Published in final edited form as:

Hum Mutat. 2019 September ; 40(9): 1314–1320. doi:10.1002/humu.23825.

Predicting venous thromboembolism risk from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges

Gregory McInnes[1], **Roxana Daneshjou**[2], **Panagiostis Katsonis**[3], **Olivier Lichtarge**[3,4,5,6], **Raj Gopal Srinivasan**[7], **Sadhna Rana**[7], **Predrag Radivojac**[8], **Sean D Mooney**[9], **Kymberleigh A Pagel**[10], **Moses Stamboulian**[10], **Yuxiang Jiang**[10], **Emidio Capriotti**[11], **Yanran Wang**[12], **Yana Bromberg**[12], **Samuele Bovo**[13], **Castrense Savojardo**[13], **Pier Luigi Martelli**[13], **Rita Casadio**[13,14], **Lipika R. Pal**[15], **John Moult**[15,16], **Steven Brenner**[17], **Russ Altman**[18]

[1]Biomedical Informatics Training Program, Stanford University, Stanford, CA

[2]Department of Dermatology, Stanford School of Medicine, Stanford, CA

[3]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX

[4]Department of Biochemistry & Molecular Biology, Baylor College of Medicine, Houston, TX

[5]Department of Pharmacology, Baylor College of Medicine, Houston, TX

[6]Computational and Integrative Biomedical Research Center, Baylor College of Medicine, Houston, TX

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

[8]Northeastern University, Boston, MA

[9]University of Washington, Seattle, WA

[10]Indiana University, Bloomington, IN

[11]BioFolD Unit, Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, Bologna, Italy

[12]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ

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

[14]CNR, Institute of Biomembrane and Bioenergetics (IBIOM), Bari, Italy

[15]Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, Maryland, Rockville, MD

[16]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD

[17]Department of Plant and Microbial biology, University of California Berkeley, Berkeley, CA

[18]Departments of Bioengineering, Biomedical Data Science, Genetics, and Medicine, Stanford University, Stanford, CA

Disclosure Statement

All authors have no conflict of interest to declare.

Abstract

Genetics play a key role in venous thromboembolism (VTE) risk, however established risk factors in European populations do not translate to individuals of African descent due to differences in allele frequencies between populations. As part of the fifth iteration of the Critical Assessment of Genome Interpretation, participants were asked to predict VTE status in exome data from African American subjects. Participants were provided with 103 unlabeled exomes from patients treated with warfarin for non-VTE causes or VTE and asked to predict which disease each subject had been treated for. Given the lack of training data, many participants opted to use unsupervised machine learning methods, clustering the exomes by variation in genes known to be associated with VTE. The best performing method using only VTE related genes achieved an AUC of 0.65. Here we discuss the range of methods used in the prediction of VTE from sequence data and explore some of the difficulties of conducting a challenge with known confounders. Additionally, we show that an existing genetic risk score for VTE that was developed in European subjects works well in African Americans.

Keywords

Venous thromboembolism; machine learning; phenotype prediction; exomes; prediction challenge

1 Introduction

There are 300,000 to 900,000 cases of venous thromboembolism (VTE) a year in the United States alone (Beckman, et al, 2010). VTE captures both deep vein thrombosis (DVT) and pulmonary embolism (PE). There are differences in the incidence of VTE based on ancestry; individuals of African ancestry have a 30–60% higher incidence of VTEs than people of European ancestry (Roberts et al., 2009; Zakai & McClure, 2011). VTE risk is multifactorial, both environmental and genetic factors are involved (Feero, 2004). For individuals of European descent, the commonly seen VTE risk factors are *F5* R506Q (rs6025:G>A; three to fivefold increased risk of VTE in carriers) and *F2* G20210A (rs1799963:G>A; two to threefold increased risk of VTE in carriers) (Middeldorp & van Hylckama Vlieg, 2008; Rosendaal & Reitsma, 2009).

However, the genetic variants that confer risk in populations of European descent are nearly absent in African Americans, and population-specific genetic factors influencing the higher VTE rate are not well characterized (Dowling, et al., 2003). A recent study identified a population-specific genetic risk factor in African Americans, but much of the genetic risk is still undiscovered (Daneshjou et al., 2016). Previous work has been done to develop genetic risk models for VTE in European populations, but no such risk model exists for individuals of African descent and the existing models have not been tested in African populations (Soria et al. 2014).

The Critical Assessment of Genomic Interpretation (CAGI) aims to objectively assess the prediction of phenotypic impacts of genetic variation. In the fifth iteration of CAGI, participants were challenged to predict the VTE status of 103 African American individuals from exome data. This dataset was used as part of a warfarin dosage prediction challenge in

CAGI 4, where participants were asked to predict the precise warfarin dosage of each individual (Daneshjou et al., 2017). VTE often requires long term use of anticoagulants. The dataset comprised 66 individuals with a VTE diagnosis and 37 individuals on warfarin for non-VTE causes (such as atrial fibrillation prophylaxis, AF). Thus, we were able to repurpose this data CAGI 5 and participants were asked to distinguish between individuals that were prescribed warfarin for a clotting disorder versus those that were prescribed warfarin for non-VTE purposes.

2 Methods

2.1 Data distribution

Participants were provided exome data for all 103 subjects in the VCF file format as well as corresponding covariate data. The covariate data included was subject age, height, weight, sex, and drug regimen (aspirin, amiodarone, and warfarin dose). Amiodarone is an antiarrhythmic drug used to treat atrial fibrillation, which could be a clear sign that the subject belonged in the AF group. However, only one subject was on amiodarone, so this conferred no predictive advantage to the participants. Participants consented to the CAGI data use agreement.

2.2 Predicting phenotypes

Participants were asked to make VTE status predictions for all 103 subjects in the provided data. No labelled training data was provided. Participants were required to return a text file with predicted disease status and confidence in the prediction for each subject. They were also provided with a validation script to check their output formatting. Participants were asked to provide a brief description of their prediction methods for each submission. The prediction results were presented at the CAGI 5 meeting.

2.3 Data quality

The data had previously undergone rigorous QC using ancestry informative markers to confirm self-reported ancestry and identity by state (IBS) analysis in order to ensure that samples were not related, as previously described (Daneshjou et al., 2014).

2.4 Assessing predicted phenotypes

In order to assess the submissions of each group, several evaluation metrics were used. Predictions were evaluated using area under the ROC curve (AUC), accuracy, sensitivity, specificity, and F1 scores. Some participants submitted binary class predictions rather than probabilities. In order to fairly evaluate predictions across all groups the predictions that were submitted as probabilities were binarized using a cutoff of 0.5, where a score greater than 0.5 indicates a VTE prediction. Accuracy, sensitivity, specificity, recall, and F1 scores were then computed with the binarized data, whereas AUC was calculated on the submitted scores.

2.5 Establishing a baseline

A baseline prediction score was calculated using the multilocus genetic risk score proposed by Soria et al. The proposed method uses a linear model of 17 loci across nine VTE related genes. To compute the scores the number of the alternate alleles at each site was multiplied by the corresponding coefficients proposed by Soria et al. As with the participant submitted scores, the genetic risk scores were binarized using a threshold of 0.5 prior to calculating accuracy, sensitivity, specificity, recall, and F1 scores, and the raw scores were used to compute AUC.

3 Results

We assessed 14 submissions of phenotype predictions from seven groups. As no training dataset was provided, most participants chose to use unsupervised models trained on variants from genes previously reported to be associated with the phenotypes. Some groups used burden based scoring methods, scoring samples by the frequency of damaging variants in selected genes.

Each of the participants formulated their own strategy for predicting phenotype from the exome data. Although each was unique, there were many similarities between the methods employed (Figure 1). All submissions but one primarily used the genetic data, each group first selected genes related to the phenotypes of interest from a disease-gene database, then used the variants in those genes for downstream analysis. Half of the fourteen submissions employed an unsupervised approach, clustering the variants from the selected genes using a variety of approaches. Clustering methods included principle component analysis, k-means clustering, and a single submission using a deep learning-based approach with autoencoders. Six of the groups employed scoring-based methods to the variants within the selected genes to calculate an overall burden score for each subject. A single submission did not use the genotype data at all and trained a logistic regression classifier to predict VTE status based clinical covariates.

The dataset was originally collected to study the genetics of warfarin dosage and had been previously published on and the original publication reports that VTE status is significantly associated with warfarin dosage (Daneshjou et al., 2016, Supplementary Figure S1). Warfarin dosage was provided to the participants as a covariate for each subject. The known relationship between VTE status and warfarin dose in the dataset was exploited by several groups in their predictions. The most extreme case classified individuals as VTE patients if they were on a high warfarin dose, and classified individuals on a low warfarin dose as AF. This was the best performing method overall achieving 72% accuracy. Since warfarin dosage is largely influenced by genetics, several groups included genes involved in warfarin pharmacokinetics and pharmacodynamics in their models. Overall, five of the 14 submissions used knowledge that warfarin dosage is associated with VTE status in this dataset in some form.

Of the nine submissions that did not use warfarin dosage to inform their predictions, all utilized either an unsupervised, clustering-based, approach to distinguish the two classes, or used various methods to score variants based on predicted deleteriousness. The top

performing group that did not inform their predictions with warfarin dosage information achieved an AUC of 0.65. This method selected genes associated with VTE, pulmonary embolism, and deep vein thrombosis (25 genes total) from DisGeNET and performed k-means clustering on variants determined to be non-neutral by SNAP (Pinero et al, 2017; Bromberg et al, 2008). The distribution of AUC scores for all predictions can be seen in Figure 2 and a complete list of the scores for each submission is presented in Table 1. Details of all prediction methods can be found in the supplementary material.

A baseline prediction accuracy was generated using a linear model proposed by Soria et al. The baseline model outperformed all submissions that did not use warfarin, achieving a prediction accuracy of 67% and an AUC of 0.71.

4 Discussion

This CAGI exome prediction challenge has yielded several insights into the genetics of VTE in African Americans as well as insights into the challenges conducting prediction challenges.

Predicting VTE risk from genetic sequence is a difficult task and the challenge offered in CAGI 5 was no exception. Participants were asked to differentiate exomes of individuals suffering from VTE and those who may be treated with warfarin for a different indication. This task was further complicated by the lack of training data to for participants to validate their proposed methods. This led most participants to develop methods using existing biological knowledge to perform feature selection.

Most participants opted to use clustering-based approaches to predict VTE status. This was a prudent decision given the lack of training data and the stated goal of distinguishing two traits. The other common approach was to score variants within genes based on their predicted deleteriousness, then to create a final score for each individual based on the number of deleterious variants. Although the best performing method used a clustering approach (k-means), there was no clear advantage to using clustering methods over scoring methods.

All groups subset the exome to genes known to be associated with VTE or AF to use for downstream predictions. The groups with the top two highest scoring submissions both used DisGeNET to select phenotype associated genes. There is clear value in limiting the search space of the genome and DisGeNET seems to be a useful asset for selecting phenotype associated genes. Groups 1 and 5 (which accounted for 5 of the top 6 submissions that did not use warfarin), both used DisGeNET to select genes for their predictions but then applied different methods to predicting subject status from those genes.

One major point of contention in conducting this challenge was the confounding effect of warfarin on the prediction task. Participants were asked to make predictions about VTE status with the given dataset but were not advised to avoid using warfarin dosage which may be a confounding variable unique to this dataset and not broadly correlated with VTE risk. It is therefore reasonable that the participants sought the best performance possible and used the previously published correlation between VTE and warfarin dose in this dataset. This

does, however, go against the spirit of the CAGI experiment, which is meant to better understand the impact of genetics on phenotype and this challenge in particular aimed to improve our collective ability to predict VTE from genetic data. For this reason, we have divided the prediction results between those who used warfarin dosage information, either directly (using the subjects warfarin dose) or indirectly (through the inclusion of warfarin dose related genes), in their models.

The point was raised at the CAGI conference that if this confounding factor was known, why give the warfarin dosage to participants at all? This was due to a miscommunication between the data providers and the conference organizers. However, it may have made little difference as the participants were provided with the entire exome and there is a strong genetic relationship between warfarin dosage and genetics. In hindsight, it may have been better for the challenge to not provide warfarin dosage to the participants and to remove genes related to warfarin pharmacokinetics and pharmacodynamics from the exomes. Alternatively, it may have been better to explicitly inform the participants to avoid using any knowledge about warfarin in their predictions.

In order to assess the submitted predictions against an existing gold standard we calculated genetic risk scores for each exome using the method proposed by Soria et al. The genetic risk scores calculated using this method achieved an AUC of 0.71, greater than that of any submitted method that did not use warfarin dose in their predictions. This method was developed using data from individuals of European descent and had not been previously validated in individuals of African descent. The AUC achieved by this method in African Americans exceeds the reported AUC in the original study population (0.677). A different, five locus, genetic risk score had previously been tested in African populations and found to not perform as well as it did in whites (Folsom et al, 2015). This suggests that the method proposed by Soria et al may be clinically useful in predicting VTE in African Americans and could warrant further clinical validation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank and acknowledge the CAGI planning committee and participants. The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650. G.M. is supported by BD2K grant number T32 LM 012409. Y.B. and Y.W. were supported by the NIH U01 GM115486 grant.

Funding

The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650. G.M. is supported by BD2K grant number T32 LM 012409. Y.B. and Y.W. were supported by the NIH U01 GM115486 grant.

References

Bromberg Y, Yachdav G, & Rost B (2008). SNAP predicts effect of mutations on protein function. Bioinformatics, 24(20), 2397–2398. 10.1093/bioinformatics/btn435 [PubMed: 18757876]

Daneshjou R, Gamazon ER, Burkley B, Cavallari LH, Johnson JA, Klein TE, … Perera MA (2014). Genetic variant in folate homeostasis is associated with lower warfarin dose in African Americans. Blood, 124(14), 2298–2305. 10.1182/blood-2014-04-568436 [PubMed: 25079360]

Daneshjou R, Cavallari LH, Weeke PE, Karczewski KJ, Drozda K, Perera MA, … Altman RB (2016). Population-specific single-nucleotide polymorphism confers increased risk of venous thromboembolism in African Americans. Molecular Genetics & Genomic Medicine, 4(5), 513–520. 10.1002/mgg3.226 [PubMed: 27652279]

Daneshjou R, Wang Y, Bromberg Y, Bovo S, Martelli PL, Babbi G, … Morgan AA (2017). Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Human Mutation, 38(9), 1182–1192. 10.1002/humu.23280 [PubMed: 28634997]

Dowling NF, Austin H, Dilley A, Whitsett C, Evatt BL, & Hooper WC (2003). The epidemiology of venous thromboembolism in Caucasians and African-Americans: the GATE Study. Journal of Thrombosis and Haemostasis : JTH, 1(1), 80–87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12871543 [PubMed: 12871543]

Feero WG (2004). Genetic thrombophilia. Primary Care, 31(3), 685–709, xi. 10.1016/j.pop.2004.04.014 [PubMed: 15331254]

Folsom AR, Tang W, Weng L-C, Roetker NS, Cushman M, Basu S, & Pankow JS (2016). Replication of a genetic risk score for venous thromboembolism in whites but not in African Americans. Journal of Thrombosis and Haemostasis, 14(1), 83–88. 10.1111/jth.13193 [PubMed: 26565658]

Middeldorp S, & van Hylckama Vlieg A (2008). Does thrombophilia testing help in the clinical management of patients? British Journal of Haematology, 143(3), 321–335. 10.1111/j.1365-2141.2008.07339.x [PubMed: 18710381]

Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, … Furlong LI (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research, 45(D1), D833–D839. 10.1093/nar/gkw943 [PubMed: 27924018]

Roberts LN, Patel RK, & Arya R (2009). Venous thromboembolism and ethnicity. British Journal of Haematology, 146(4), 369–383. 10.1111/j.1365-2141.2009.07786.x [PubMed: 19552721]

Rosendaal FR, & Reitsma PH (2009). Genetics of venous thrombosis. Journal of Thrombosis and Haemostasis, 7, 301–304. 10.1111/j.1538-7836.2009.03394.x [PubMed: 19630821]

Soria JM, Morange P, Vila J, Souto JC, Moyano M, Trégouët D, … Elosua R (2014). Multilocus Genetic Risk Scores for Venous Thromboembolism Risk Assessment. Journal of the American Heart Association, 3(5), e001060 10.1161/JAHA.114.001060 [PubMed: 25341889]

Zakai NA, & McClure LA (2011). Racial differences in venous thromboembolism. Journal of Thrombosis and Haemostasis, 9(10), 1877–1882. 10.1111/j.1538-7836.2011.04443.x [PubMed: 21797965]

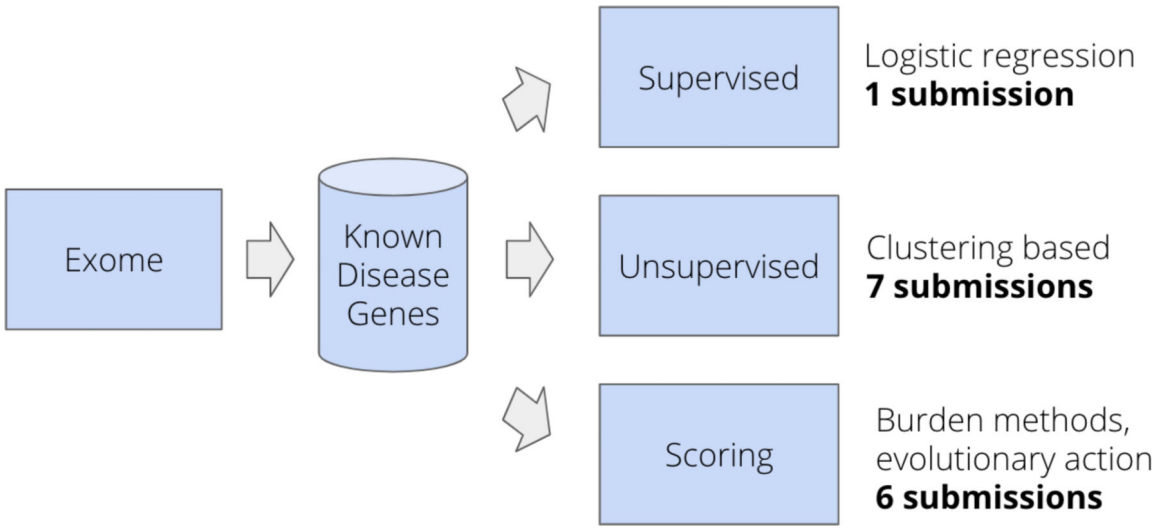

Figure 1.

General participant workflow. Each group formed their own approach to predicting phenotypes of the exomes, but there were some similarities across all submissions. All groups subset the exome into genes known to be involved in the phenotypes of interest, then made predictions based on the variants in those subset genes. Some groups generated scores for each individual based on burden of variants of a certain class. Others clustered the genotypes alone and segmented the clusters into predicted phenotypes.

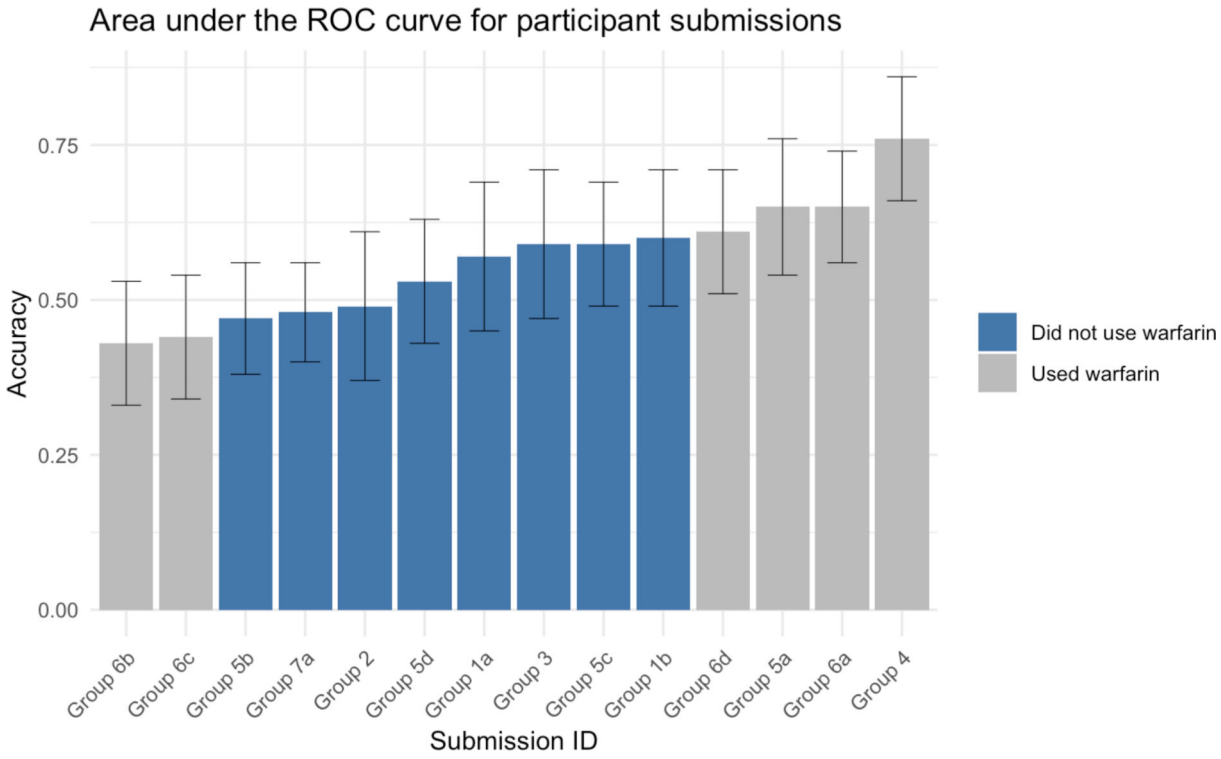

Figure 2.

Area under the ROC (receiver operating characteristic) curve for all submissions. Submissions that used knowledge of warfarin confounding in the dataset (either by including the warfarin dose or including genes involved in warfarin pharmacogenetics) are shown in red, submissions that did not use the warfarin confounding in any way are shown in blue. The error bars indicate the 95% confidence interval of the AUC.

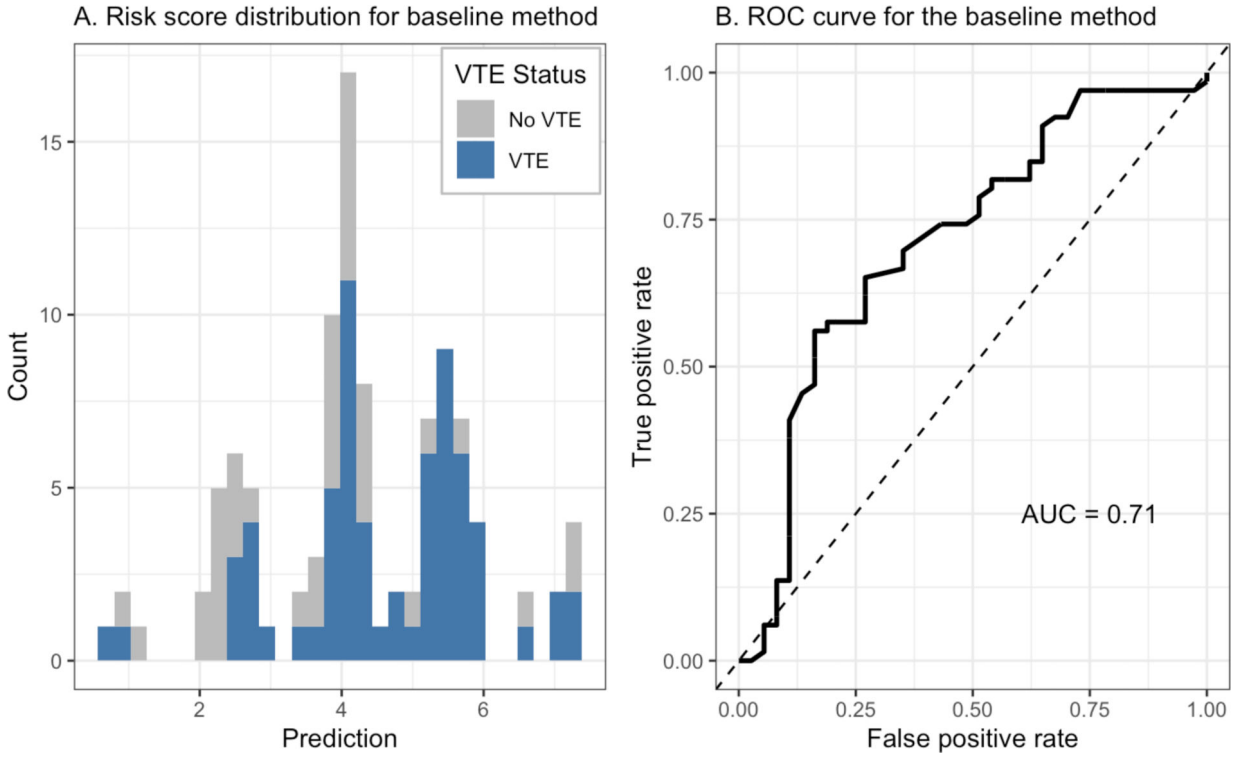

Figure 3.

Performance of baseline method on prediction of venous thromboembolism. Here we show the risk scores and predictive performance of the genetic risk score developed by Soria et al. We show the distribution of risk scores across all patients with subjects with VTE shown in blue and those without VTE shown in gray (left). We also show a ROC curve to illustrate the predictive performance of the baseline method (right).

Table 1.

Evaluation metrics for all submissions and for the baseline method. The table is broken up by submissions that used the known warfarin confounding, those that did not, and the baseline method. Within each group scores are sorted by AUC. Accuracy, sensitivity, specificity, and F1 are calculated using a cutoff of 0.5 for all predictions.

Description	Submission	Approach	AUC	Accuracy	Sensitivity	Specificity	F1
Did not use warfarin in prediction	Group 5a	Unsupervised	0.65	0.51	0.26	0.95	0.40
	Group 1b	Scoring	0.60	0.60	0.59	0.59	0.65
	Group 5c	Unsupervised	0.59	0.63	0.70	0.49	0.70
	Group 3	Scoring	0.59	0.34	0.23	0.54	0.31
	Group 1a	Scoring	0.57	0.47	0.30	0.76	0.42
	Group 5d	Unsupervised	0.53	0.59	0.73	0.32	0.69
	Group 2	Scoring	0.49	0.41	0.12	0.92	0.21
	Group 7a	Unsupervised	0.48	0.41	0.21	0.76	0.31
	Group 5b	Unsupervised	0.47	0.53	0.65	0.30	0.64
Used warfarin in prediction	Group 4	Supervised	0.76	0.70	0.71	0.65	0.75
	Group 6a	Scoring	0.65	0.72	0.85	0.46	0.79
	Group 6d	Unsupervised	0.61	0.64	0.70	0.51	0.71
	Group 6c	Unsupervised	0.44	0.47	0.53	0.35	0.56
	Group 6b	Unsupervised	0.43	0.47	0.56	0.30	0.57
Soria et al.	Baseline	Genetic risk score	0.71	0.67	0.68	0.65	0.73