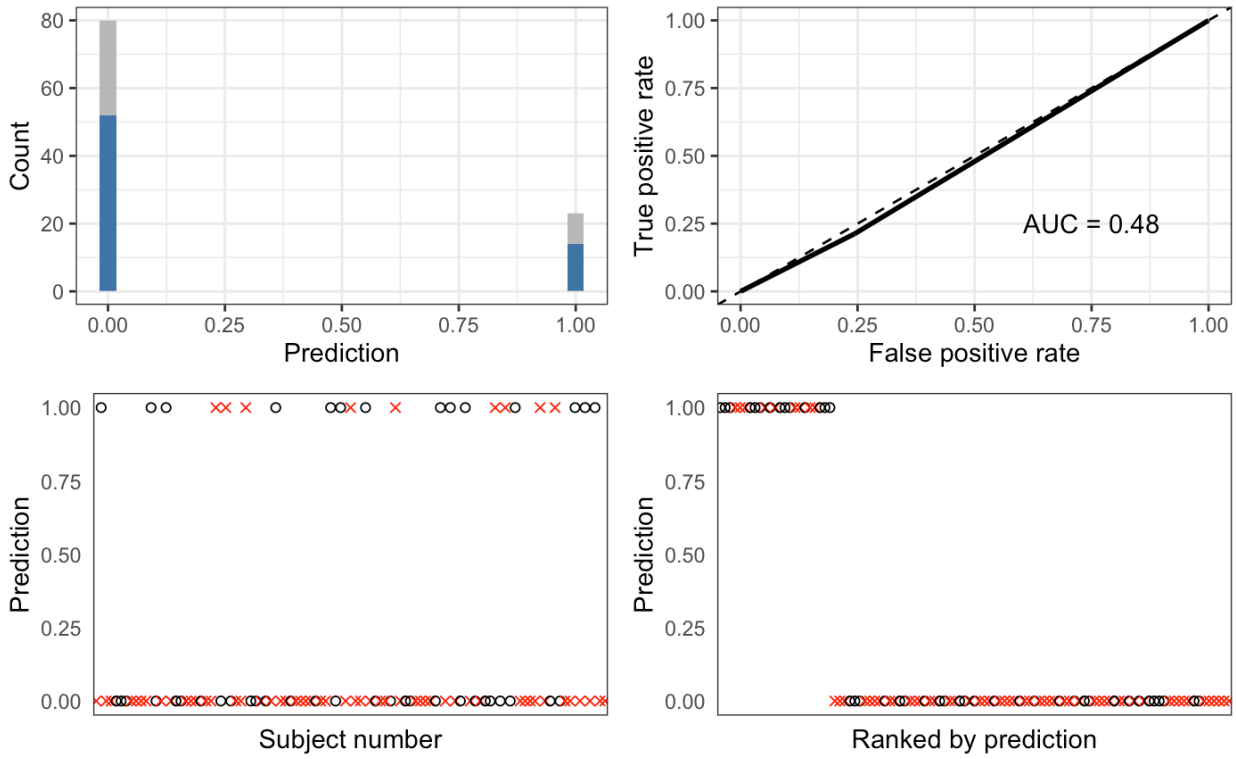
Supplementary Figure S13: Evaluation plots for the third submission from group 6. Refer to supplementary figure 2 for a description of the plots.



Supplementary Figure S14: Evaluation plots for the fourth submission from group 6. Refer to supplementary figure 2 for a description of the plots.

Group 7

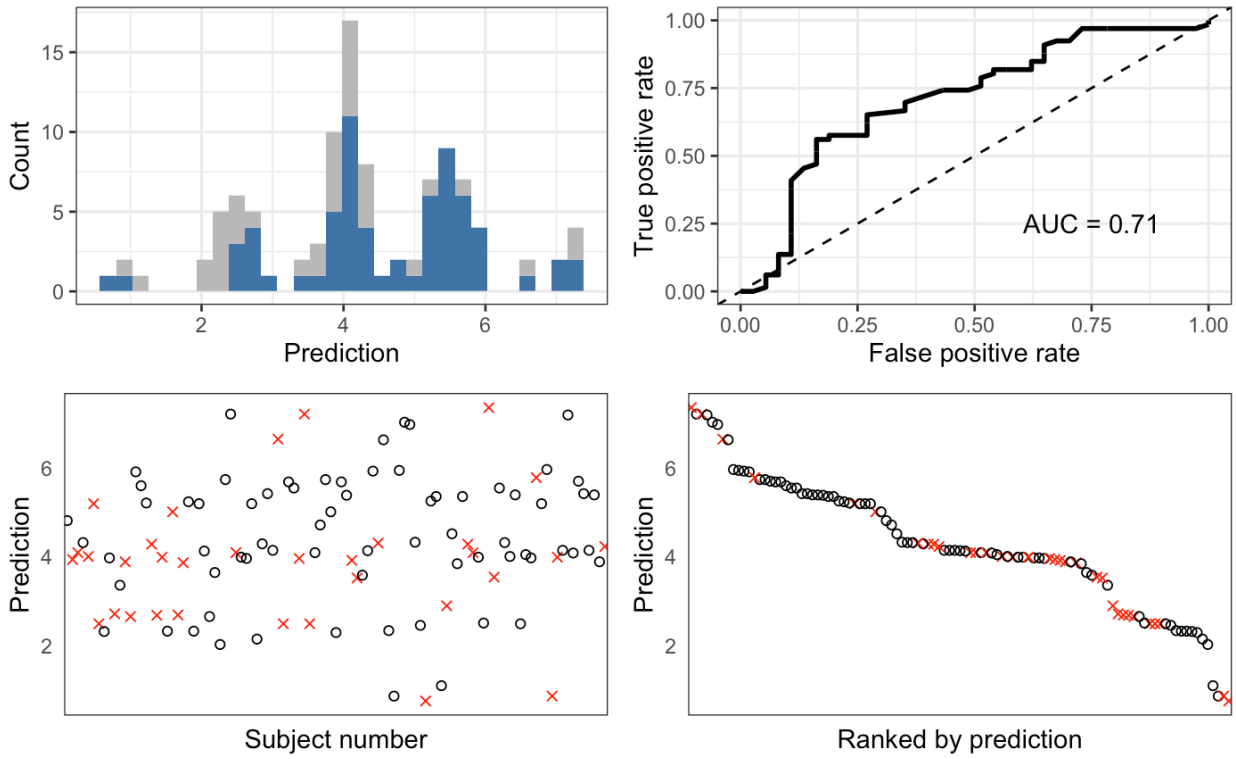
Group 7 did not submit a summary for the manuscript. In their submission they reported that they used a convolutional autoencoder to generate a latent representation of each sample then cluster the samples using agglomerative clustering.



Supplementary Figure S15: Evaluation plots for the submission from group 7. Refer to supplementary figure 2 for a description of the plots.

Baseline

Additional figures are included here for the baseline method described in the main text developed by Soria et al.


Count
15
10
5
0
2
4
6
Prediction
True positive rate
1.00
0.75
0.50
0.25
0.00
AUC = 0.71
0.00
0.25
0.50
0.75
1.00
False positive rate
Prediction
6
4
2
Subject number
Prediction
6
4
2
Ranked by prediction