View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Serveur académique lausannois

PLOS ONE

RESEARCH ARTICLE

Risk prediction of developing venous thrombosis in combined oral contraceptive users

Aaron McDaid[1,2], Emmanuelle Logette[3], Valérie Buchillier[3], Maude Muriset[3], Pierre Suchon[4,5], Thierry Daniel Pache[3], Goranka Tanackovic[3], Zoltán Kutalik[1,2], Joëlle Michaud[3]*

1 Institute of Social and Preventive Medicine, University Hospital of Lausanne, Lausanne, Switzerland, **2** Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland, **3** Gene Predictis SA, EPFL Innovation Park, 1015 Lausanne, Switzerland, **4** AIx Marseille Univ, INSERM, INRA, NORT, Marseille, France, **5** APHM, Hôpital de la Timone, Service d'hématologie biologique, Marseille, France

* jam@genepredictis.com

Check for updates

OPEN ACCESS

Citation: McDaid A, Logette E, Buchillier V, Muriset M, Suchon P, Pache TD, et al. (2017) Risk prediction of developing venous thrombosis in combined oral contraceptive users. PLoS ONE 12 (7): e0182041. https://doi.org/10.1371/journal.pone.0182041

Editor: Pablo Garcia de Frutos, Institut d'Investigacions Biomediques de Barcelona, SPAIN

Received: April 10, 2017

Accepted: July 11, 2017

Published: July 27, 2017

Copyright: © 2017 McDaid et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was funded by Canton de Vaud, Service de promotion économique et du commerce (https://www.vd.ch/themes/economie/developpement-economique/promotion économique/): EL, VB, MM, GT, JM. This funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. It was also funded by Gene Predictis

Abstract

Background

Venous thromboembolism (VTE) is a complex multifactorial disease influenced by genetic and environmental risk factors. An example for the latter is the regular use of combined oral contraceptives (CC), which increases the risk to develop VTE by 3 to 7 fold, depending on estrogen dosage and the type of progestin present in the pill. One out of 1'000 women using CC develops thrombosis, often with life-long consequences; a risk assessment is therefore necessary prior to such treatment. Currently known clinical risk factors associated with VTE development in general are routinely checked by medical doctors, however they are far from being sufficient for risk prediction, even when combined with genetic tests for *Factor V Leiden* and *Factor II G20210A* variants. Thus, clinical and notably genetic risk factors specific to the development of thrombosis associated with the use of CC in particular should be identified.

Methods and findings

Step-wise (logistic) model selection was applied to a population of 1622 women using CC, half of whom (794) had developed a thromboembolic event while using contraceptives. 46 polymorphisms and clinical parameters were tested in the model selection and a specific combination of 4 clinical risk factors and 9 polymorphisms were identified. Among the 9 polymorphisms, there are two novel genetic polymorphisms (rs1799853 and rs4379368) that had not been previously associated with the development of thromboembolic event. This new prediction model outperforms (AUC 0.71, 95% CI 0.69–0.74) previously published models for general thromboembolic events in a cross-validation setting. Further validation in independent populations should be envisaged.

SA: AM, ZK. Authors EL, VB, MM, GT and JM are employed by Gene Predictis SA. TDP is President of the Board of directors of Gene Predictis SA. Gene Predictis SA provided support in the form of salaries for authors (EL, VB, MM, GT and JM) but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the "author contributions" section.

Competing interests: This study was funded in part by Gene Predictis SA. EL, VB, MM, GT and JM are employed by Gene Predictis SA. TDP is President of the Board of directors of Gene Predictis SA. TDP, GT and JM are among shareholders of Gene Predictis SA and are inventors in four filled patents associated with this study (EP16203285, EP16203286, EP17163894 and EP17163897). AM and ZK received grant from Gene Predictis SA. PS declares no competing interests. The commercial affiliation to Gene Predictis SA does not alter our adherence to PLOS ONE policies on sharing data and materials.

Conclusion

We identified two new genetic variants associated to VTE development, as well as a robust prediction model to assess the risk of thrombosis for women using combined oral contraceptives. This model outperforms current medical practice as well as previously published models and is the first model specific to CC use.

Introduction

Venous thromboembolism (VTE), which includes deep vein thrombosis (DVT) and pulmonary embolism (PE), occurs in 1–2 per 1'000 individuals per year. The incidence increases with age, from 1 in 100'000 in children to 1 in 10'000 individuals in the reproductive age, 1 in 1'000 individuals at the age 50 to 60 and 1 in 100 over 75 years old [1]. VTE is a complex multifactorial disease influenced by several acquired or inherited conditions. The acquired conditions include a large number of risk factors such as surgery and trauma, prolonged immobilization, cancer, myeloproliferative disorders, and even pregnancy and post-partum [2]. Weight, age, smoking status and hormonal treatment are all additional environmental factors associated with an increased risk of VTE.

The inherited conditions include mutations in the diverse well-known clotting anticoagulant or thrombolytic factors genes, such as the *Factor V Leiden* (*F5*) gene, and the *prothrombin Factor II* (*F2*) gene. Such mutations can also be present in genes coding for proteins C and S, however despite the fact that they increase the risk of developing venous thrombosis significantly, they are rare and most of them are practically private [3]. Other likely inherited causes include a possible increase in the expression of procoagulant factors such as factor VIII, von Willebrand factor, and factors IX and XI [4]. In addition, non-O ABO blood groups, with the exception of the A2 group, were demonstrated to increase the risk of developing thrombosis. Many other additional genetic variants, present in the genes *FGG*, *GP6*, *KNG1*, *PROCR*, *SLC44A2*, *STXBP5* and *TSPAN15*, among others, were associated with an increased risk of venous thrombosis [5].

Over 100 million women worldwide use combined estroprogestative contraceptives (CC), due to their very high effectiveness in reducing the risk of unwanted pregnancy and their beneficial effect on diverse symptoms related to women's cycle. Nonetheless, these contraceptives also increase the risk of blood clotting substantially, which can ultimately lead to DVT and PE [6]. Newer generations of CC, the so-called 3rd and 4th generation CC (pills containing norgestimate, gestodene, desogestrel or drospirenone as progestin), are usually better tolerated by women but importantly, they increase the risk of developing VTE even more than the older preparations of the so-called 2nd generation (levonorgestrel containing-pills).

The incidence of thrombosis among CC users is around 1‰ per year [7]. In France alone, where over 3 million women aged 15–49 use CC, the National Agency for the Safety of Drugs and Health Products reports every year over 2'500 cases of DVT, 850 cases of PE, and 20 cases of death linked to contraceptive pills. Taking into account the incidence of thrombosis among contraceptive pill users and number of women using them, it is estimated that 22'000 DVT related to CC occur each year in Europe. Thus, one of the major challenges for healthcare professionals is to identify women at risk of developing blood clotting disease related to CC such as DVT and PE, and advise them on alternative contraception methods.

As the standard of care nowadays, prescribing physicians assess the risk of thrombosis using clinical parameters, mostly focusing on age, body mass index, smoking habits and

personal and familial history of DVT or related diseases that are known risk factors for VTE development. However, diverse studies demonstrate that clinical informations, notably familial history, are insufficient to reliably estimate risk of VTE [8, 9]. When the familial history of thrombosis is positive, physicians might use the first-level laboratory test for thrombophilia screening that includes analysis of only 2 genetic risk factors: the *F5-Leiden* and the *F2* mutations; eventually, some laboratories, also include genetic tests allowing to assess for the ABO blood group. Widely-accepted evidence of haemostatic abnormalities associated with thrombophilia includes the following parameters: antithrombin deficiency, protein S deficiency, protein C deficiency, *F5-Leiden* mutation, *F2* mutation, non-O ABO blood group and high levels of factor VIII dysfibrinogenaemia [10]. Though *F5-Leiden* and *F2* mutation are well-established risk factors for thrombosis development, they explain less then one third of the inherited risk to develop thrombosis. Precisely, *F5-Leiden* is present among 20% of patients that develop thrombosis, whereas only 6% of patients carry the *F2* mutation [10]. Therefore, genomic assessement that takes into account other polymorphisms associated with VTE development is mandatory.

Materials & methods

Population studied

The population described in this study has been designed to investigate the clinical and genetic factors that affect the risk of VTE in women taking CC. The study includes 794 female cases who have developed at least one episode of VTE while taking CC. These cases are part of the previously described PILl Genetic RIsk Monitoring (PILGRIM) study [11], in which the method used to confirm the occurence of thrombosis is defined. 828 control women were also collected from different sources: 523 are part of the PILGRIM study; 174 are part of the CoLaus study [12], 56 were recruited between 1997 and 1998 in south of France among healthy volunteers and the remaining controls were recruited by established medical clinicians between 2012 and 2016 among Swiss population. The last two groups of controls include any woman of childbearing age who was using CC at the time of collection and did not have a thrombotic event prior to sample collection time. These women were not recruited as part of thrombophilia screening and are unrelated to thrombotic patients. Nonetheless some of them have described a family history at the time of collection (19/128). The PILGRIM study controls presented a selection bias due to having been recruited as part of thrombophilia screening due to family history [11] and several variables (family history, F5 and F2) were not used as such, as described below. All control women are taking CC but have not developed VTE by the time of the genotyping investigation. This study involved human subjects and was carried out in accordance with the tenets of the Declaration of Helsinki; all participants signed an informed consent and data were anonymised. The procedures regarding the collect of PILGRIM samples were reviewed and approved by the Assistance Publique des Hopitaux de Marseille insitutional review committee. The CoLaus study was approved by the Ethics Committee of Lausanne University.

Genotyping

46 SNPs were selected according to their association with VTE development or hormone metabolism (principaly estrogens) as described in the literature. These 46 SNPs were genotyped using Illumina GoldenGate technology and assessed using Illumina BeadXpress and GenomeStudio V2011.1 software. Clusters for each SNP were curated manually and undetermined samples were further genotyped using Sanger sequencing. SNP rs1053878 was genotyped using RFLP-PCR; in more details, the DNA region was amplified with the following

primers (Forward: 5′-GCCACCGTGTCCACTACTATG-3′ and Reverse: 5′-GTCCACGCA CACCAGGTAAT-3′) and the amplicons were digested with PvuII restriction enzyme. Controls from the CoLaus cohort were previously genotyped as described [13]. For the CoLaus controls, proxys ($r^2 > 85\%$) were used for 9 SNPs (rs4572916 for rs10029715, rs8176704 for rs1053878, rs3736455 for rs13146272, rs6018 for rs1800595, rs4253417 for rs2289252, rs110 38993 for rs3136516, rs2169682 for rs7082872, rs687621 for rs8176719 and rs2069952 for rs9574). Genotyping data for rs1799963 was missing in the CoLaus study. Genotpying data for rs6025 and rs1799963 of the 523 control samples from the PILGRIM study were ignored to avoid selection bias due to having been recruited as part of thrombophilia screening due to family history [11]. Allele frequencies of the 46 SNPs in the controls were consistent with the ones observed in the European subsample of the 1000 Genomes panel [14].

Clinical characteristics

Age and smoking status were determined at the time of VTE for cases and at the time of DNA collection for controls. BMI was dertermined at the time of consultation for both cases and controls. Family history was defined as positive when at least one first-degree relative has suffered VTE. Information on family history for the 523 control women from the PILGRIM study was not used as such because of the recruitment bias [11]. All women included in this study took oral combined contraceptive.

Statistical analyses

The study population was randomly divided 10'000 times into a training set and a test set of equal size. For every sample split missing values were imputed by a random draw from the non-missing values present in the control samples. Once missing values were imputed, we applied step-wise logistic regression model selection (as long as the Akaike Information Criterion (AIC) was improved) to each training set to select variables and assign coefficients. The fitted model was then applied to the test sets to estimate the predictions of the model in an out-of-sample setting. Across the 10'000 runs the average number of selected variables was 18.1. Two variables were selected over 99.9% of the time (rs6025 and rs1799963). When a run did not select a variable (i.e. we had no evidence that the coefficient is significantly different from zero) its coefficient was set to zero, equivalent to an odds-ratio of 1. The final model coefficients are estimated as the median values of the coefficients across the 10,000 runs. This model consists of 13 variables (including 9 SNPs) with non-zero median values. The corresponding standard error (SE) for each of the 13 coefficients is the median standard error across those runs (out of the 10,000) when the variable was selected into the model. Confidence intervals and p-values were derived from the coefficients and standard errors (SEs) in the standard manner.

We compared our 9-variable genetic prediction model (including only SNPs) to previously published genetic models [8, 15]. For a fair evaluation, in each random data split (and imputation) the coefficients of each model (including our 9-variable genetic model) were estimated in the training set and the predictions were evaluated in the test set based on the Area Under the receiver operator characteristic Curve (AUC). AUC is equal to the probability that the predictor value of a positive test ranks higher than that of a negative test in order to discriminate the women at risk to the women without risk. The AUC ranges from 0.5 (50%—no predictive value) to 1 (100%—perfect discrimination) [16].The final AUC for each model is its median AUC across the 10,000 random data splits.

Results

The clinical characteristics of the population of women taking CC described here are defined in Table 1. Age distribution is similar between cases and controls as demonstrated by Wilcoxon rank-sum test (p-value = 0.1). Five parameters are statistically different between both populations, including 3 clinical variables (BMI, family history and smoking status) and two thrombophilia markers, FV-Leiden and prothrombin (F2). Although we cannot demonstrate that these differences are not due to selection bias, all five characteristics are known risk factors for VTE development, thus the minor observed differences are not surprising. The modest differences observed in our samples reinforce the current evidence that clinical information is not sufficient to distinguish women at risk to develop VTE [8, 9].

46 SNPs selected according to their association with VTE development or hormone metabolism (principaly estrogens), as described in the literature, were successfully genotyped in the 1622 women involved in this study. Familial history and genotyping data of rs6025 and rs1799963 of 523 control women were treated as missing in order to avoid an ascertainment bias. To make sure that the frequency of each SNP in our control population corresponds to the frequencies expected in a general Caucasian population, we compared the allele frequencies (AF) observed in our control population to the ones reported in 1000 genomes project [14]. The frequencies were similar (S1 Table, Fisher p-values > 0.01) for all but two SNPs (rs1593812 and rs429358, which were discarded from further analysis) suggesting that our

Table 1. Clinical characteristics of the population.

	Cases (n)	%/SD	Controls (n)	%/SD	p-value[6]
Total number	794	49%	828[1]	51%	
VTE	794				
DVT	600	75.5%			
PE	194	24.5%			
Age (mean)	32 [17–49] +	SD: ± 9.6+	31.5 [18–51] +	SD: ± 9.0+	0.1
BMI (mean)	24 [18–37] +	SD: ± 5.2+	23 [17.5–33.5] +	SD: ± 4.2+	**6.6E-06**
Family history of VTE	222	28%	19(317)[2]	15(38)[2]%	**1.7E-03**
Smoking	260	33%	206	25%	**4.7E-04**
Cancer	6	0.7%	2[3]	0.2%	0.4
Autoimmune disease	8	1%	4[3]	0.7%	0.65
Thrombophilia factors:					
Protein C	20	2.5%	7[3]	1.6%	0.1
Protein S	10	1.2%	12[3]	0%	0.15
Antithrombin	6	0.8%	2[1]	0.5%	0.4
F5-Leiden	132	16.5%	10(98)[4]	3(13)[4]%	**3.3E-09**
Prothrombin (F2)	80	10%	3(64)[5]	2(8)[4]%	**3.8E-03**

+ 95%CI (in brackets) and Standard deviation (SD) are indicated for these parameters

[1] The total number of controls differs depending on the variable as indicated in [2] and [3].

[2] This variable was set as missing or was missing for 700 control women as indicated in M&M and the total number of controls used here is 128 controls. The number indicated in brackets is the original number before correction for bias.

[3] This parameter is missing for 305 control women.

[4] This variable was set as missing for 523 control women as indicated in M&M and the total number of controls used here is 305 controls. The number indicated in brackets is the original number before correction for bias.

[5] This variable was set as missing for 697 control women as indicated in M&M and the total number of controls used here is 131 controls. The number indicated in brackets is the original number before correction for bias.

[6] p-values calculated using Wilcoxon rank-sum test to compare cases to controls.

https://doi.org/10.1371/journal.pone.0182041.t001

Table 2. Clinical and genetic parameters selected in the Pill Protect® model.

Variable	Gene (when applicable)	Effect allele (when applicable)	Mean (frequency) among cases	Mean (frequency) among controls	OR	95% CI	p-value
Family history of VTE			28[1]	15[1]	2.13	1.61–2.83	1.4E-07
Smoking			33[1]	25[1]	1.63	1.27–2.09	1.3E-04
BMI			24[2]	23[2]	1.07	1.04–1.09	3.2E-07
Age			32[2]	31.5[2]	1.01	1.001–1.03	0.03
rs6025	*F5*	A	0.09	0.02	6.46	4.04–10.3	5.8E-15
rs1799963	*F2*	A	0.05	0.01	5.32	3.01–9.31	7.39E-09
rs8176719	*ABO*	I	0.50	0.41	1.52	1.28–1.80	1.71E-06
rs2289252	*F11*	T	0.49	0.42	1.34	1.14–1.58	3.5E-04
rs1799853	*CYP2C9*	T	0.15	0.12	1.54	1.21–1.94	3.5E-04
rs9574	*PROCR*	G	0.57	0.51	1.25	1.07–1.47	0.0052
rs8176750	*ABO*	D	0.05	0.07	0.60	0.42–0.85	0.0043
rs4379368	*SUGCT*	T	0.11	0.08	1.35	1.03–1.80	0.032
rs710446	*KNG1*	G	0.46	0.43	1.22	1.04–1.43	0.016

[1] Percentage of cases and controls with the corresponding clinical factor

[2] Mean of the corresponding clinical factor across the cases or controls

https://doi.org/10.1371/journal.pone.0182041.t002

control population reflects a general Caucasian population and that the genotyping is of high quality.

Logistic regression models were fitted step-wise to find the optimal (in terms of AIC) multivariate model in the 10,000 training sets. By averaging these 10,000 models, we identified 4 clinical variables as risk factors contributing to the prediction of the risk of VTE in our population. Age, BMI, smoking status and family history were selected and had significant p-values (Table 2, p-values < 0.05). 9 out of the 44 tested SNPs were in the averaged model and also significantly associated with the development of thrombosis (Table 2). The reported p-values survive 5% false discovery rate (FDR) control. Among these nine SNPs, as expected, *F5-Leiden* (OR = 6.46, CI = [3.46–8.37]) and *F2* (OR = 5.32, CI = [2.66–7.9]) mutations are long-known risk VTE factors. Further five SNPs including rs2289252 (*F11*), rs710446 (*KNG1*), rs9574 (*PROCR*) and rs8176719/rs8176750 (tagging ABO subtypes) have been recently associated with VTE (Table 2) [17–20]. The final two of the nine polymorphisms, rs1799853 (*CYP2C9*) and rs4379368 (*SUGCT*), have not been described before to impact VTE development. No interactions among selected parameters were identified to be significant.

We estimated the out-of-sample performance of the set of 13 combined parameters as well as the clinical- and genetic-only models separately (Pill Protect® models). The ROC curves for the clinical, genetic and combined models are represented in Fig 1. The clinical model gives an AUC of 0.61 (0.58–0.64) and the genetic variables alone give an AUC of 0.68 (0.65–

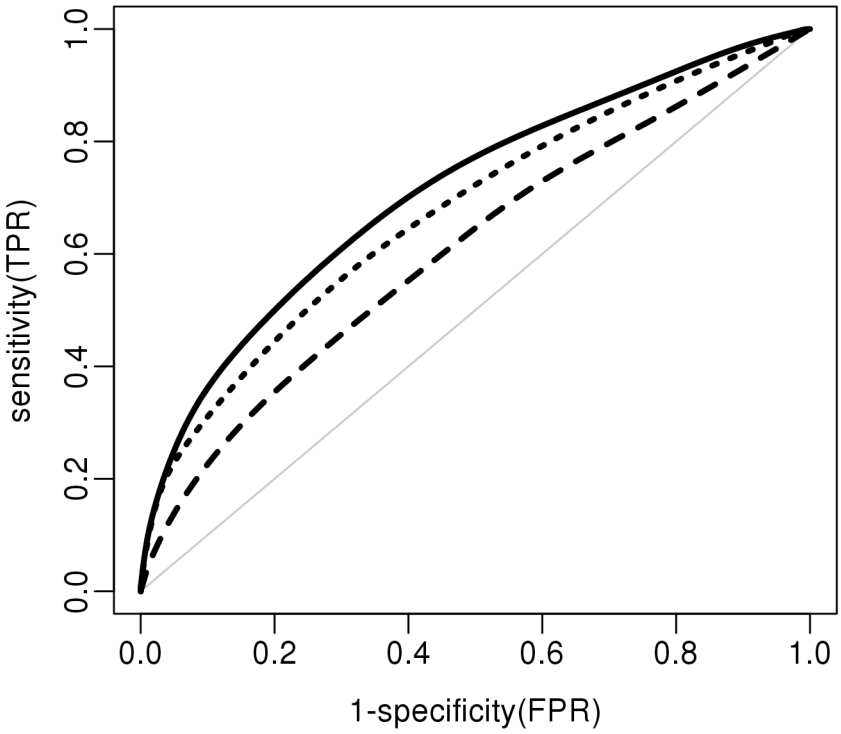

Fig 1. ROC (AUC) curves for Pill Protect models. The clinical (dashed line), genetic (dotted line) and combined models (black line) are indicated. The light grey line represents the reference line (AUC 0.5).

https://doi.org/10.1371/journal.pone.0182041.g001

0.71). Combining both clinical and genetic parameters increase the AUC to 0.71 (0.69–0.74). To compare these results with the current best practice, based on an oral anamnesis of the patient, we estimated coefficients for the clinical variables from a meta-analysis of the literature using weighted means (S2 Table, MD algorithm). In some cases, medical doctors may request a thrombophilia status that includes the genotyping information for *F5-Leiden* and *F2* mutations (rs6025 and rs1799963). We, therefore, also compared our model to a model that contains the previous clinical variables and coefficients for these two SNPs obtained from the literature (S2 Table, MD-gen algorithm). The MD model reached an AUC of 0.61 in our studied population that is similar to our clinical-only model. After adding genetic information to the MD model, the MD-gen model reaches an AUC of 0.67 in our studied population. Our combined model achieved significantly higher performance than any model we could derive from the literature (Table 3).

Table 3. Out-of sample AUC values for various published- and our models applied to our studied population.

Model	AUC	95% CI
Pill Protect® clinical model	0.61	0.58–0.64
Pill Protect® genetic model	0.68	0.65–0.70
Pill Protect® combined model	0.71	0.69–0.74
MD model	0.61	0.60–0.62
MD-gen model	0.67	0.66–0.68
Bruzelius genetics	0.65	0.63–0.68
De Haan genetics	0.64	0.62–0.68

https://doi.org/10.1371/journal.pone.0182041.t003

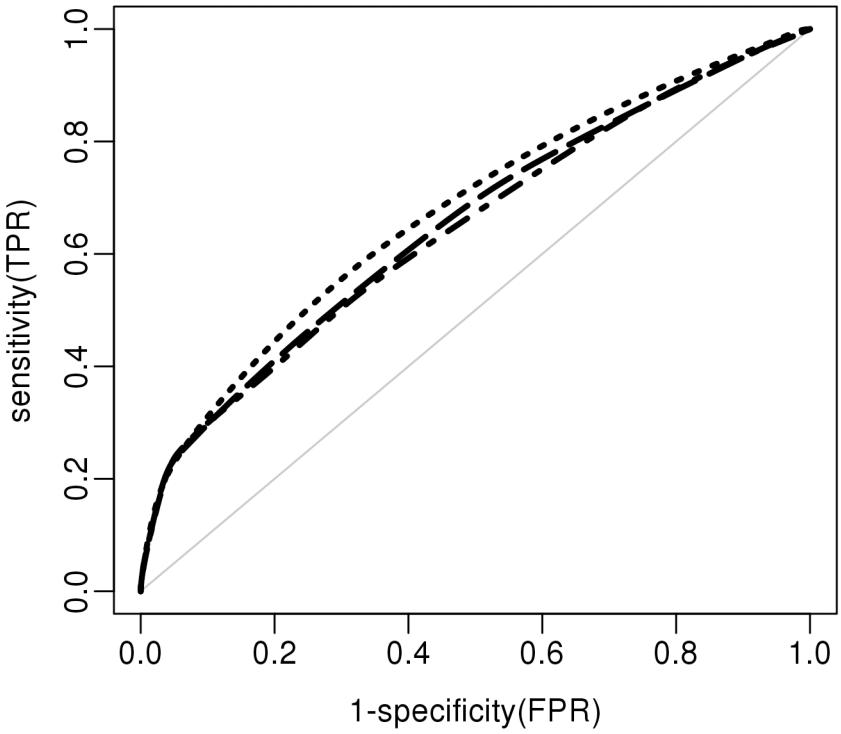

Fig 2. ROC (AUC) curves for the Pill Protect® and published genetic models. The models described in De Haan et al. (dot-dashed line), in Bruzelius et al. (long dashes) and in this paper (dotted line) are indicated. The light grey line represents the reference line (AUC 0.5).

https://doi.org/10.1371/journal.pone.0182041.g002

Previous studies have modelled VTE risk using different combinations of parameters although none are specific to the use of CC. Because the whole set of clinical parameters used by the other models was not available, we compared only the genetic models. The genetic score described by De Haan et al. [8] is based on 5 SNPs (rs6025, rs1799963, rs8176719, rs2066865, rs2036914). All of these SNPs are present in our current study, although only 3 of them are used in our final model. Applying this 5-SNP model to our population yielded an AUC of 0.64 (0.62–0.68) (Table 3 and Fig 2), which is less than the described AUC on MEGA and LETS cohorts (0.69 and 0.67 respectively) due to winners curse. The genetic score described by Bruzelius et al. [15] is based on 7 SNPs (rs6025, rs1799963, rs514659, rs2289252, rs1799810, rs710446, rs2066865) and 4 interactions. Among the 7 SNPs, 4 are present and one (rs514659) has a good proxy (rs8176719) in our Pill Protect® model, and one (rs2066865) is not part of our model but is present in our dataset. The last SNP (rs1799810) is absent from our dataset and was, therefore, not used in the comparison. However, given the small (and least significant) coefficient reported by Bruzelius, it would probably not affect significantly the performance. This is confirmed by the fact that genetic score associated with this set of six SNPs reaches an AUC of 0.65 (0.63–0.68) in our study (Table 3 and Fig 2), which is very similar to what was described by Bruzelius et al. (0.66; [0.64–0.68]). Still both AUC values are significantly below the 0.68 AUC of our 9-SNP genetic model.

Discussion

In this study, we determined a new combination of parameters that predicts the risk of VTE in women using CC. This combination outperformed significantly previously published models

as well as the clinical evaluation currently used by medical doctors. We also identified two new genetic markers associated with the development of VTE in our population.

The risk of VTE development upon CC use is presently assessed by an oral anamnesis and based on physician's experience. In the presence of clinical risk factors and/or family history, some medical doctors may request a thrombophilia status that tests for the well-established markers FV-Leiden and Prothrombin. The current incidence of 1‰ of VTE per year in CC users indicates that the risk assessement needs to be improved. Analysing genetic and clinical data for a population of women using CC, we were able to calculate a risk score that outperfoms a model that simulates the current empirical approach even when combined with additional information on FV-Leiden and Prothrombin. Our predictor Pill Protect®, including 9 genetic markers in combination with 4 clinical factors, was able to reach an out-of-sample AUC of 0.71. The use of these 9 genetic markers also outperformed the combination of markers previously published by others [8, 15]. A thrombophilia status would be complementary to the risk score approach described here, as a functional test in the presence of clinical suscpiscion, in order to take into account rare mutations such as the one present in *Protein S*, *Protein C* or *Antithrombin* genes.

Our study presents some limitations regarding the identification of rare polymorphisms and rare mutations due to the size of the population and the limited genotyping approach which was not genome-wide. The combination of the 13 genetic and clinical parameters improves the current methodology. Further investigation using a genome-wide approach on a larger cohort would be necessary to capture additional weaker effects.

We identified two polymorphisms that had not previously been associated with the development of VTE. We demonstrated that they are key in the development of VTE in CC users and future studies will address their role in the general population. The first one (rs1799853) is an established genetic markers in the field of pharmacogenetics also called *2. It affects the activity of the enzyme CYP2C9 encoded by the corresponding gene. The cytochrome CYP2C9 is involved in the metabolism of ethynyl estradiol present in most of the combined pills [21]. We hypothesize that a decreased activity of the metabolism would lead to an increased systemic level of ethynyl estradiol and therefore to an increased risk of coagulation. Interestingly, the impact of this SNP on VTE in our data seems to be stronger than that of several previously described markers.

The second novel polymorphism (rs4379368) is present in the gene coding for another enzyme SUGCT. This transferase has been previously associated with migraine susceptibility using genome-wide association study [22]. It is well established that migraine is a risk factor for arterial diseases [23] and more recently migraine has also been associated with the development of VTE [24]. The combination of migraine and hormone treatment increases further the risk of cardiovascular diseases. It remains to determine, however, whether migraine as a risk factor would improve the performance of our combined model because the information was not available in our population.

The combination of the nine SNPs identified here as well as the identification of two SNPs newly associated with VTE is specific to women who use CC due to the study design. Hence further studies will be performed to confirm that these two novel SNPs and their combination would also associate with VTE in the general population.

Our model selection has three key aspects: (1) Excluding individuals with missing values would have drastically reduced the available sample size. Hence, we chose to perform 10,000 random imputations of the missing data and averaged results over the various randomly filled data sets. One could have envisioned more sophisticated data imputation, but multivariate linear imputation of the missing data would not have improved the multivariate predictive model performance. (2) To perform out-of-sample evaluation we used a cross-validation framework,

where for each of the 10,000 sets we split the data into two equally-sized groups and used one group ('training set') to estimate the coefficients and the other group ('test set') to provide predictions for the AUC. We could have reported the results from a single data split, however that would not have used optimally the available data and would be prone to random fluctuations depending on the split. (3) The individual p-values of the selected 13 variables survive 5% FDR control (given the total number of tested variables); hence less than one of them is expected to be a false positive finding. Our cross-validation framework (with zeroing out the coefficients of unselected variables) was designed to protect our coefficient estimates from winners curse. Further work will establish meaningful clinical thresholds in order to translate this model into a clinical test.

In conclusion, we identified new genetic markers for VTE development among CC users and determined a new and robust combination of clinical and genetic parameters to predict VTE risk in CC users. Although further validation in independent populations should be envisaged, this combination outperforms all previously published genetic risk score in our cross-validation setting.

Supporting information

S1 Table. Polymorphism frequencies. MAF stands for Minor Allele Frequency. MAF in the 1000 genomes project is indicated as MAF 1000K. MAF in the studied population with missing values is indicated as MAF controls. Fisher test was used to estimate the deviation between both frequencies. Two SNPs present significant differences between both frequencies, they are indicated in bold.
(DOCX)

S2 Table. Meta-analysis of the literature. The coefficients for each indicated variable are coming from a meta-analysis of the literature.
(DOCX)

Acknowledgments

The authors also express their gratitude to the participants in the Lausanne CoLaus study and to the investigators who have contributed to the recruitment, in particular the co-PIs, Peter Vollenweider, Gérard Waeber, Martin Preisig and research nurses of CoLaus: Yolande Barreau, Anne-Lise Bastian, Binasa Ramic, Martine Moranville, Martine Baumer, Marcy Sagette, Jeanne Ecoffey, Sylvie Mermoud. The authors also thank Professor P. Morange for giving us access to the PILGRIM study samples.

Author Contributions

Conceptualization: Goranka Tanackovic, Zoltán Kutalik, Joëlle Michaud.

Data curation: Aaron McDaid, Thierry Daniel Pache, Joëlle Michaud.

Formal analysis: Aaron McDaid.

Investigation: Emmanuelle Logette, Valérie Buchillier, Maude Muriset, Joëlle Michaud.

Methodology: Aaron McDaid, Zoltán Kutalik, Joëlle Michaud.

Resources: Pierre Suchon, Thierry Daniel Pache.

Supervision: Goranka Tanackovic, Zoltán Kutalik, Joëlle Michaud.

Writing – original draft: Aaron McDaid, Goranka Tanackovic, Zoltán Kutalik, Joëlle Michaud.

Writing – review & editing: Emmanuelle Logette, Valérie Buchillier, Maude Muriset, Pierre Suchon, Thierry Daniel Pache.

References

1. Rosendaal FR. Causes of venous thrombosis. Thromb J. 2016; 14(Suppl 1):24. https://doi.org/10.1186/s12959-016-0108-y PMID: 27766050; PubMed Central PMCID: PMCPMC5056464.
2. Seligsohn U, Lubetsky A. Genetic susceptibility to venous thrombosis. N Engl J Med. 2001; 344 (16):1222–31. https://doi.org/10.1056/NEJM200104193441607 PMID: 11309638.
3. Millar DS, Johansen B, Berntorp E, Minford A, Bolton-Maggs P, Wensley R, et al. Molecular genetic analysis of severe protein C deficiency. Hum Genet. 2000; 106(6):646–53. PMID: 10942114.
4. Cushman M. Inherited risk factors for venous thrombosis. Hematology Am Soc Hematol Educ Program. 2005:452–7. https://doi.org/10.1182/asheducation-2005.1.452 PMID: 16304419.
5. Morange PE, Suchon P, Tregouet DA. Genetics of Venous Thrombosis: update in 2015. Thromb Haemost. 2015; 114(5):910–9. https://doi.org/10.1160/TH15-05-0410 PMID: 26354877.
6. Vinogradova Y, Coupland C, Hippisley-Cox J. Use of combined oral contraceptives and risk of venous thromboembolism: nested case-control studies using the QResearch and CPRD databases. BMJ. 2015; 350:h2135. https://doi.org/10.1136/bmj.h2135 PMID: 26013557; PubMed Central PMCID: PMCPMC4444976.
7. Reid R, Leyland N, Wolfman W, Allaire C, Awadalla A, Best C, et al. SOGC clinical practice guidelines: Oral contraceptives and the risk of venous thromboembolism: an update: no. 252, December 2010. Int J Gynaecol Obstet. 2011; 112(3):252–6. PMID: 21416656.
8. de Haan HG, Bezemer ID, Doggen CJ, Le Cessie S, Reitsma PH, Arellano AR, et al. Multiple SNP testing improves risk prediction of first venous thrombosis. Blood. 2012; 120(3):656–63. https://doi.org/10.1182/blood-2011-12-397752 PMID: 22586183.
9. Suchon P, Al Frouh F, Henneuse A, Ibrahim M, Brunet D, Barthet MC, et al. Risk factors for venous thromboembolism in women under combined oral contraceptive. The PILI Genetic RIsk Monitoring (PILGRIM) Study. Thromb Haemost. 2016; 115(1):135–42. https://doi.org/10.1160/TH15-01-0045 PMID: 26290123.
10. Rosendaal FR, Reitsma PH. Genetics of venous thrombosis. J Thromb Haemost. 2009; 7 Suppl 1:301–4. https://doi.org/10.1111/j.1538-7836.2009.03394.x PMID: 19630821.
11. Suchon P, Al Frouh F, Ibrahim M, Sarlon G, Venton G, Alessi MC, et al. Genetic risk factors for venous thrombosis in women using combined oral contraceptives: update of the PILGRIM study. Clin Genet. 2017; 91(1):131–6. https://doi.org/10.1111/cge.12833 PMID: 27414984.
12. Firmann M, Mayor V, Vidal PM, Bochud M, Pecoud A, Hayoz D, et al. The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. BMC Cardiovasc Disord. 2008; 8:6. https://doi.org/10.1186/1471-2261-8-6 PMID: 18366642; PubMed Central PMCID: PMCPMC2311269.
13. Kutalik Z, Johnson T, Bochud M, Mooser V, Vollenweider P, Waeber G, et al. Methods for testing association between uncertain genotypes and quantitative traits. Biostatistics. 2011; 12(1):1–17. https://doi.org/10.1093/biostatistics/kxq039 PMID: 20543033.
14. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. https://doi.org/10.1038/nature15393 PMID: 26432245; PubMed Central PMCID: PMCPMC4750478.
15. Bruzelius M, Bottai M, Sabater-Lleal M, Strawbridge RJ, Bergendal A, Silveira A, et al. Predicting venous thrombosis in women using a combination of genetic markers and clinical risk factors. J Thromb Haemost. 2015; 13(2):219–27. https://doi.org/10.1111/jth.12808 PMID: 25472531.
16. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006; 27(8):861–74. https://doi.org/10.1016/j.patrec.2005.10.010
17. Heit JA, Armasu SM, Asmann YW, Cunningham JM, Matsumoto ME, Petterson TM, et al. A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q. J Thromb Haemost. 2012; 10(8):1521–31. https://doi.org/10.1111/j.1538-7836.2012.04810.x PMID: 22672568; PubMed Central PMCID: PMCPMC3419811.
18. Li Y, Bezemer ID, Rowland CM, Tong CH, Arellano AR, Catanese JJ, et al. Genetic variants associated with deep vein thrombosis: the F11 locus. J Thromb Haemost. 2009; 7(11):1802–8. https://doi.org/10.1111/j.1538-7836.2009.03544.x PMID: 19583818.

19. Morange PE, Oudot-Mellakh T, Cohen W, Germain M, Saut N, Antoni G, et al. KNG1 Ile581Thr and susceptibility to venous thrombosis. Blood. 2011; 117(13):3692–4. https://doi.org/10.1182/blood-2010-11-319053 PMID: 21270443.

20. Saposnik B, Reny JL, Gaussem P, Emmerich J, Aiach M, Gandrille S. A haplotype of the EPCR gene is associated with increased plasma levels of sEPCR and is a candidate risk factor for thrombosis. Blood. 2004; 103(4):1311–8. https://doi.org/10.1182/blood-2003-07-2520 PMID: 14576048.

21. Wang B, Sanchez RI, Franklin RB, Evans DC, Huskey SE. The involvement of CYP3A4 and CYP2C9 in the metabolism of 17 alpha-ethinylestradiol. Drug Metab Dispos. 2004; 32(11):1209–12. https://doi.org/10.1124/dmd.104.000182 PMID: 15304426.

22. Anttila V, Winsvold BS, Gormley P, Kurth T, Bettella F, McMahon G, et al. Genome-wide meta-analysis identifies new susceptibility loci for migraine. Nat Genet. 2013; 45(8):912–7. https://doi.org/10.1038/ng.2676 PMID: 23793025; PubMed Central PMCID: PMCPMC4041123.

23. Sacco S, Pistoia F, Degan D, Carolei A. Conventional vascular risk factors: their role in the association between migraine and cardiovascular diseases. Cephalalgia. 2015; 35(2):146–64. https://doi.org/10.1177/0333102414559551 PMID: 25505017.

24. Peng KP, Chen YT, Fuh JL, Tang CH, Wang SJ. Association between migraine and risk of venous thromboembolism: A nationwide cohort study. Headache. 2016; 56(8):1290–9. https://doi.org/10.1111/head.12885 PMID: 27411732.