BMJ

BMJ 2011;343:d4656 doi: 10.1136/bmj.d4656

RESEARCH

Development and validation of risk prediction algorithm (QThrombosis) to estimate future risk of venous thromboembolism: prospective cohort study

Julia Hippisley-Cox *professor of clinical epidemiology and general practice* , Carol Coupland *associate professor in medical statistics*

Division of Primary Care, University Park, Nottingham NG2 7RD, UK

Abstract

Objectives To derive and validate a new clinical risk prediction algorithm (QThrombosis, www.qthrombosis.org) to estimate individual patients' risk of venous thromboembolism.

Design Prospective open cohort study using routinely collected data from general practices. Cox proportional hazards models used in derivation cohort to derive risk equations evaluated at 1 and 5 years. Measures of calibration and discrimination undertaken in validation cohort.

Setting 564 general practices in England and Wales contributing to the QResearch database.

Participants Patients aged 25-84 years, with no record of pregnancy in the preceding 12 months or any previous venous thromboembolism, and not prescribed oral anticoagulation at baseline: 2 314 701 in derivation cohort and 1 240 602 in validation cohort.

Outcomes Incident cases of venous thromboembolism, either deep vein thrombosis or pulmonary embolism, recorded in primary care records or linked cause of death records.

Results The derivation cohort included 14 756 incident cases of venous thromboembolism from 10 095 199 person years of observation (rate of 14.6 per 10 000 person years). The validation cohort included 6913 incident cases from 4 632 694 person years of observation (14.9 per 10 000 person years). Independent predictors included in the final model for men and women were age, body mass index, smoking status, varicose veins, congestive cardiac failure, chronic renal disease, cancer, chronic obstructive pulmonary disease, inflammatory bowel disease, hospital admission in past six months, and current prescriptions for antipsychotic drugs. We also included oral contraceptives, tamoxifen, and hormone replacement therapy in the final model for women. The risk prediction equation explained 33% of the variation in women and 34% in men in the validation cohort evaluated at 5 years The D statistic was 1.43 for women and 1.45 for men. The receiver operating curve statistic was 0.75 for both sexes. The model was well calibrated.

Conclusions We have developed and validated a new risk prediction model that quantifies absolute risk of thrombosis at 1 and 5 years. It can help identify patients at high risk of venous thromboembolism for prevention. The algorithm is based on simple clinical variables which the patient is likely to know or which are routinely recorded in general practice records. The algorithm could be integrated into general practice clinical computer systems and used to risk assess patients before hospital admission or starting medication which might increase the risk of venous thromboembolism.

Introduction

Each year, over 25 000 people in England die from venous thromboembolism developed in hospital.[1] This number is more than the combined total of deaths from breast cancer, AIDS, and traffic accidents, and more than 25 times the number of deaths from meticillin resistant *Staphylococcus aureus*.[1] Venous thromboembolism is an important and preventable cause of morbidity and mortality,[2] with almost a third of survivors experiencing long term effects.[3 4] To improve survival and to prevent complications, the occurrence of venous thromboembolism needs to be reduced.[5]

Independent risk factors for venous thromboembolism have been identified[6-8] and prophylaxis exists for high risk individuals.[9] In 2010, the UK National Institute for Health and Clinical Excellence (NICE) issued new guidance to improve the prevention of venous thromboembolism for patients by use of cost effective interventions.[10] The guidance highlighted the need for new research to develop and validate risk prediction models to predict absolute risk of venous thromboembolism, taking account of patient factors, comorbidity, and concurrent medication, and incorporating venous thromboembolism events arising in a community setting.[10] It recommended the development of new risk prediction models by use of primary care research databases.[10] Such databases, with linked cause of death data, contain robust information on many of the relevant exposures and outcomes.[11] They are also represent the populations where such a model is likely to be used to inform treatment decisions, including the use of prophylaxis[10] and the use of medication which might increase venous thromboembolism risk.

Correspondence to: J Hippisley-Cox juliahippisleycox@gmail.com

Once validated, suitable clinical risk prediction models can be integrated into clinical computer systems to help systematically identify individuals at high risk of venous thromboembolism and to alert clinicians to those who might benefit from interventions.[12 13] Although there are currently no validated algorithms to predict risk for venous thromboembolism designed for use in primary care, computerised clinical decision support could improve appropriate use of thromboprophylaxis in a hospital setting.[14]

We developed and validated a risk prediction algorithm (QThrombosis) to estimate the individual absolute risk of venous thromboembolism to target prophylaxis to the people most likely to benefit.[10 15] The algorithm could also be used to inform treatment decisions regarding use of medication that may increase risk of venous thromboembolism, such as the combined oral contraceptive pill,[7 8 16-18 11] hormone replacement therapy,[7 8 19-21] and antipsychotic medication.[8] We developed the algorithm to estimate the risk of individuals developing venous thromboembolism up to five years into the future, rather than the current risk of having venous thromboembolism in patients presenting with symptoms such as swollen legs.

Methods

Study design and data source

We did a prospective open cohort study in a large population of primary care patients using the QResearch database (version 29). We included all general practices in England and Wales that had been using their computer systems with Egton Medical Information Systems (EMIS) for at least a year. We randomly allocated two thirds of practices to a derivation dataset and the remaining third to a validation dataset using the random number utility in Stata.

We identified an open cohort of patients aged 25-84 years drawn from patients registered with general practices between 1 January 2004 and 30 April 2010. We excluded patients who did not have a postcode related Townsend score, patients with a history of venous thromboembolism, and those who had been prescribed oral anticoagulation drugs at any time before the study start date. We also excluded women with recorded evidence of pregnancy in the preceding 12 months, because the risk of venous thromboembolism in pregnancy is likely to require separate analysis using data where gestation, mode, and date of delivery dates are well recorded. Entry to the cohort was the latest of two dates: the study start date (1 January 2004) or 12 months after the patient registered with the practice. We censored patients at the earliest date of a diagnosis of venous thromboembolism, death, deregistration with the practice, last upload of computerised data, five years after study entry, or the study end date (30 April 2010).

Clinical outcomes

Our clinical outcome was incident diagnosis of venous thromboembolism including either deep vein thrombosis or pulmonary embolism, recorded either on the patients' general practice record using the relevant Read diagnostic codes or on their linked Office of National Statistics cause of death record using the relevant International Classification of Diseases (ICD)-9 codes or ICD-10 diagnostic codes. We used codes similar to those used in previous studies where possible.[16]

Risk factors

We examined predictor variables based on established risk factors for venous thromboembolism, focusing on those that are likely to be recorded in the patient's electronic record and that the patient is likely to know (box 1).

We defined recent events as events recorded in the 12 months before study entry, and we categorised them as: within the past 6 months, 6-12 months ago, or not in the past 12 months. We defined current medication use as at least one prescription in the 30 days preceding study entry for antipsychotic drugs and tamoxifen, since most prescriptions are issued monthly. We defined current use of oral contraceptives and hormone replacement therapy as at least one prescription in the past six months, since most prescriptions are issued for this period.

Derivation and validation of the models

We developed and validated the risk prediction algorithm using established methods.[10 12 30-33] We used multiple imputation to replace missing values for body mass index and smoking status, and used these values in our main analyses.[34-37] We carried out five imputations. We used Cox proportional hazards models to estimate the coefficients for risk factors for men and women separately, using robust variance estimates to allow for the clustering of patients within general practices. We used Rubin's rules to combine the results across the imputed datasets.[38] We used fractional polynomials to model non-linear risk relations with continuous variables.[39] We fitted a full model initially and retained variables if they had a hazard ratio of more than 0.80 or less than 1.20 (for binary variables) and were significant at the 0.01 level. To simplify the model, we then focused on variables for the most common conditions and medications and combined similar variables with comparable hazard ratios where possible. We compared Akaike information criteria for models with and without Townsend score to determine the score's contribution.

We examined interactions between predictor variables and age. We used the regression coefficients for each variable from the final model as weights, which we combined with the baseline survivor function evaluated for each year up to five years to derive risk equations at each year of follow up.[40] We estimated the baseline survivor function based on zero values of centred continuous variables, with all binary predictor values set to zero, using the methods implemented in Stata.

We used multiple imputation in the validation cohort to replace missing values for body mass index and smoking. We then applied the risk equations for men and women obtained from the derivation cohort to the validation cohort and calculated measures of discrimination. We calculated the R^2 statistic[41] (estimated variation in time to venous thromboembolism), the D statistic[42] (a measure of discrimination where higher values indicate better discrimination), and the area under the receiver operating characteristic curve (receiver operating curve statistic) at one and five years. We assessed calibration (comparing the mean predicted risks at one and five years with the observed risk by tenth of predicted risk. We obtained the observed risk by using the Kaplan-Meier estimate evaluated at one and five years.

We applied the algorithm to the validation cohort to define the thresholds for the 0.5%, 1%, 5%, and 10% of patients at highest estimated risk of venous thromboembolism at one and five years. We used all the available data on the database to maximise the power and generalisability of the results. We used Stata (version 11) for all analyses.

Box 1 Predictor variables based on established risk factors for venous thromboembolism

- Age[7 8] (continuous)
- Body mass index[7 8 11 22] (continuous)
- Smoking status[7 8 22 11] (non-smoker; ex-smoker; light, moderate, or heavy smoker)
- Townsend deprivation score[7 8] (continuous)
- Varicose veins[6 8] (yes/no)
- Congestive cardiac failure[8 23 24 25] (yes/no)
- Rheumatoid arthritis[7 24] (yes/no)
- Chronic renal disease[7 8] (yes/no)
- Inflammatory bowel disease[8 24 26 27] (yes/no)
- Cancer[6-8 23 28] (lung, gastrointestinal, pancreas, renal, breast, prostate, other)
- Recent hospital admission[6 8] (yes/no)
- Recent hip fracture or hip surgery (or both)[8] (yes/no)
- Current use of antipsychotic drugs[7 8] (none, atypical, typical)
- Current use of tamoxifen[8 28] (yes/no)
- Current use of hormone replacement therapy[7 8 19-21] (none, equine or non-equine hormone replacement therapy)
- Use of antiplatelets (yes/no)
- Cardiovascular disease[6-8] (stroke, transient ischaemic attack, or coronary heart disease)
- Atrial fibrillation (yes/no)
- Asthma[7 8] (yes/no)
- Chronic obstructive pulmonary disease[7 8] (yes/no)
- Family history of venous thromboembolism[29] (yes/no)

Results

Overall study population

Overall, 564 QResearch practices in England and Wales met our inclusion criteria, of which 375 were randomly assigned to the derivation dataset with the remainder assigned to a validation cohort. We identified 2 598 829 patients aged 25-84 years in the derivation cohort. We excluded 152 719 (5.9%) patients without a recorded Townsend score, 26 211 (1.0%) on oral anticoagulation treatment, 85 306 (3.3%) with evidence of pregnancy in the preceding 12 months, 168 (0.01%) with a missing date for venous thromboembolism, and 19 724 (0.8%) with a history of venous thromboembolism. These exclusions left 2 314 701 patients for analysis.

We identified 1 354 517 patients aged 25-84 years in the validation cohort. We excluded 44 973 (3.3%) patients without a recorded Townsend score, 13 815 (1.0%) on oral anticoagulation, 44 318 (3.3%) with evidence of pregnancy in the preceding 12 months, 113 with a missing date for venous thromboembolism, and 10 696 (0.6%) with a history of venous thromboembolism. These exclusions left 1 240 602 patients for analysis.

The baseline characteristics of each cohort were similar (table 1). As in previous studies,[12 13 30] the patterns of missing data supported the use of multiple imputation to replace missing values for smoking and body mass index (not shown, available from the authors).

Rates of incident venous thromboembolism

In the derivation cohort, we identified 14 756 incident cases of venous thromboembolism arising from 10 095 199 person years of observation, giving a rate of 14.6 per 10 000 person years. Of these 14 756 cases, 5799 (39.3%) were pulmonary embolism and 8957 (60.7%) were deep vein thrombosis. We identified 14 039 (95.1%) cases of venous thromboembolism from general practice records and 717 (4.9%) solely from linked Office of National Statistics death records.

In the validation cohort, we identified 6913 incident cases of venous thromboembolism arising from 4 632 694 person years of observation, giving a rate of 14.9 per 10 000 person years. Of these 6913 cases, we identified 6526 (94.4%) from general practice records and 387 (5.6%) solely from linked Office of National Statistics death records.

Predictor variables

After fitting a full model, we combined variables that were similar with comparable hazard ratios where possible. For example, we combined the various types of cancer into one variable (any cancer) since the hazard ratios for the individual types of cancer were similar and the 95% confidence intervals overlapped. Similarly, we combined the variables for current use of typical and atypical antipsychotic drugs into one variable and equine and non-equine hormone replacement therapy into another variable. For example, the adjusted hazard ratio for women was 1.27 (95% confidence interval 1.16 to 1.40) for typical antipsychotic drugs and 1.69 (1.36 to 2.11) for atypical antipsychotic drugs. The adjusted hazard ratio for both types of antipsychotic drugs combined was 1.55 (1.32 to 1.81). We also combined two variables for hip fracture or operation and recent hospital admission.

Table 2 shows the predictor variables selected for the final simplified models for men and women. We found no significant interaction terms with age.

The risk of venous thromboembolism in women was linked to increasing age, body mass index, and quantity of cigarettes smoked every day. Risks were also raised in women with varicose veins (40% increase), congestive cardiac failure (40%), chronic renal disease (60%), any cancer (85%), chronic obstructive airways disease (41%), inflammatory bowel disease

(45%), and those admitted to hospital in the past six months (86%).

The risk of venous thromboembolism increased in women prescribed antipsychotic drugs, (55% increase), oral contraceptives (33%), hormone replacement therapy (20%), and tamoxifen (48%). Although the risk of venous thromboembolism rose with increasing levels of deprivation, the effect was not marked and did not substantially affect the model fit. Therefore, we did not include deprivation in the final model.

Our final model for men included similar variables except for those variables specific to women (hormone replacement therapy, oral contraceptives, and tamoxifen). The magnitudes of the adjusted hazard ratios were generally similar to those found for women.

In our multivariate analysis, we found no significant change in risk in men or women for: current antiplatelet therapy, atrial fibrillation, cardiovascular disease, asthma, or family history of venous thromboembolism (although the number of patients with family history of venous thromboembolism recorded was very low).

Validation

Discrimination

The validation statistics (table 3) showed that the risk prediction algorithm explained 33% of the variation in time to venous thromboembolism for women and 34% for men in the validation cohort when evaluated over five years. At five years, the D statistic was 1.43 for women and 1.45 for men. The receiver operating curve statistic was 0.75 in both sexes. The performance of the algorithm over five years was marginally better than the performance over one year (table 3).

Calibration

The figure compares the mean predicted risks with the observed risks at one and five years, by tenths of the distribution of predicted risk, to assess the calibration of the model in the validation cohort. We found a close similarity between the mean predicted risks and the observed risks at one and five years within every tenth of predicted risk, indicating that the algorithm was well calibrated. For example, in the top tenth of predicted risk for women, the mean predicted five year risk was 2.78% and the observed risk was 2.70%, giving a ratio of 1.03. For men, the corresponding figures were 2.46% and 2.35% , giving a ratio of 1.05.

Thresholds and risk stratification

Since the QThrombosis algorithm is new (box 2), we had no established thresholds for defining a high risk group. Therefore, we calculated cut-offs to define the top 0.5%, 1%, 5%, and 10% for absolute risk of venous thromboembolism based on the estimated risks at one and five years in the validation cohort (men and women combined).

Table 4 shows the cut-offs, and the total number of patients that would fall into each group based on the one and five year risk. It also shows the total number of incident cases of venous thromboembolism occurring in the groups and the overall total number of cases of venous thromboembolism. For example, the 90th centile defined a high risk group with a five year risk score of more than 15 per 1000. There were 2441 new cases of venous thromboembolism within this group over five years, which accounted for 35% of all new cases of venous thromboembolism. In other words, the sensitivity was 35% for this cut-off. The positive predictive value at this cut-off was 2%. The 99th centile defined a high risk group with a five year risk score of more than 38 per 1000. There were 350 new cases of venous thromboembolism in this group over five years. The sensitivity based on the 99th centile was 5% and the positive predictive value was 2.8%.

Clinical example 1

A 39 year old woman, who is a heavy smoker, has a body mass index of 36.7 and a history of varicose veins, and is currently taking the oral contraceptive pill. She has a one year thrombosis risk of 0.2% and a five year risk of 1.1%. A similar woman not currently prescribed the oral contraceptive pill has a one year risk of 0.15% and a five year risk of 0.9%.

Clinical example 2

A 54 year old woman, who is a moderate smoker, has a body mass index of 36.7, a history of varicose veins, and chronic obstructive airways disease. She has been admitted to hospital in the past six months, and is currently prescribed hormone replacement therapy. She has a one year thrombosis risk of 0.8% and a five year risk of 4.5%. A similar woman not currently prescribed hormone replacement therapy has a one year risk of 0.7% and a five year risk of 3.8%.

Clinical example 3

A 78 year old man, who is a heavy smoker, has a body mass index of 29.4, chronic obstructive airways disease, chronic renal disease, and congestive cardiac failure, and is currently prescribed an antipsychotic drug. He has a one year thrombosis risk of 4.2% and a five year risk of 20.7%. A similar man not currently prescribed antipsychotic drugs has a one year risk of 2.3% and a five year risk of 11.8%.

Clinical example 4

An 80 year old woman, who is an ex-smoker, has a body mass index of 27.6, congestive cardiac failure, chronic renal disease, and breast cancer. She has been admitted to hospital in the past six months and is currently prescribed tamoxifen and an antipsychotic drug. She has a one year thrombosis risk of 7.1% and a five year risk of 33.5%. A similar woman not prescribed tamoxifen or antipsychotic drugs has a one year risk of 3.2% and a five year risk of 16.4%.

Discussion

Venous thromboembolism is a common, lethal condition which can be prevented with the appropriate use of effective interventions in high risk individuals.[10] We have developed and validated a new risk prediction algorithm designed to predict the absolute risk of venous thromboembolism in a large representative primary care population. This algorithm could be used to identify patients at highest risk of venous thromboembolism and those most likely to benefit from intervention, such as change in medication, mechanical prophylaxis, or thromboprophylactic medication. The algorithm is not, however, designed to assess the current risk of venous thromboembolism—for example, in a symptomatic patient presenting with a swollen leg. We think our study provides new information which helps address gaps in evidence highlighted by recent NICE guidance.[10]

Although our study has focused on the formal development and validation of the algorithm, we can see several clinical situations where the algorithm embedded in a clinical risk calculator might be useful. Firstly, it could be used to identify patients at

Box 2 QThrombosis web calculator

A simple web calculator implements the QThrombosis algorithm and is publicly available. It also has the open source software for download (www.qthrombosis.org)

increased risk of venous thromboembolism on or before hospital admission or before long haul flights, so that prophylaxis can be considered in a more systematic way.[10]

Secondly, the algorithm could be used when considering medication which might increase venous thromboembolism risk, such as the oral contraceptive pill, tamoxifen, hormone replacement therapy, or antipsychotic drugs. For example, a woman might be interested to know her absolute level of risk and how it might change with medication, and this risk can be assessed against the expected benefits of the medication.

Thirdly, the algorithm could be used to identify high risk groups of patients suitable for further testing, closer monitoring, or preventative treatment. Pragmatic randomised trials can establish the true benefits of preventive treatment in individuals at high risk of thromboembolic events and the exact cut-offs for treatment where the benefits will outweigh the risks. We have presented clinical examples of estimated absolute risk at one and five years although the algorithm can calculate risks at 1, 2, 3, 4, or 5 years; therefore, relevant risks can be calculated depending on the clinical situation and intervention being considered.

Other studies of risk models

While other studies have examined risk factors for venous thromboembolism, studies specifically designed to develop and validate risk prediction algorithms for venous thromboembolism are lacking. We identified only one cohort study which developed a risk prediction score for venous thromboembolism over a 10 year observation period starting in 1993. Heinemann and colleagues reported the Bavarian thromboembolic risk which examined venous thromboembolism risk in 4337 young women (18-55 years) using genetic information and self reported outcomes via a questionnaire followed up by a telephone interview.[43] The sample was limited by size since it included only 34 cases of venous thromboembolism. However, it was the first such study and included a careful analysis of many candidate variables including age, body mass index, varicose veins, use of hormone replacement therapy, family history of cardiovascular disease, oral contraceptive use, smoking status, educational attainment, reproductive history, and laboratory measurements, but no acute events such as surgery or immobilisation. Most of these variables were not,important and did not improve prediction, or were considered impractical to use. Hence, their final model included only three variables (age, body mass index, and family history) in addition to genetic information. While the Bavarian thromboembolic risk algorithm was not published or validated, their analysis and commentary did support the use of such an algorithm for future risk stratification. The study also concluded that the algorithm could be based on clinical data alone at least until better genetic information is more available.[43]

More recently, a predictive model for chemotherapy associated thrombosis has been developed in 2701 patients with selected cancers undergoing chemotherapy in a hospital outpatient setting followed up over a median of 2.5 months.[44] The algorithm was based on five predictive variables: site of cancer, body mass index, use of erythropoiesis stimulating agent, haemoglobin more than 100g/L, and leucocyte count of more than 11×10^9/L. The same team validated the predictive model in 1365 patients from the same study[44] with a C statistic of 0.7 (which is lower than the receiver operative curve value of 0.75 from our present study). Our study is more suitable for use in a general primary care population.

Strengths

The strengths and weaknesses of general practice databases for the development and validation of clinical risk prediction algorithms have been described in detail elsewhere.[12 32] In summary, key strengths include size; duration of follow-up; representativeness; and lack of selection, recall, and respondent bias. UK general practices have good levels of accuracy and completeness in recording clinical diagnoses and prescribed medications,[45 46]

We think our study has good face validity since it has been undertaken in the setting where most patients in the UK are assessed, treated, and followed up. Our study also includes established exposures known to increase risk of venous thromboembolism. We have been able to include a long list of predictor variables and establish which factors remain independent after adjustment and their relative importance. The strength of the association between cancer and risk of venous thromboembolism is similar to that reported elsewhere.[28] The size of our study is particularly important since venous thromboembolism is uncommon in certain population groups.

We have also developed the algorithm in one cohort and validated it in a separate cohort that represents the patients likely to be considered for preventative measures. Although the validation cohort is derived from general practices using the same clinical computer system (EMIS), they were physically discrete. Also, since this computer system is used in over half of UK general practices, our results are likely to generalise well. A separate independent validation study using another general practice database is planned (to be undertaken by another independent team).

Limitations

Limitations included lack of formally adjudicated outcomes, information bias, potential for missing data, and residual confounding. Our database has linked cause of death from the UK Office of National Statistics, and our study is therefore likely to have picked up most cases of venous thromboembolism thereby minimising ascertainment bias. Patients who die from venous thromboembolism in hospital will have the cause of death recorded on their death certificate and therefore will be included on the linked cause of death data. Other patients who have been diagnosed with but do not die from venous thromboembolism in hospital will have the information recorded in hospital discharge letters which are sent to the general practitioner and then entered into the patient's electronic record.

The recorded clinical diagnoses of venous thromboembolism were not independently verified for the study, but in other studies the inclusion of possible venous thromboembolism cases gave similar adjusted odds ratios to that based on confirmed cases.[11] The observed incidence rate in our population was close to the 11.7 per 10 000 person years reported elsewhere[47] and in line with the estimated 5-10 per 10 000 person years for women of reproductive age.[48] Although we are reliant on the accuracy

of information recorded by primary care physicians, we think that the quality of information is likely to be good since previous studies have validated similar outcomes and exposures using questionnaire data and found levels of completeness and accuracy to be high.[11]

Although genes associated with increased risk of venous thromboembolism have been identified,[49] such information is not routinely collected or recorded in electronic records and hence we were unable to include genetic information in our analysis. We concluded that although the absence of genetic information may be a limitation of our study, it is unlikely to have a major effect on the use of the QThrombosis algorithm, because such information is unlikely to be available for use in a general risk calculator.

We designed this study to identify patients at high risk of venous thromboembolism who might require prophylaxis before a hospital procedure or other event rather than to help diagnose it in symptomatic patients. Similarly, our study was not designed to estimate how the risk of thromboembolism might change during the course of a hospital episode. Further analysis of this kind might be possible once the QResearch database is linked to secondary care data.

Furthermore, our findings cannot offer hypotheses on individual mechanisms in the genesis of thromboembolism for comorbid conditions or individual drugs. For some clinical conditions (such as cancer), biological mechanisms have already been proposed.[50] Even in a dataset of this size, the number of patients exposed to some individual drugs is too small to estimate separate hazard ratios. We cannot rule out the possibility of residual confounding by indication, which is a further barrier to making definitive risk comparisons between individual drugs. However, in terms of our overall findings, the direction and the magnitude of the hazard ratios associated with risk of venous thromboembolism are broadly in line with those reported elsewhere,[29] with largest effects seen in association with cancer,[28] recent hip fracture, hip surgery, and hospital admission.[15] Family history was not associated with risk of venous thromboembolism, yet this finding is likely to reflect the small numbers of patients with this information recorded.

We thank the general practices who contributed to the QResearch database, and EMIS for their expertise in establishing, developing, and supporting the database.

Funding: There was no external funding for the majority of the work undertaken on this project, which JHC and CC undertook in their ClinRisk roles. This analysis arose from a study of risks and benefits of hormone replacement therapy which was originally funded by David Stables (medical director of EMIS) in 2008, since venous thromboembolism was one of the outcomes.

Competing interests: All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any additional organisation for the submitted work; JHC is professor of clinical epidemiology at the University of Nottingham and unpaid director of QResearch, a not-for-profit organisation which is a joint partnership between the University of Nottingham and EMIS (commercial IT supplier for 60% of general practices in the UK). JHC is also a paid director of ClinRisk Limited, which produces open and closed source software to ensure the reliable and updatable implementation of clinical risk algorithms within clinical computer systems to help improve patient care. CC is associate professor of medical statistics at the University of Nottingham and a paid consultant statistician for ClinRisk Limited; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: The project has been independently reviewed in accordance with the QResearch agreement with Trent multicentre ethics committee.

Contributors: JHC initiated the study, undertook the literature review, data extraction, data manipulation, and primary data analysis, and wrote the first draft of the paper. CC contributed to the design, analysis, interpretation, and drafting of the paper.

Data sharing: The algorithms presented in this study will be released as open source software under the GNU Lesser General Public Licence version 3. The open source software allows use by anyone without charge under the terms of the GNU lesser General Public License version 3. Closed source software can be licensed at a fee.

1 Committee, HoCH. The prevention of thromboembolism in hospitalised patients. 1st ed. House of Commons, 2005: 112.
2 Kyrle PA, Eichinger S. Deep vein thrombosis. *Lancet* 2005;365:1163-74.
3 Prandoni P, Lensing AW, Cogo A, Cuppini S, Villalta S, Carta M, et al. The long-term clinical course of acute deep venous thrombosis. *Ann Intern Med* 1996;125:1-7.
4 Mohr DN, Silverstein MD, Heit JA, Petterson TM, O'Fallon WM, Melton LJ. The venous stasis syndrome after deep venous thrombosis or pulmonary embolism: a population-based study. *Mayo Clin Proc* 2000;75:1249-56.
5 Heit JA, O'Fallon WM, Petterson TM, Lohse CM, Silverstein MD, Mohr DN, et al. Relative impact of risk factors for deep vein thrombosis and pulmonary embolism: a population-based study. *Arch Intern Med* 2002;162:1245-8.
6 Heit JA, Silverstein MD, Mohr DN, Petterson TM, O'Fallon WM, Melton LJ 3rd. Risk factors for deep vein thrombosis and pulmonary embolism: a population-based case-control study. *Arch Intern Med* 2000;160:809-15.
7 Hippisley-Cox J, Coupland C. Unintended effects of statins in men and women in England and Wales: population based cohort study using the QResearch database. *BMJ* 2010;340:c2197.
8 Parker C, Coupland C, Hippisley-Cox J. Antipsychotic drugs and risk of thromboembolism: nested case-control study. *BMJ* 2010;341:c4245.
9 Geerts WH, Heit JA, Clagett GP, Pineo GF, Colwell CW, Anderson FA Jr, et al. Prevention of venous thromboembolism. *Chest* 2001;119(suppl 1):132-75S.
10 National Institute for Clinical Excellence. Venous thromboembolism: reducing the risk. Reducing the risk of venous thromboembolism (deep vein thrombosis and pulmonary embolism) in patients admitted to hospital. NICE guideline 92. National Institute for Clinical Excellence, 2010:50.
11 Lawrenson R, Todd JC, Leydon GM, Williams TJ, Farmer RD. Validation of the diagnosis of venous thromboembolism in general practice database studies. *Br J Clin Pharmacol* 2000;49:591-6.
12 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336:1475.
13 Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 2009;338:b880.
14 Galanter WL, Thambi M, Rosencranz H, Shah B, Falck S, Lin F-J, et al. Effects of clinical decision support on venous thromboembolism risk assessment, prophylaxis, and prevention at a university teaching hospital. *Am J Health Syst Pharm* 2010;67:1265-73.
15 Heit JA. The epidemiology of venous thromboembolism in the community: implications for prevention and management. *J Thrombosis Thrombolysis* 2006;21:23-9.
16 Lidegaard O, Lokkegaard E, Svendsen AL, Agger C. Hormonal contraception and risk of venous thromboembolism: national follow-up study. *BMJ* 2009;339:b2890.
17 Jick SS, Kaye JA, Russmann S, Jick H. Risk of nonfatal venous thromboembolism with oral contraceptives containing norgestimate or desogestrel compared with oral contraceptives containing levonorgestrel. *Contraception* 2006;73:566-70.
18 Van Hylckama Vlieg A, Helmerhorst FM, Vandenbroucke JP, Doggen CJM, Rosendaal FR. The venous thrombotic risk of oral contraceptives, effects of oestrogen dose and progestogen type: results of the MEGA case-control study. *BMJ* 2009;339:b2921.
19 Smith NL, Heckbert SR, Lemaitre RN, Reiner AP, Lumley T, Weiss NS, et al. Esterified estrogens and conjugated equine estrogens and the risk of venous thrombosis. *JAMA* 2004;292:1581-7.
20 Canonico M, Plu-Bureau G, Lowe GDO, Scarabin P-Y. Hormone replacement therapy and risk of venous thromboembolism in postmenopausal women: systematic review and meta-analysis. *BMJ* 2008;336:1227.
21 Roberts H. Type of hormone replacement therapy and risk of venous thromboembolism. *BMJ* 2008:336:1203.
22 Lutsey PL, Virnig BA, Durham SB, Steffen LM, Hirsch AT, Jacobs DR Jr, et al. Correlates and consequences of venous thromboembolism: the Iowa Women's Health Study. *Am J Public Health* 2010;100:1506-13.
23 Samama MM, Dahl OE, Quinlan DJ, Mismetti P, Rosencher N. Quantification of risk factors for venous thromboembolism: a preliminary study for the development of a risk assessment tool. *Haematologica* 2003;88:1410-21.
24 Alikhan R, Cohen AT, Combe S, Samama MM, Desjardins L, Eldor A, et al. Risk factors for venous thromboembolism in hospitalized patients with acute medical illness: analysis of the MEDENOX Study. *Arch Intern Med* 2004;164:963-8.
25 Ng TMH, Tsai F, Khatri N, Barakat MN, Elkayam U. Venous thromboembolism in hospitalized patients with heart failure: incidence, prognosis, and prevention. *Circ Heart Fail* 2010;3:165-73.
26 Matthew JG, Joe W, Timothy RC. Venous thromboembolism during active disease and remission in inflammatory bowel disease: a cohort study. *Lancet* 2010;375:657-63.
27 Miehsler W, Reinisch W, Valic E, Osterode W, Tillinger W, Feichtenschlager T, et al. Is inflammatory bowel disease an independent and disease specific risk factor for thromboembolism? *Gut* 2004;53:542-8.
28 Lee AYY, Levine MN. Venous thromboembolism and cancer: risks and outcomes. *Circulation* 2003;107:17-21.

What is already known on this topic

Venous thromboembolism is a major cause of preventable morbidity and mortality

Independent risk factors have been identified and UK guidelines encourage identification of high risk patients and effective use of preventative measures

There are currently no validated risk prediction algorithms to identify high risk individuals which are suitable for use in primary care

What this study adds

We have developed and validated a new risk prediction algorithm which identifies patients at high risk of venous thromboembolism with good discrimination and calibration

The algorithm uses simple clinical variables which patients are likely to know or which are routinely recorded in general practice records

The algorithm could be integrated into general practice systems to risk assess patients on medication. It could also be used before hospital admission, long haul flights, or the initiation of medication which might increase the risk of venous thromboembolism.

29 Couturaud F, Leroyer C, Julian JA, Kahn SR, Ginsberg JS, Wells PS, et al. Factors that predict risk of thrombosis in relatives of patients with unprovoked venous thromboembolism. *Chest* 2009;136:1537-45.
30 Hippisley-Cox J, Coupland C. Predicting the risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of the QFractureScore. *BMJ* 2009;339:b4229.
31 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007;335:136.
32 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart* 2008;94:34-9.
33 Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ* 2009;339:b2584.
34 Schafer J, Graham J. Missing data: our view of the state of the art. *Psychological Methods* 2002;7:147-77.
35 Group TAM. Academic medicine: problems and solutions. *BMJ* 1989;298:573-9.
36 Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. *J Epidemiol Community Health* 2007;60:979.
37 Moons KGM, Donders RART, Stijnen T, Harrell FJ. Using the outcome for imputation of missing predictor values was preferred. *J Epidemiol Community Health* 2006;59:1092.
38 Rubin DB. *Multiple imputation for non-response in surveys*. John Wiley, 1987.
39 Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* 1999;28:964-74.
40 Hosmer D, Lemeshow S. *Applied logistic regressiom* . John Wiley & Sons Inc, 1989.
41 Royston P. Explained variation for survival models. *Stata J* 2006;6:1-14.
42 Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;23:723-48.
43 Heinemann L, DoMinh T, Assmann A, Schramm W, Schurmann R, Hilpert J, et al. VTE risk assessment—a prognostic model: BATER Cohort Study of young women. *Thrombosis J* 2005;3:5.
44 Khorana AA, Kuderer NM, Culakova E, Lyman GH, Francis CW. Development and validation of a predictive model for chemotherapy-associated thrombosis. *Blood* 2008;111:4902-7.
45 Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. *BMJ* 1991;302:766-8.
46 Majeed A. Sources, uses, strengths and limitations of data collected in primary care in England. *Health Stat* 2004:5-14.
47 Silverstein MD, Heit JA, Mohr DN, Petterson TM, O'Fallon WM, Melton LJ 3rd. Trends in the incidence of deep vein thrombosis and pulmonary embolism: a 25-year population-based study. *Arch Intern Med* 1998;158:585-93.
48 Heinemann LA, Dinger JC. Range of published estimates of venous thromboembolism incidence in young women. *Contraception* 2007;75:328-36.
49 Koeleman P, Reitsma P, Allaart C, Bertina R. Activated protein C resistance as an additional risk factor for thrombosis in protein C-deficient families. *Blood* 1994;84:1031-5.
50 Furie B, Furie BC. Mechanisms of thrombus formation. *N Engl J Med* 2008;359:938-49.

Accepted: 16 June 2011

Cite this as: *BMJ* 2011;343:d4656

Tables

Table 1| Baseline characteristics of patients in derivation and validation cohorts

	Derivation cohort (n=2 314 701)	Validation cohort (n=1 240 602)
Male	1 190 500 (51.4)	637 482 (51.4)
Female	1 124 201 (48.6)	603 120 (48.6)
Mean age (SD)	47.6 (15.7)	47.7 (15.7)
Mean Townsend score (SD)	−0.1 (3.5)	−0.1 (3.5)
BMI recorded	1 828 253 (79.0)	975 051 (78.6)
Mean BMI (SD)	26.4 (4.8)	26.4 (4.8)
Smoking status recorded	2 168 137 (93.7)	1 160 241 (93.5)
BMI and smoking status recorded	1 818 859 (78.6)	970 083 (78.2)
Smoking status		
Non-smoker	1 187 610 (51.3)	634 966 (51.2)
Ex-smoker	430 425 (18.6)	231 139 (18.6)
Smoking amount not recorded	64 622 (2.8)	42 591 (3.4)
Light smoker (<10 cigarettes/day)	168 021 (7.3)	82 943 (6.7)
Moderate smoker (10-19 cigarettes/day)	189 805 (8.2)	98 492 (7.9)
Heavy smoker (≥20 cigarettes/day)	127 654 (5.5)	70 110 (5.7)
Medical and family history		
Family history of venous thromboembolism	67 (0.0)	27 (0.0)
Varicose veins	41 054 (1.8)	21 737 (1.8)
Congestive cardiac failure	15 081 (0.7)	7910 (0.6)
Rheumatoid arthritis	16 601 (0.7)	8918 (0.7)
Chronic renal disease	5957 (0.3)	3275 (0.3)
Inflammatory bowel disease	14 910 (0.6)	7761 (0.6)
Lung cancer	1526 (0.1)	723 (0.1)
Gastrointestinal cancer	7633 (0.3)	3981 (0.3)
Pancreatic cancer	197 (0.0)	104 (0.0)
Renal cancer	3209 (0.1)	1685 (0.1)
Breast cancer (women only)	14 004 (0.6)	7399 (0.6)
Prostate cancer (men only)	6326 (0.3)	3326 (0.3)
Other cancers	18 316 (0.8)	9582 (0.8)
Current typical antipsychotic drugs	51 343 (2.2)	28 471 (2.3)
Current atypical antipsychotic drugs	13 031 (0.6)	6841 (0.6)
Current oral contraceptives (women only)	98 690 (4.3)	53 654 (4.3)
Current hormone replacement therapy (women only)	52 218 (2.3)	28 291 (2.3)
Current tamoxifen (women only)	7116 (0.3)	3849 (0.3)
Hip fracture or replacement in past 182 days	2777 (0.1)	1318 (0.1)
Hospital admission in past 182 days	27 657 (1.2)	16 106 (1.3)

BMI=body mass index. Figures in the tables are number (%) unless otherwise specified.

Table 2| **Adjusted hazard ratios (95% CI) for final models in derivation cohort**

	Women			Men		
	Events (No)	Adjusted hazard ratio* (95% CI)	P value	Events (No)	Adjusted hazard ratio* (95% CI)	P value
Smoking status						
Non-smoker	4533	1.00	–	3148	1.00	–
Ex-smoker	1689	1.07 (1.01 to 1.15)	0.030	2238	1.06 (0.995 to 1.13)	0.070
Light smoker	443	1.22 (1.09 to 1.37)	0.001	618	1.22 (1.09 to 1.35)	<0.001
Moderate smoker	558	1.17 (1.05 to 1.29)	0.003	592	1.37 (1.22 to 1.52)	<0.001
Heavy smoker	375	1.34 (1.18 to 1.52)	<0.001	562	1.49 (1.33 to 1.66)	<0.001
Medical history						
Varicose veins	407	1.40 (1.24 to 1.58)	<0.001	172	1.38 (1.18 to 1.63)	<0.001
Congestive cardiac failure	206	1.40 (1.2 to 1.62)	<0.001	168	1.33 (1.13 to 1.57)	0.001
Chronic renal disease	46	1.60 (1.17 to 2.19)	0.003	62	1.92 (1.50 to 2.44)	<0.001
Any cancer	573	1.85 (1.69 to 2.03)	<0.001	505	2.18 (1.97 to 2.41)	<0.001
Chronic obstructive airways disease	360	1.41 (1.24 to 1.62)	<0.001	429	1.62 (1.45 to 1.80)	<0.001
Inflammatory bowel disease	87	1.45 (1.15 to 1.82)	0.002	94	1.5 (1.18 to 1.91)	0.001
Hospital admission in past six months	244	1.86 (1.63 to 2.14)	<0.001	209	1.93 (1.64 to 2.27)	<0.001
Current medication						
Antipsychotic drugs†	187	1.55 (1.32 to 1.81)	<0.001	121	1.84 (1.51 to 2.23)	<0.001
Tamoxifen†	97	1.48 (1.19 to 1.84)	<0.001	NA	NA	NA
Oral contraceptives†	229	1.33 (1.12 to 1.58)	0.001	NA	NA	NA
Hormone replacement therapy†	447	1.20 (1.08 to 1.34)	0.001	NA	NA	NA

95% CI=95% confidence intervals. NA=not applicable. *Hazard ratios adjusted for all other terms in the table, age, and body mass index (BMI). Models included fractional polynomial terms for age and BMI. For women, terms were $age^{-0.5}$, ln(age), BMI^{-2}, BMI^{-2}ln(BMI). For men, terms were age^{3}, age^{3}ln(age), BMI^{-2}, BMI^{-2}ln(BMI). †Compared with patients without this characteristic.

Table 3| Validation statistics for risk prediction algorithm in validation cohort

	Mean (95% CI) evaluated at one year	Mean (95% CI) evaluated at five years
Women		
R^2 statistic (%)*	28.02 (24.60 to 31.44)	32.78 (31.08 to 34.48)
D statistic†	1.28 (1.17 to 1.39)	1.43 (1.37 to 1.49)
ROC statistic†	0.71 (0.70 to 0.73)	0.75 (0.74 to 0.76)
Men		
R^2 statistic (%)	31.11 (27.57 to 34.64)	33.51 (31.71 to 35.30)
D statistic	1.38 (1.26 to 1.49)	1.45 (1.39 to 1.51)
ROC statistic	0.72 (0.70 to 0.74)	0.75 (0.74 to 0.76)

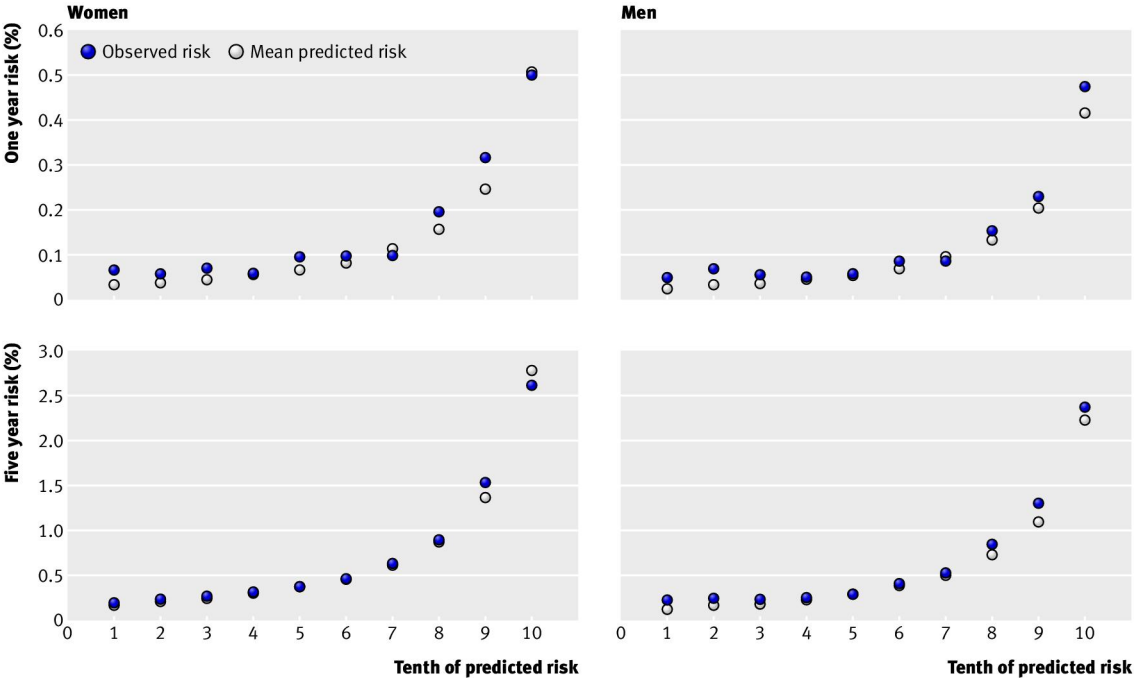
*Statistic shows explained variation (higher values indicate that more variation is explained). †ROC=receiver operating curve. ROC statistic is a measure of discrimination (higher values indicate better discrimination).

Table 4| **Incident cases of venous thromboembolism in groups of predicted risk within one and five years, in validation cohort (men and women combined)**

	Risk threshold per 1000	No of patients in risk group	No of patients in risk group with new VTE diagnosis	Total no of new VTE diagnoses	Sensitivity (%)	Positive predictive value (%)
Five year risk						
Top 10% risk score	15	124 060	2441	6913	35.3	2.0
Top 5% risk score	21	62 027	1400	6913	20.3	2.3
Top 1% risk score	38	12 405	350	6913	5.1	2.8
Top 0.5% risk score	48	6203	191	6913	2.8	3.1
One year risk						
Top 10% risk score	3	124 042	595	1686	35.3	0.5
Top 5% risk score	4	62 018	356	1686	21.1	0.6
Top 1% risk score	7	12 406	101	1686	6.0	0.8
Top 0.5% risk score	9	6202	57	1686	3.4	0.9

VTE=venous thromboembolism.

Figure



Mean predicted risks and observed risks of venous thromboembolism by tenth of predicted risk, applying risk prediction scores to the validation cohort