Diagnostic prediction models for suspected pulmonary embolism: systematic review and independent external validation in primary care

Janneke M T Hendriksen,[1] Geert-Jan Geersing,[1] Wim A M Lucassen,[2] Petra M G Erkens,[3] Henri E J H Stoffers,[3] Henk C P M van Weert,[2] Harry R Büller,[4] Arno W Hoes,[1] Karel G M Moons[1]

[1]Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, 3508 GA Utrecht, Netherlands

[2]Department of General Practice, Academic Medical Center, Amsterdam, Netherlands

[3]Department of Family Medicine, CAHPRI School for Public Health and Primary Care, Maastricht University, Maastricht, Netherlands

[4]Department of Vascular Medicine, Academic Medical Center, Amsterdam, Netherlands

Correspondence to:
J M T Hendriksen
j.m.t.hendriksen-9@umcutrecht.nl

ABSTRACT

OBJECTIVE
To validate all diagnostic prediction models for ruling out pulmonary embolism that are easily applicable in primary care.

DESIGN
Systematic review followed by independent external validation study to assess transportability of retrieved models to primary care medicine.

SETTING
300 general practices in the Netherlands.

PARTICIPANTS
Individual patient dataset of 598 patients with suspected acute pulmonary embolism in primary care.

MAIN OUTCOME MEASURES
Discriminative ability of all models retrieved by systematic literature search, assessed by calculation and comparison of C statistics. After stratification into groups with high and low probability of pulmonary embolism according to pre-specified model cut-offs combined with qualitative D-dimer test, sensitivity, specificity, efficiency (overall proportion of patients with low probability of pulmonary embolism), and failure rate (proportion of pulmonary embolism cases in group of patients with low probability) were calculated for all models.

RESULTS
Ten published prediction models for the diagnosis of pulmonary embolism were found. Five of these models could be validated in the primary care dataset: the original Wells, modified Wells, simplified Wells, revised Geneva, and simplified revised Geneva models. Discriminative ability was comparable for all models (range of C statistic 0.75-0.80). Sensitivity ranged from 88% (simplified revised Geneva) to 96% (simplified Wells) and specificity from 48% (revised Geneva) to 53% (simplified revised Geneva). Efficiency of all models was between 43% and 48%. Differences were observed between failure rates, especially between the simplified Wells and the simplified revised Geneva models (failure rates 1.2% (95% confidence interval 0.2% to 3.3%) and 3.1% (1.4% to 5.9%), respectively; absolute difference −1.98% (−3.33% to −0.74%)). Irrespective of the diagnostic prediction model used, three patients were incorrectly classified as having low probability of pulmonary embolism; pulmonary embolism was diagnosed only after referral to secondary care.

CONCLUSIONS
Five diagnostic pulmonary embolism prediction models that are easily applicable in primary care were validated in this setting. Whereas efficiency was comparable for all rules, the Wells rules gave the best performance in terms of lower failure rates.

WHAT IS ALREADY KNOWN ON THIS TOPIC

A low Wells score (≤4) in combination with a negative qualitative point of care D-dimer test safely excludes pulmonary embolism in about 4/10 patients in primary care

More diagnostic prediction models for pulmonary embolism have been developed and validated in secondary care, but clinical performance of these other models in primary care is unknown

WHAT THIS STUDY ADDS

The (simplified) versions of the Wells rule and the revised Geneva scores exclude 4/10 patients on the basis of a low predicted risk of pulmonary embolism and a negative point of care D-dimer test

Use of the original and simplified revised Geneva score, in combination with a point of care D-dimer test, was associated with a higher proportion of pulmonary embolism patients missed, compared with use of the Wells rules in primary care

Introduction

Pulmonary embolism is a potentially fatal condition if left untreated. Its presentation can be relatively mild, sometimes even mimicking myalgia or a simple cough. This causes pulmonary embolism to be a diagnosis that is easily missed.[1 2] As a result, physicians have a low threshold for suspicion and subsequent referral for further diagnostics.[3 4] Referred patients will be exposed to the burden, costs, and even potential iatrogenic damage of diagnostic techniques such as spiral computed tomography or contrast nephropathy.[5] However, only in a small subset (about 10-15%) of all suspected cases are emboli actually confirmed during diagnostic investigation.[6]

Several non-invasive diagnostic prediction models have been developed for safe exclusion of pulmonary embolism and are usually followed by D-dimer testing.[7] Physicians can use these models as a strategy to enhance the efficiency of the diagnostic process by precluding those patients with a low probability of pulmonary embolism from further diagnostic tests, without compromising on safety (that is, missing cases of pulmonary embolism). Such diagnostic strategies can reduce the number of unnecessary computed tomography scans by 35%, with only 1-2% of missed cases in the group of patients with a low probability of pulmonary embolism.[7]

In many countries, general practitioners are the first physicians to encounter patients with symptoms suggestive of pulmonary embolism. Risk stratification

is valuable in deciding which patients to refer. All diagnostic models for safe exclusion of pulmonary embolism have been developed and validated in hospital or acute care settings. However, diagnostic prediction models developed in a particular setting often perform less well when applied in another setting. Therefore, models derived in hospital or acute care settings cannot simply be implemented in primary care.[8-14] Reasons for this poorer performance include differences in the case mix and the prevalence of pulmonary embolism due to the unselected population, as well as differences in physicians' experience of patients with suspected pulmonary embolism.[9 10 15 16] Hence, when transferring diagnostic models or strategies across healthcare settings, evaluation of their performance in this other setting is necessary first. This form of external validation is referred to as domain or setting validation,[8 10 17] or as quantification of the transportability of prediction models.[13 18]

The recent AMUSE-2 study (Amsterdam, Maastricht, Utrecht Study on thrombo-Embolism)[19] has been the first to prospectively quantify the transportability of the, perhaps best known, secondary care derived diagnostic prediction model for pulmonary embolism (that is, the Wells pulmonary embolism rule,[20] combined with point of care D-dimer testing) in a primary care setting. Various other diagnostic pulmonary embolism prediction models that may also be valuable for primary care have been developed but have not been validated in a primary care population.

The aim of this study was therefore to assess the clinical performance in a primary care setting of all existing diagnostic models developed for patients with suspected pulmonary embolism. We firstly did a systematic review and critical appraisal of all available diagnostic models for pulmonary embolism, as recommended by guidelines on prediction models research.[21] Next, the diagnostic models easily applicable in primary care were validated in the AMUSE-2 dataset—that is, a large independent prospectively constructed cohort of patients presenting to their general practitioner with complaints suggestive of pulmonary embolism.

Methods
Updated systematic review
For our systematic review and critical appraisal of the existing diagnostic models for pulmonary embolism, we followed the recent methodological guidance by the Prognosis Methods Group of the Cochrane Collaboration.[21-24]

Firstly, we framed the review question and design by using the CHARMS checklist for systematic reviews of prediction models (see appendix box A).[21] We then repeated the systematic search previously performed for an aggregate meta-analysis by Lucassen et al and used the same study selection criteria.[7] We searched for studies on development and validation of diagnostic prediction models published between January 2010 and October 2014. Details on the search syntax can be found in appendix figure A. We then critically appraised studies on the development of diagnostic prediction

models by using the CHARMS checklist (appendix table A). All retrieved papers were examined by two independent reviewers (JH, GJG) and a third independent reviewer (KGMM) in case of disagreement.

Given the scope of our systematic review (see appendix box A), we assessed all diagnostic prediction models for pulmonary embolism, retrieved by our search, on their applicability in a primary care domain. Accordingly, the diagnostic predictors or tests included in the diagnostic model needed to be measurable at the general practitioner's office. Variables such as signs and symptoms, items from history taking, pulse rate, or blood pressure are easily and quickly obtained in primary care, whereas results from (arterial) blood gas analyses, chest radiographs, or advanced electrocardiograph interpretations generally are not. Diagnostic models with predictors that cannot easily be obtained in primary care were excluded from the main analyses.

Validation cohort
Population characteristics
The AMUSE-2 cohort was designed to prospectively validate the Wells pulmonary embolism rule in a Dutch primary care setting. The study took place between 1 July 2007 and 31 December 2010. In short, it included 662 adult patients presenting at one of the participating general practices with complaints raising suspicion of pulmonary embolism (that is, acute dyspnoea, pain on inspiration, or unexplained cough; all at the discretion of the including physicians). Of these patients, 64 met one of the predefined exclusion criteria of anticoagulant treatment at presentation, pregnancy, or unwillingness or inability to provide written informed consent, leaving 598 patients for further evaluation. More details of this cohort and the sample size calculation are described elsewhere.[19]

Predictors
In all participants, the general practitioner assessed relevant information on general health and specific cardiopulmonary and signs and symptoms of deep venous thrombosis by systematically filling out a pre-specified case record form. Subsequently, a qualitative point of care D-dimer test (Simplify D-dimer; Clearview, Inverness Medical, Bedford, UK) was performed. This test returns a visual dichotomous outcome; a positive test result is indicated by a pink-purple coloured line that appears on the disposable device within 10 minutes of application. This corresponds to a D-dimer concentration above 80 ng/mL. Only a control line will be visible if the test is negative. In case of an inconclusive test result, we classified the result as positive.

All predictors in the validation cohort were assessed blinded for the outcome. Exact definitions and measurement methods of the predictors in the validation cohort have been described previously.[19]

Outcome
The study protocol recommended referral of all patients with suspected pulmonary embolism to secondary care, regardless of the outcome of the Wells

rule or D-dimer test. In secondary care, the regular diagnostic pathway according to local hospital guidelines was followed, with no explicit blinding for the general practitioner's findings. This usually comprised a combination of estimated probability and quantitative laboratory based D-dimer testing and was followed by diagnostic imaging if indicated. The primary outcome was the presence of venous thromboembolism (either deep venous thrombosis or pulmonary embolism), as based on a composite reference standard of all diagnostic imaging tests performed in the hospital (spiral computed tomography, ventilation-perfusion scanning, pulmonary angiography, leg ultrasonography, and clinical probability assessment as performed in secondary care, with or without D-dimer testing) and including any occurrence of venous thromboembolic events during three months of follow-up in primary care.

Data analysis
For all diagnostic prediction models, we retrospectively calculated the individual score of each included patient on the basis of the presence or absence of the model's predictors. We compared the overall discriminative ability of the models, using the C statistic (that is, the area under the curve) of the receiver operating characteristics curve, with 95% confidence intervals. We assessed differences between the C statistics with the DeLong method.[25]

To stratify all participants in the validation cohort into categories of low or high probability of having pulmonary embolism, we used the diagnostic pathway as recommended by guidelines (fig 1): first the stratification based on the cut-off value of each diagnostic model as suggested in the development papers for the model, followed by a D-dimer test in case of a low predicted probability of pulmonary embolism. A subsequent negative D-dimer test implied a low predicted probability of pulmonary embolism and no need for referral to secondary care. In all other cases, the high predicted probability of pulmonary embolism meant referral for further objective testing.
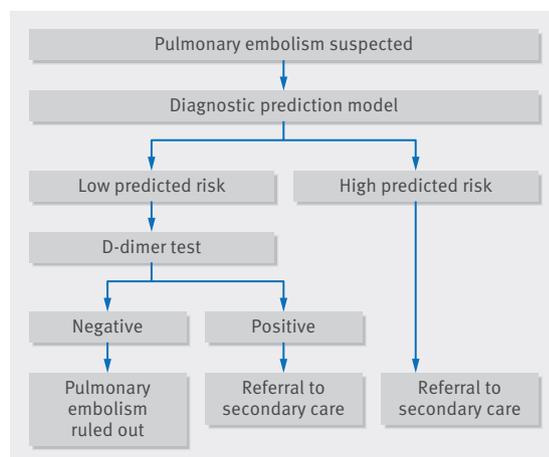
Fig 1 | Flow scheme of diagnostic pathway in suspected pulmonary embolism in primary care

For the Wells rule, two cut-off values have been proposed: low risk (≤4) and very low risk (<2). Given the prevalence in this primary care cohort and following previous publications on this topic, we chose to use the first of these cut-off values for all subsequent analyses.[19 26] We then calculated the common diagnostic accuracy measures of the models (that is, sensitivity, specificity, and positive and negative predictive values).

From a clinical point of view, a diagnostic prediction model should ideally classify as many patients as possible in the non-referral group, but not at the expense of an increase in pulmonary embolism events missed in this group. Therefore, we evaluated the clinical performance of each model (combined with D-dimer) by focusing on the efficiency and the failure rate. We defined efficiency as the proportion of patients in the whole cohort stratified to the group with low predicted probability of pulmonary embolism (that is: (true negatives (tn)+false negatives (fn))/total cohort). We defined failure rate as the proportion of these patients with low predicted probability of pulmonary embolism ultimately diagnosed as having pulmonary embolism on the basis of our composite reference standard (that is: fn/(tn+fn)). We calculated differences in failure rates between the models, with the surrounding 95% confidence intervals. We then varied the cut-off values as proposed in the model development studies to evaluate the influence of different cut-off values on the outcome measures. Failure rates were displayed in a forest plot, together with the failure rates found in previous validation studies of the different models with D-dimer testing.

Finally, we constructed calibration plots for the diagnostic prediction models. With calibration plots, the agreement between the predicted and observed probability of pulmonary embolism can be visualised. In the absence of a reported intercept for the models, we re-estimated the intercept in the validation cohort by using the linear predictor as offset in a logistic regression model including the model coefficients.

In the dataset, we imputed missing values for predictors by using multiple imputation techniques before our analyses.[27 28] Imputation was performed to minimise the effect of the bias associated with selectively ignoring these patients. In case of a high percentage of missing values, no imputation was performed. Instead, the main analysis was carried out with all missing values assigned as the predictor being absent. In a sensitivity analysis, all missing values were assigned as the predictor being present.

We used IBM SPSS version 21 for descriptive statistical analyses. We used R version 3.2.0 for forest plots and the calculation of differences in C statistics (DeLong method) and failure rates.

Reporting
The results of this validation study were reported in adherence to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) guideline (see appendix table C).[29 30]

Patient involvement

No patients were involved in setting the research question or the outcome measures; nor were they involved in recruitment or in the design and implementation of the study. There are no plans to involve patients in dissemination.

Results

Our systematic literature search identified four model validation studies but no newly developed models in addition to the 10 models previously found with the search by Lucassen (fig 2). Of these 10 models, we excluded two models (the Pisa rule[31] and the original Geneva score[32]) from further analysis on the basis of the predetermined criteria for validation in our primary care cohort (fig 2). The original, modified, and simplified Wells rules, the revised Geneva scores, the revised Pisa rule, the pulmonary embolism rule-out criteria (PERC) rule, and the Charlotte rule fulfilled the predetermined primary care criteria.[20 33-38] However, owing to complete missing information on the predictors oxygen saturation and body temperature in our dataset, we were not able to validate the revised Pisa rule, the PERC rule, and the Charlotte rule, leaving five models for further evaluation. See appendix table B for an overview of all diagnostic models identified in this study, including the five models that were tested in the individual patient dataset of our validation cohort: the original, modified, and simplified Wells rules and the original and simplified revised Geneva scores.

Of the four recently published validation studies identified by our search, only one study reported the results of the combination of the prediction model and D-dimer testing and was used in this current update.[19] The observed failure rates from previous development and validation studies, and this validation of the prediction models, are shown in a forest plot (fig 3).

External validation

In the validation cohort, venous thromboembolism was diagnosed in 73 of 598 patients (72 pulmonary embolism, 1 deep venous thrombosis) during the complete three month follow-up period (prevalence 12%). Table 1 shows the baseline characteristics of the cohort, as well as the baseline characteristics observed in the development studies of the validated diagnostic prediction models. The main differences between the four development cohorts and the current validation cohort include the prevalence of the outcome, mean age, and percentage of male participants.

All models had moderate to good discriminative ability, with a C statistic ranging from 0.75 (simplified revised Geneva score) to 0.80 (original and modified Wells rules) (largest difference $P=0.038$; simplified revised Geneva v original Wells) (table 2 and appendix figure B).

For the originally suggested thresholds of the three Wells rules (original Wells, modified Wells, and simplified Wells, all with D-dimer testing if low predicted probability of pulmonary embolism), sensitivity was around 95%; it was slightly lower for the Geneva score models (88-90%). All five diagnostic models showed a specificity of approximately 50% (table 3). The simplified revised Geneva model was observed to be the most efficient rule (48%; 287 non-referred patients in cohort of 598 patients with suspected pulmonary embolism), but it was also associated with the highest failure rate (nine missed events in 287 non-referred patients; 3.1%). The largest difference in failure rates was observed between the simplified Wells and simplified revised Geneva score (−1.98%, 95% confidence interval −3.33% to −0.74%) (table 4). Interestingly, three of these nine patients were missed by all rules under study. These three patients are described in more detail in the box.

Overall, a one point lower cut-off (low risk original Wells ≤3 and D-dimer negative) affected the failure rate of each model little (failure rate 0.9-2.9%) but hampered its efficiency, especially for the simplified Wells rule (table 5). Conversely, an increased cut-off was more efficient, ranging from 49% to 54%, but more pulmonary embolism events were missed.
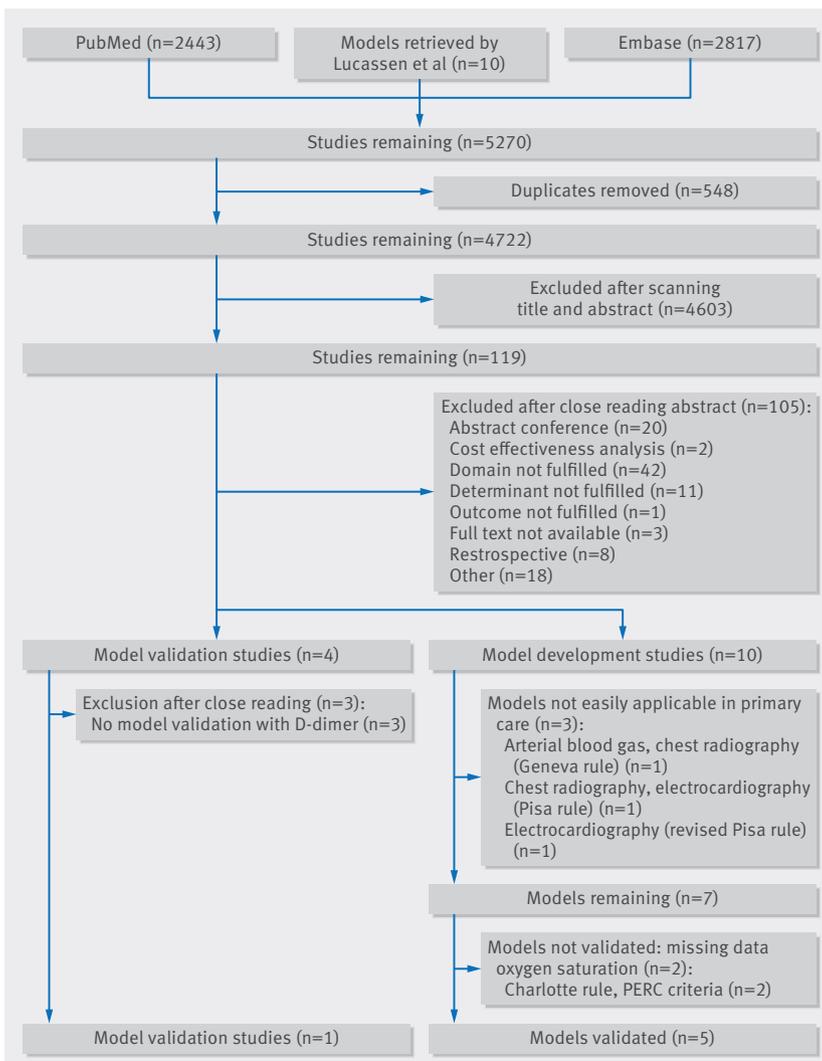

Fig 2 | Overview of selection of studies that developed or validated prediction models for diagnosis of pulmonary embolism, based on literature search in PubMed and Embase. PERC=pulmonary embolism rule-out criteria
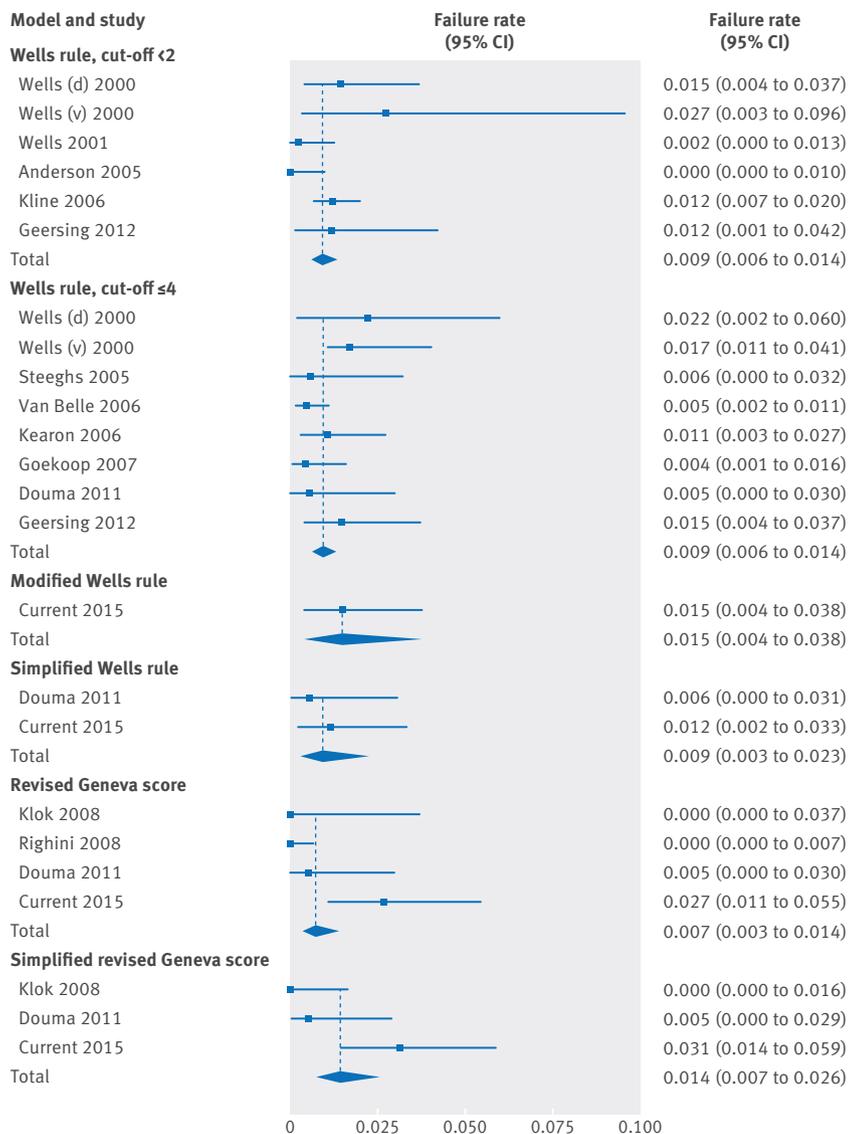
Fig 3 | Forest plot of failure rates in development and validation studies of diagnostic prediction models, if combined with D-dimer testing

As can be appreciated from appendix figure C, the calibrations of the Wells rules and the revised Geneva score were good and comparable. Calibration of the simplified revised Geneva score was slightly worse, especially in the highest tenth of predicted risk.

Discussion

Our systematic review identified 10 previously developed diagnostic prediction models to rule out pulmonary embolism. Of these, we evaluated five models for their transportability to primary care in an independent cohort of 598 patients. We found that all five models could rule out pulmonary embolism in about four in every 10 patients. However, we observed a difference in favour of the three Wells rules compared with the revised Geneva models in terms of safety. The proportion of cases of pulmonary embolism in those patients identified as being at low risk was substantially lower for the Wells rules.

Strengths and limitations

Strengths of this study are that—to the best of our knowledge—this is the first study to validate multiple diagnostic strategies for ruling out pulmonary embolism in the primary care setting. Furthermore, the validated models were selected on the basis of an extensive literature search and critical appraisal of the models. For this, we used the guidance of the CHARMS checklist of the Cochrane Collaboration.[21] Moreover, we focused on the clinical performance measures of efficiency and failure rate; we believe that our results are relevant to daily practice, as these measures reflect the effect on clinical practice of using a diagnostic prediction model. Nevertheless, for full appreciation of our results, some limitations need to be considered.

Firstly, our study was conducted in the Netherlands, a country with a well developed primary care structure in which general practitioners are the healthcare gatekeepers. Although in many west European countries, Canada, Australia, New Zealand, and parts of the United States, general practitioners play a rather comparable role in the healthcare system, our results may be less generalisable to healthcare settings where primary care medicine is less well developed.

Secondly, our case record form was primarily designed to score the items of the original Wells pulmonary embolism rule with the aim of validating this diagnostic model in a primary care setting. Accordingly, participating general practitioners were not asked to score all items of other diagnostic prediction models at the moment of inclusion, although results for the scores of the other models could be calculated post hoc. However, in almost 60% of patients in our cohort data were lacking on pain on deep vein palpation, which is part of the Geneva models. In the main analysis, we assigned all missing values to "no pain." In a sensitivity analysis, we repeated all analyses with all missing values considered as "pain present." Results did not alter, indicating a limited influence of the frequently missing item in the whole prediction rule.

Thirdly, five more prediction models have been developed but were not validated in this study. Two models did not meet our a priori defined primary care criteria. As they rely heavily on predictors not frequently assessed in primary care, such chest radiography results, the applicability of these models in primary care is limited. These predictors were not routinely collected in our validation cohort. Also, we did not have information on oxygen saturation and fever. Following guidelines on validation of prediction models, we wanted to refrain from simplification of the models by excluding some predictors and then refitting the new model on our data, as that would lead to the development of even more models. Simply omitting these missing predictors from the full models without further refitting would lead to structural underestimation of risks and poor clinical performance. Therefore, we refrained from validation of these five diagnostic prediction models in our primary care cohort.

Fourthly, debate is ongoing about the accepted failure rates in clinical practice. Although no true

      5

Table 1 | Baseline characteristics of participants in external validation cohort (AMUSE-2; patients presenting in primary care, with suspected pulmonary embolism), and baseline characteristics of development studies of five diagnostic prediction models. Values are numbers (percentages) unless stated otherwise

Characteristics	AMUSE 2 validation cohort (n=598)	Original Wells[20] (n=1260)	Modified/simplified Wells[33] (n=3306)	Original revised Geneva[34] (n=965)	Simplified revised Geneva[35] (n=1049)
Male sex	173 (28.9)	NR	1409 (42.6)	403 (41.8)	414 (39.5)
Mean (SD) age, years	48 (16)	NR	53 (18)	61 (19)	56 (SD NR)
Venous thromboembolic events	73 (12.2)	NA	NA	NA	241 (23.0)
Pulmonary embolism cases	72 (12.0)	222 (17.6)	674 (20.4)	222 (23.0)	241 (23.0)
Items Wells rule					
Clinical signs and symptoms of DVT	57 (9.5)	NR	190 (5.7)	NR	NR
Alternative diagnosis less likely	333 (55.7)	NR	2032 (61.5)	NR	NR
Heart rate >100 bpm	111 (18.6)	NR	867 (26.2)	NR	NR
Immobilisation or recent surgery	94 (15.7)	NR	610 (18.5)	232 (24.0)	NR
Previous pulmonary embolism or DVT	84 (14.0)	160 (12.7)	480 (14.5)	166 (17.2)	NR
Haemoptysis	21 (3.5)	NR	176 (5.3)	43 (4.5)	NR
Malignancy	26 (4.3)	NR	476 (14.4)	89 (9.2)	NR
Items revised Geneva score					
Age >65 years	103 (17.2)	NR	NR	NR	NR
Heart rate:					
<75 bpm	187 (31.3)	NR	NR	NR	NR
75-94 bpm	259 (43.3)	NR	NR	NR	NR
>94 bpm	152 (25.4)	NR	NR	NR	NR
Lower limb pain	96 (16.1)	NR	NR	138 (14.3)	NR
Oedema/swelling	48 (8.0)	NR	NR	NR	NR
Pain and swelling*†	22 (3.7)	NR	NR	51 (5.3)	NR
Pain on lower limb deep venous palpation†	34 (5.7)	NR	NR	NR	NR
Pain and swelling*‡	24 (4.0)	NR	NR	NR	NR
Pain on lower limb deep venous palpation‡	391 (65.4)	NR	NR	NR	NR
D-dimer measurement					
Qualitative D-dimer positive	259 (43.3)	NR	NA§	NA	NA§

DVT=deep venous thrombosis; NA=not applicable; NR=not reported.
*"Pain and swelling" is a composite variable, formed by variable "Oedema/swelling" present and/or variable "pain on deep vein (DV) palpation" present.
†Missing values for "pain on DV palpation" (n=357 (59.7%)) were assigned as "pain on DV palpation" absent.
‡Missing values for "pain on DV palpation" (n=357 (59.7%)) were assigned as "pain on DV palpation" present.
§Quantitative D-dimer test instead of qualitative point of care D-dimer test performed.

consensus exists, some people consider an upper 95% confidence interval boundary of a failure rate higher than 3% to be too high, although previous landmark studies in the field of diagnosing pulmonary embolism found failure rates with an upper 95% confidence interval boundary close to 4%.[20] Others focused on the point estimate of the failure rate instead, which should be under 2% given that even the most invasive diagnostic procedure (digital subtraction angiography) cannot diagnose all pulmonary embolism events and misses

about 2% of cases.[39] We observed failure rates of the Wells rules of 1.2-1.5%, which is well below this 2% point estimate but with a 95% confidence interval that crosses 3%. Importantly, however, these boundaries for the 95% confidence interval are not necessarily the "true" values but reflect the remaining uncertainty around the point estimate.

Fifthly, for scoring the item "pulmonary embolism most likely," Wells originally suggested that information on electrocardiography, routine laboratory tests, and

Table 2 | Head to head comparison of C statistics of five diagnostic prediction models (without D-dimer testing) to rule out pulmonary embolism, validated in primary care AMUSE-2 cohort (n=598), using DeLong method[25]

Models compared (model 1 v model 2)	C statistic		Estimated difference (95% CI)	P value
	Model 1	Model 2		
Original Wells v simplified Wells	0.801	0.787	0.014 (−0.004 to 0.032)	0.132
Original Wells v modified Wells	0.801	0.798	0.003 (−0.001 to 0.007)	0.114
Original Wells v original revised Geneva	0.801	0.756	0.045 (−0.007 to 0.097)	0.091
Original Wells v simplified revised Geneva	0.801	0.748	0.053 (0.003 to 0.103)	0.038
Simplified Wells v modified Wells	0.787	0.798	−0.011 (−0.028 to 0.006)	0.203
Simplified Wells v original revised Geneva	0.787	0.756	0.031 (−0.017 to 0.079)	0.207
Simplified Wells v simplified revised Geneva	0.787	0.748	0.039 (−0.005 to 0.083)	0.084
Modified Wells v original revised Geneva	0.798	0.756	0.043 (−0.010 to 0.096)	0.113
Modified Wells v simplified revised Geneva	0.798	0.748	0.050 (0.000 to 0.100)	0.048
Original revised Geneva v simplified revised Geneva	0.756	0.748	0.008 (−0.010 to 0.026)	0.388

Table 3 | Diagnostic accuracy measures of five diagnostic prediction models, combined with point of care D-dimer testing, to rule out pulmonary embolism, validated in primary care AMUSE-2 cohort (n=598), with (95% confidence intervals)

Measure	Original Wells ≤4	Modified Wells ≤2	Simplified Wells ≤1	Original revised Geneva ≤5*	Simplified revised Geneva ≤2*
Sensitivity	95% (87% to 98%)	95% (87% to 98%)	96% (88% to 99%)	90% (81% to 96%)	88% (78% to 94%)
Specificity	51% (47% to 55%)	50% (46% to 55%)	49% (45% to 53%)	48% (44% to 53%)	53% (49% to 57%)
Positive predictive value	21% (17% to 26%)	21% (17% to 26%)	21% (17% to 25%)	20% (15% to 24%)	21% (16% to 26%)
Negative predictive value	99% (96% to 100%)	99% (96% to 100%)	99% (97% to 100%)	97% (95% to 99%)	97% (94% to 99%)
Efficiency	46% (41% to 50%)	45% (41% to 49%)	43% (39% to 48%)	44% (40% to 48%)	48% (44% to 52%)
Failure rate	1.5% (0.4% to 3.7%)	1.5% (0.4% to 3.8%)	1.2% (0.2% to 3.3%)	2.7% (1.1% to 5.4%)	3.1% (1.4% to 5.9%)

*Main analysis used composite variable "pain and swelling," in which missing values for "pain on deep vein (DV) palpation" were assigned as "pain on DV palpation" absent; results for analysis in which all missing values were assigned as "pain on DV palpation" present are identical and are not presented separately.

Table 4 | Differences in failure rates between different diagnostic prediction models in AMUSE-2 cohort (with 95% confidence intervals based on 200 bootstrap samples)

Comparison (model 1 v model 2)	Difference between failure rates (model 1 v model 2)
Original Wells v simplified Wells	0.32% (−0.08% to 1.04%)
Original Wells v modified Wells	−0.02% (−0.04% to 0.00%)
Original Wells v original revised Geneva	−1.21% (−2.49% to 0.14%)
Original Wells v simplified revised Geneva	−1.67% (−3.21% to −0.27%)
Simplified Wells v modified Wells	−0.33% (−1.06% to 0.06%)
Simplified Wells v original revised Geneva	−1.53% (−2.80% to −0.36%)
Simplified Wells v simplified revised Geneva	−1.98% (−3.33% to −0.74%)
Modified Wells v original revised Geneva	−1.20% (−2.47% to 0.18%)
Modified Wells v simplified revised Geneva	−1.65% (−3.21% to −0.25%)
Original revised Geneva v simplified revised Geneva	−0.45% (−1.23% to 0.21%)

Characteristics of three cases missed by all five diagnostic prediction models combined with D-dimer testing when validated in AMUSE-2 cohort, consisting of 598 primary care patients with suspected pulmonary embolism

Case 1—75 year old man

Venous thromboembolism in history; heart rate 59 bpm

Simplify D-dimer negative

Scores of models:
- Original Wells—1.5 (previous PE)
- Modified Wells—1.0 (previous PE)
- Simplified Wells—1.0 (previous PE)
- Original revised Geneva—4.0 (age >65, previous PE)
- Simplified revised Geneva—2.0 (age >65, previous PE)

Case 2—30 year old woman

Use of oral contraceptives; heart rate 92 bpm

Simplify D-dimer negative

Scores of models:
- Original Wells—3.0 (PE most likely diagnosis)
- Modified Wells—2.0 (PE most likely diagnosis)
- Simplified Wells—1.0 (PE most likely diagnosis)
- Original revised Geneva—3.0 (heart rate 75-94 bpm)
- Simplified revised Geneva—1.0 (heart rate 75-94 bpm)

Case 3—25 year old woman

Use of oral contraceptives; heart rate 80 bpm

Simplify D-dimer negative

Scores of models:
- Original Wells—0.0
- Modified Wells—0.0
- Simplified Wells—0.0
- Original revised Geneva—3.0 (heart rate 75-94 bpm)
- Simplified revised Geneva—1.0 (heart rate 75-94 bpm)

PE=pulmonary embolism; bpm=beats per minute.

chest radiographs would be needed. These items are often not readily available in primary care, which may hamper the scoring of this subjective item in the Wells rule in primary care. Given that we observed good results for the Wells rule in primary care, it would be of interest to know how general practitioners actually interpret this subjective item. One hypothesis could be that this subjective item reflects the presence or absence of risk factors for venous thromboembolism, such as recent long haul flights or oestrogen use. However, when comparing the distribution of known risk factors for the subjective item scored or not, we observed no clear differences (data not shown). Although this remains speculative, we believe that this subjective item "pulmonary embolism most likely" may reflect contextual knowledge from general practitioners on how symptoms are usually presented by their patients. Qualitative studies suggested that a consideration of "pulmonary embolism most likely" might come into the mind if a given patient presents symptoms differently ("out of the ordinary") compared with previous consultations.[1] Such contextual knowledge is often stronger in primary care, given the longstanding relationship that general practitioners often have with their patients.

Sixthly, despite the fact that most patients with the eventual diagnosis of pulmonary embolism were identified as having a high probability of pulmonary embolism by the validated diagnostic prediction models, three patients were missed by all of the rules. As shown in the box, these patients' characteristics were diverse, but two were young women who were taking oral contraception.

Finally, we used a qualitative point of care D-dimer test in this study. These tests are known to have a relatively lower sensitivity than laboratory based quantitative tests.[40 41] Potentially, this resulted in a higher number of false negative results. However, specificity of point of care tests is higher, which favours the efficiency of the test. Moreover, the ease of performing a point of care test on the spot and having the results available within 15 minutes are convenient in a primary care setting. The trade-off between harms (increase in false negatives) and benefits (increased efficiency, quick point of care testing) should ideally be assessed in a cost effectiveness analysis.

Relation to development and validation studies

Multiple studies have evaluated the diagnostic performance of the five diagnostic prediction models under

Table 5 | Diagnostic accuracy measures (with 95% CIs) of five diagnostic prediction models* combined with point of care D-dimer testing, to rule out pulmonary embolism, validated in primary care AMUSE-2 cohort (n=598)

Measure	Original Wells ≤5	Original Wells ≤3	Modified Wells ≤3	Modified Wells ≤1	Simplified Wells ≤2	Simplified Wells =0	Original revised Geneva ≤6	Original revised Geneva ≤4	Simplified revised Geneva ≤3	Simplified revised Geneva ≤1
Sensitivity	89% (80% to 95%)	95% (87% to 98%)	89% (80% to 95%)	97% (91% to 99%)	85% (75% to 92%)	99% (93% to 100%)	89% (80% to 95%)	92% (83% to 97%)	86% (76% to 93%)	93% (85% to 98%)
Specificity	59% (54% to 63%)	50% (46% to 55%)	59% (54% to 63%)	32% (28% to 36%)	59% (55% to 64%)	21% (18% to 25%)	54% (44% to 53%)	38% (34% to 42%)	59% (55% to 64%)	36% (32% to 41%)
PPV	23% (18% to 28%)	21% (17% to 26%)	23% (18% to 28%)	17% (13% to 20%)	23% (18% to 28%)	15% (12% to 18%)	21% (17% to 26%)	13% (10% to 16%)	23% (18% to 28%)	17% (13% to 21%)
NPV	97% (95% to 99%)	99% (96% to 100%)	97% (95% to 99%)	99% (96% to 100%)	97% (94% to 98%)	99% (95% to 100%)	97% (95% to 99%)	97% (94% to 99%)	97% (94% to 99%)	97% (94% to 99%)
Efficiency	53% (49% to 57%)	45% (41% to 49%)	53% (49% to 57%)	28% (25% to 32%)	54% (50% to 58%)	19% (16% to 22%)	49% (45% to 53%)	34% (30% to 38%)	54% (50% to 58%)	33% (29% to 37%)
Failure rate	2.5% (1.1% to 4.9%)	1.5% (0.4% to 3.8%)	2.5% (1.1% to 4.9%)	1.2% (0.3% to 4.2%)	3.4% (1.7% to 6.0%)	0.9% (0.0% to 4.9%)	2.8% (1.2% to 5.5%)	2.9% (1.1% to 6.3%)	3.1% (1.5% to 5.6%)	2.6% (0.8% to 5.9%)

NPV=negative predictive value; PPV=positive predictive value.
*Chosen cut-off values for ruling out pulmonary embolism ("low PE probability") are 1 point higher or lower than cut-offs recommended for published models.

study in a secondary (or tertiary) care setting. The failure rates observed in our validation study in primary care are largely in line with the previous studies regarding the performance of the Wells rules in combination with D-dimer testing, as can be appreciated from figure 3: most secondary care studies yielded a failure rate of the diagnostic approach of combining the Wells rule with D-dimer testing around 1% (range 0.5-2.2%), whereas in our study we observed a failure rate close to 1.5% for the Wells rules combined with D-dimer testing.

Contrastingly, for the (simplified) revised Geneva scores, the failure rates observed in this primary care cohort are slightly higher than those in the previous studies (fig 3). Reasons for this are not entirely clear, but one can think of the differences in cohort characteristics that might have influenced the diagnostic performance. Importantly, the revised Geneva scores have been developed in a tertiary care setting, with a preselected and thus more severe disease presentation. Conversely, an unselected patient population and a lower prevalence of pulmonary embolism are typical features of a primary care setting, and such differences in case mix of patients can influence the diagnostic performance of the rule.

Clinical implications and conclusions
On the basis of our findings of a (slightly) higher failure rate for the Geneva rules, we propose that these rules are less suitable for use in this particular setting, specifically when compared with the Wells rules. We would suggest that general practitioners use the (simplified version of the) Wells rule, combined with a (point of care) D-dimer test. Pulmonary embolism can be excluded in about four in every 10 patients with suspected pulmonary embolism, with an acceptably low failure rate of below 2%.

1   Barais M, Morio N, Cuzon Breton A, et al. "I can't find anything wrong: it must be a pulmonary embolism": diagnosing suspected pulmonary embolism in primary care, a qualitative study. PLoS One 2014;9:e98112.
2   Schiff G, Hasan O, Kim S, et al. Diagnostic error in medicine: analysis of 583 physician-reported errors. Arch Intern Med 2009;169:1881-7.
3   Prasad V, Rho J, Cifu A. The diagnosis and treatment of pulmonary embolism: a metaphor for medicine in the evidence-based medicine era. Arch Intern Med 2012;172:955-8.
4   Wiener R, Schwartz L, Woloshin S. Time trends in pulmonary embolism in the United States: evidence of overdiagnosis. Arch Intern Med 2011;171:831-7.
5   Lencioni R, Fattori R, Morana G, Stacul F. Contrast-induced nephropathy in patients undergoing computed tomography (CONNECT)—a clinical problem in daily practice? A multicenter observational study. Acta Radiol 2010;51:741-50.
6   Le Gal G, Bounameaux H. Diagnosing pulmonary embolism: running after the decreasing prevalence of cases among suspected patients. J Thromb Haemost 2004;2:1244-6.
7   Lucassen W, Geersing GJ, Erkens PM, et al. Clinical decision rules for excluding pulmonary embolism: a meta-analysis. Ann Intern Med 2011;155:448-60.
8   Altman D, Vergouwe Y, Royston P, Moons K. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.
9   Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice BMJ 2009;338:b606.

10 Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;98:691-8.

11 Hendriksen JM, Geersing GJ, Moons KG, de Groot JA. Diagnostic and prognostic prediction models. *J Thromb Haemost* 2013;11:129-41.

12 Altman D, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19:453-73.

13 Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006;144:201-9.

14 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130:515-24.

15 Knottnerus JA. Between iatrotropic stimulus and interiatric referral: the domain of primary care research. *J Clin Epidemiol* 2002;55:1201-6.

16 Oudega R, Hoes AW, Moons KG. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. *Ann Intern Med* 2005;143:100-7.

17 Toll D, Janssen K, Vergouwe Y, Moons K. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol* 2008;61:1085-94.

18 Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015;68:279-89.

19 Geersing GJ, Erkens PM, Lucassen W, et al. Safe exclusion of pulmonary embolism using the Wells rule and qualitative D-dimer testing in primary care: prospective cohort study. *BMJ* 2012;345:e6564.

20 Wells PS, Andersen DR, Rodger M, et al. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer. *Thromb Haemost* 2000;83:416-20.

21 Moons KG, De Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11:e1001744.

22 Riley RD, Ridley G, Williams K, Altman DG, Hayden J, de Vet HC. Prognosis research: toward evidence-based results and a Cochrane methods group. *J Clin Epidemiol* 2007;60:863-5.

23 Collins GS, Moons KG. Comparing risk prediction models. *BMJ* 2012;344:e3186.

24 Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. *BMJ* 2009;339:b4184.

25 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837-45.

26 Douma RA, Mos IC, Erkens PM, et al. Performance of 4 clinical decision rules in the diagnostic management of acute pulmonary embolism. *Ann Intern Med* 2011;154:709-18.

27 Janssen KJ, Vergouwe Y, Donders AR, et al. Dealing with missing predictor values when applying clinical prediction models. *Clin Chem* 2009;55:994-1001.

28 Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. *Stat Med* 1991;10:585-98.

29 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594.

30 Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1-73.

31 Miniati M, Monti S, Bottai M. A structured clinical model for predicting the probability of pulmonary embolism. *Am J Med* 2003;114:173-9.

32 Wicki J, Perneger TV, Junod AF, Bounameaux H, Perrier A. Assessing clinical probability of pulmonary embolism in the emergency ward: a simple score. *Arch Intern Med* 2001;161:92-7.

33 Gibson NS, Sohne M, Kruip MJ, et al. Further validation and simplification of the Wells clinical decision rule in pulmonary embolism. *Thromb Haemost* 2008;99:229-34.

34 Le Gal G, Righini M, Roy PM, et al. Prediction of pulmonary embolism in the emergency department: the revised Geneva score. *Ann Intern Med* 2006;144:165-71.

35 Klok FA, Mos IM, Nijkeuter M, et al. Simplification of the revised Geneva score for assessing clinical probability of pulmonary embolism. *Arch Intern Med* 2008;168:2131-6.

36 Miniati M, Bottai M, Monti S, Salvadori M, Serasini L, Passera M. Simple and accurate prediction of the clinical probability of pulmonary embolism. *Am J Respir Crit Care Med* 2008;178:290-4.

37 Kline JA, Mitchell AM, Kabrhel C, Richman PB, Courtney DM. Clinical criteria to prevent unnecessary diagnostic testing in emergency department patients with suspected pulmonary embolism. *J Thromb Haemost* 2004;2:1247-55.

38 Kline JA, Nelson RD, Jackson RE, Courtney DM. Criteria for the safe use of D-dimer testing in emergency department patients with suspected pulmonary embolism: a multicenter US study. *Ann Emerg Med* 2002;39:144-52.

39 Perrier A, Roy PM, Sanchez O, et al. Multidetector-row computed tomography in suspected pulmonary embolism. *N Engl J Med* 2005;352:1760-8.

40 Geersing GJ, Janssen KJ, Oudega R, et al. Excluding venous thromboembolism using point of care D-dimer tests in outpatients: a diagnostic meta-analysis. *BMJ* 2009;339:b2990.

41 Stein PD, Hull RD, Patel KC, et al. D-dimer for the exclusion of acute venous thrombosis and pulmonary embolism. *Ann Intern Med* 2004;140:589-602.