BMC Pulmonary Medicine

RESEARCH

Open Access

Comparative diagnostic accuracy of pre-test clinical probability scores for the risk stratification of patients with suspected pulmonary embolism: a systematic review and Bayesian network meta-analysis

Ali Etemadi[1], Mohammadmobin Hosseini[1], Hamed Rafiee[1], Amir Mahboubi[1], Tara Mahmoodi[1], Toshiki Kuno[2], Yaser Jenab[1], Claire E. Raphael[3], Wilbert S. Aronow[4], Kaveh Hosseini[1*] and Jay Giri[5]

Abstract

Background The primary evaluation of pulmonary embolism (PE) is complicated by the presence of various pre-test clinical probability scores (pCPS) with different cut-offs, all equally recommended by guidelines. This lack of consensus has led to practice variability, unnecessary imaging, and worse patient outcomes. We aim to provide more definitive insights through a holistic comparison of available pCPS.

Methods PubMed, Embase and Web of Science, and Google Scholar were searched for studies evaluating pCPS in patients clinically suspected of PE until June 2023. Risk of bias was evaluated using QUADAS-2. Included pCPS were evaluated based on their diagnostic accuracy in: (1) Ruling-out PE (2) Utilization of imaging, and (3) Differentiating between patients needing d-dimer from imaging. Diagnostic test accuracy indices were synthesized using beta-binomial Bayesian methods.

Results Forty studies (37,027 patients) were included in the meta-analysis. Three-tier revised Geneva (RG) and three-tier Wells performed similarly in ruling-out PE (negative likelihood ratio (LR-) [95% credible interval (CI)]: 0·39[0·27–0·58] vs 0·34[0·25–0·45]). However, RG performed better in utilization of imaging (LR+: 6·65[3·75–10·56] vs 5·59[3·7–8·37], $p < 0.001$) and differentiating between patients needing d-dimer vs imaging (diagnostic odds ratio (DOR): 8·03[4·35–14·1] vs. 7·4[4·65–11·84], $p < 0.001$). The two-tier Wells score underperformed in all aspects (LR-: 0·56[0·45–0·68], LR+: 2·43[1·81–3·07], DOR: 4·41[2·81–6·43]). PERC demonstrated a reliable point estimate for ruling out PE, albeit with a wide CI (LR-: 0·36[0·17–0·78]).

Conclusions RG outperforms other pCPS for primary evaluation of suspected PE. While the difference is not large, RG's independence from subjective items supports its recommendation over three-tier Wells. Two-tier Wells underperforms significantly compared to the rest of pCPS. PERC shows considerable promise for minimizing unnecessary D-dimer testing in crowded emergency departments; however, more evidence is needed before its definitive recommendation.

*Correspondence:
Kaveh Hosseini
kaveh_hosseini130@yahoo.com
Full list of author information is available at the end of the article

Protocol registration PROSPERO (CRD42023464118).

Keywords Pulmonary Embolism, Risk Assessment, Predictive Value of Tests, Clinical Decision Support System

Background

Pre-test clinical probability scores (pCPS) guide the diagnostic workup of patients with a clinical suspicion of pulmonary embolism (PE) to improve patient outcomes without overburdening healthcare resources [1]. It has been established that increased imaging doesn't predict improved patient outcomes [2]. Additionally, many studies have reported that large proportions (> 50%) of patients with suspected PE are subjected to unwarranted imaging, [3] which could have been prevented with the proper use of pCPS [4]. Despite this, pCPS remain significantly underutilized [5].

One barrier to the implementation of pCPS in practice is the confusion caused by the range of available scores with different cut-offs, all equally recommended in the latest practice guidelines [6]. This is largely due to most studies solely concentrating on comparing the performance of pCPS in sensitively ruling out PE, with major scores demonstrating similar performance [7–12]. However, pCPS performance in ruling out PE, i.e., minimizing the number of patients with PE assigned to the low-probability category, while crucial, does not paint a complete picture of their performance in clinical practice. As pCPS are not designed for final diagnosis but to guide patients toward appropriate diagnostic tests (DTs). Consequently, the ability of each pCPS to assign imaging only to the patients who truly need it, while correctly differentiating between those with high and low probabilities of PE, is equally crucial for the effective management of patients with a clinical suspicion of PE. However, this aspect remains largely underexplored.

In the present review, we outline a framework for a holistic comparison of pCPS and subsequently use this framework to compare the overall effectiveness of available pCPS in risk stratifying the general population of patients suspected of PE.

Methods

Protocol and registration

The protocol for the current review is available on PROSPERO (CRD42023464118).

Eligibility criteria

We included peer-reviewed cross-sectional studies, clinical trials, and retrospective and prospective cohorts reporting on pCPS of adult patients with a clinical suspicion of PE. To improve generalizability, study populations selected based on comorbidities (e.g., COVID) were excluded. The setting was not limited to the emergency department. The following pCPS for PE were considered: 1. Wells score (including simplified, three- or two-tier versions); 2. Geneva score (including original, revised, simplified, three- or two-tier versions); 3. PERC (applied to the entire study population or the low-probability subgroup); 4. YEARS. Given the large amount of available evidence, we restricted the studies to the English language. To be included, a primary study had to report enough DT accuracy indices to allow the calculation of true positives, true negatives, false positives, and false negatives for at least one of the mentioned index tests. For the reference test, in addition to the standard diagnostic methods for PE, such as computed tomography pulmonary angiography (CTPA) and pulmonary angiograms, we included studies that diagnosed or ruled-out PE based on guideline-recommended pathways. For a detailed description of the eligibility criteria, see Appendix 1.

Search strategy

The main searched databases were PubMed, Embase and Web of Science. The first 200 records of Google Scholar were added as a supplementary search. No date restrictions were applied. The complete details of the implemented search strategies are available in Appendix 2.

Study selection and data collection process

Four reviewers independently reviewed titles and abstracts of the first 50 records with final determination reached by group consensus. Afterward, the reviewers each independently screened titles and abstracts of all remaining records. Next, the same reviewers independently screened full-text articles for inclusion. In both stages, each record was at least reviewed by two reviewers, and conflicts were resolved by a third independent reviewer. No automation tools were used at any stage.

Each record was extracted by one reviewer and validated by another. A fifth reviewer further validated each extracted record. If needed, WebPlotDigitizer was used to extract data from the figures of the primary studies [13]. The data extraction form and the relevant definitions are available in Appendix 3. Based on their reported cut-offs, each pCPS was considered either two-tier (patients classified into low- or high-probability categories) or

three-tier (patients classified into low-, moderate-, or high-probability categories).

To maintain data independence, a decision rule was developed before data extraction to avoid the inclusion of multiple reports of one cohort (see Appendix 4) [14]. Baseline data, collected at the time of initial patient assessment, or the earliest available data from each study were included in this meta-analysis. Follow-up pCPS measurements were excluded to avoid potential biases introduced by subsequent changes in patient condition.

Risk of bias and applicability

The assessment of quality and applicability for each study was performed using the QUADAS-2 tool [15]. Each included study was evaluated by one researcher and then verified by another. Conflicts were resolved by a third reviewer. We generally followed the instructions in the background document. However, we assigned comparatively less weight to the following three signaling questions during the evaluation of overall bias risk within their respective domains. Consistent with clinical practice, the selection of the reference test is mostly guided by the results of the index test (e.g., high-probability patients typically undergo CTPA). Consequently, we considered the lack of blinding to the index test results during the interpretation of the reference test as having a lesser impact on the risk of bias. Additionally, patients other than those undergoing CTPA are typically monitored for any changes in their condition over time (generally 3–6 months). Given the impracticality and ethical concerns associated with immediately subjecting all patients to CTPA, we regarded both the time interval between the index test and the reference test, as well as the uniformity of the reference tests across all patients, as having less influence on the risk of bias. No summary score was calculated; however, any domain with high or unclear risk of bias led to the entire study being considered high or unclear risk.

Methodology for synthesis

The framework for comparing pCPS effectiveness

The main statistical challenge in evaluating the effectiveness of pCPS for risk stratification of PE is the use of three-tier scores, particularly the moderate-probability category, which is commonly used in practice. This creates a problem, as virtually all relevant statistical methods require each patient to be classified as either high- or low-probability by each score and cannot accommodate patients who fall into the moderate-probability category, which is neither.

Previous reviews have circumvented this issue by classifying moderate-probability patients as high-probability [8]. As now only the low-probability group remains consistent with the original pCPS cut-off scores, DT indices focused on these patients can be used to compare pCPS performance in ruling out PE. Specifically, negative likelihood ratio (LR-), that measures the likelihood of PE given the classification of a patient as low-probability, can be intuitively used to compare pCPS performance in ruling out PE, with lower LR- showing better performance. However, using this model ('PE-unlikely model'), no conclusions can be drawn about the pCPS performance in correctly assigning patients to imaging, as the current aggregated high-probability patient group is clinically heterogeneous and includes moderate-probability patients, who are generally not assigned to imaging.

To address this issue, we introduced another aggregated model alongside the PE-unlikely model by classifying moderate-probability patients as low-probability (PE-likely model'). Similarly, this allows for an intuitive comparison of pCPS performance in correctly assigning patients to imaging using the positive likelihood ratio (LR+), with higher values indicating better performance. Furthermore, the PE-likely model offers another advantage: the new aggregate low-probability group—composed of both low- and moderate-probability patients— follow a similar diagnostic pathway in practice, usually d-dimers; while the high-probability group in this model are usually assigned to imaging. Therefore, in this model, we can also compare models in terms of separating patients who should be assigned to different diagnostic pathways. This separation can be intuitively measured using diagnostic odds ratio (DOR), a measure of how well a pCPS discriminates between patients needing d-dimer and those needing CTPA. Given that most current recommended approaches to suspected PE primarily depend on deciding between CTPA and d-dimer, this can be considered a measure of integration into current diagnostic pathway. In addition to their intuitive application in our context, LR-, LR+, and DOR are unaffected by PE prevalence, further improving their practicality [16]. Overall, these three indices, calculated through using the two mentioned models, allows a holistic comparison of available pCPS, which reflects their effectiveness in clinical practice.

Notably, the PERC score differs from other included scores in two fundamental ways, which necessitated a more nuanced approach in our analysis. First, in the included primary studies, the PERC score was mostly evaluated in a predetermined low-probability population, whereas the other pCPS were assessed in the general population of patients with a clinical suspicion of PE. As a result, studies evaluating the PERC score in the low-probability group were analyzed independently to ensure the transitivity assumption of the network meta-analysis was met [17]. Second, unlike other scores, the

PERC score is used not to choose between imaging and d-dimer, but rather to decide between d-dimer testing and discharge. As a result, the LR + and DOR in the PE-likely model are less relevant for the PERC score. However, the LR- in the PE-unlikely model holds significant clinical value compared to other pCPS, as a false negative could lead to discharging a patient with undiagnosed PE, potentially resulting in life-threatening consequences.

Furthermore, to demonstrate our findings in a clinical context, we used the calculated indices for each pCPS to simulate how they would perform in a simulated population of 100 patients with a clinical suspicion of PE, with a true PE prevalence of 20%. In this simulated population, we compared pCPS based on how many patients with PE were missed (assigned to low-probability) and how many patients without PE were needlessly assigned to imaging (assigned to high-probability). To adjust for the significant variability in the reported range of observed PE prevalence in practice, we also calculated the number of missed PE cases and unnecessary imaging patients across simulated populations with PE prevalence rates ranging from 5 to 30%, in 5% increments [18].

Statistical analysis

Unique pCPS (considering their cut-offs) with at least 4 relevant records were included in the meta-analysis. In our primary analysis, both described models were built using the Bayesian beta-binomial analysis of variance model for network meta-analysis [19]. In addition to the more intuitive outputs associated with Bayesian models, this method better adjusts for the intrinsic correlation between sensitivity and specificity and the over-dispersion resulting from repeated measures, which provides a better framework for a combined evaluation of direct and indirect comparisons as well [20]. The Bayesian beta-binomial was used to estimate summary sensitivity and specificity with relevant 95% credible intervals (CI) which were used to derive other relevant indices. Given the uninformative priors used, the range of CIs is inversely proportional to the certainty the available data can provide. The model and Markov-chain mixing were evaluated using effective sample size and Gelman-Rubin R-hat indices. ANOVA was used to compare diagnostic metrics across tests, with Tukey's HSD for post-hoc pairwise comparisons using adjusted *p*-values.

As a sensitivity analysis, we deployed bivariate hierarchical models to estimate summary sensitivity and specificity for each test with 95% confidence intervals, the currently established method as recommended by Cochrane [21]. For a detailed description of how the final bivariate hierarchical model was designed and the steps taken to ensure model fit, please refer to Appendix 5.

Subgroup analyses were conducted based on: design (retrospective vs prospective), clinical setting (emergency department vs other settings), recruitment method (clinical suspicion vs availability of tests), PE prevalence, age, sex (male percent of study population), prior history of venous thromboembolism, type of reference standard (CTPA vs others), risk of bias (low vs high or unclear) and length of follow-up, as pre-specified in our protocol. To evaluate their impact, variables representing the mentioned subgroups were individually added as a fixed effect to the hierarchical model.

Heterogeneity (i.e., variability) was evaluated using the reported variance of random effects (VoR) of the sensitivity and specificity for each test and visually using forest and summary ROC plots [21]. Deeks' funnel plot asymmetry test was performed to assess publication bias in the included study population and for each test separately. The statistical analyses were conducted in R 4.3.3 and Python (through Google Colab).

Results

Study selection and characteristics

Details of the search and selection process are provided in Appendix 6 using the PRISMA flow template, leading to 45 studies being included in the present review [22–66]. Baseline characteristics and the number of PE positive and negative patients in defined tiers of each pCPS are presented in appendices 7 and 8, respectively.

After excluding pCPS with < 4 relevant records (e.g., YEARS) and one study with incompatible cut-off values, [41] 40 studies (4 index tests, 37,027 patients) were included in the meta-analysis [22–25, 27–35, 37–40, 42–46, 48–59, 61–66]. The four included pCPS were: 1. Three-tier Wells (cut-offs: 2,6); 2. Two-tier Wells (cut-off: 4); 3. Three-tier revised Geneva (RGS) (cut-offs: 4,11); 4. PERC (cut-off: 1). The items and their respective points for each score are presented in Appendix 9. PERC was evaluated in two separate subgroups, depending on whether it was applied to the entire study population or to the low-probability subgroup. The relevant forest and summary ROC plots for the studies included in the meta-analysis are presented in Appendices 10 and 11. A network graph depicting the number of direct comparisons is presented in Appendix 12. Given the limited number of direct comparisons of pCPS (nine records for four different direct comparisons), we combined both direct and indirect comparisons in our analyses. Most of the included studies had a low risk of bias. For a detailed assessment of risk of bias and applicability for each study using the QUADAS-2 tool see Appendix 13.

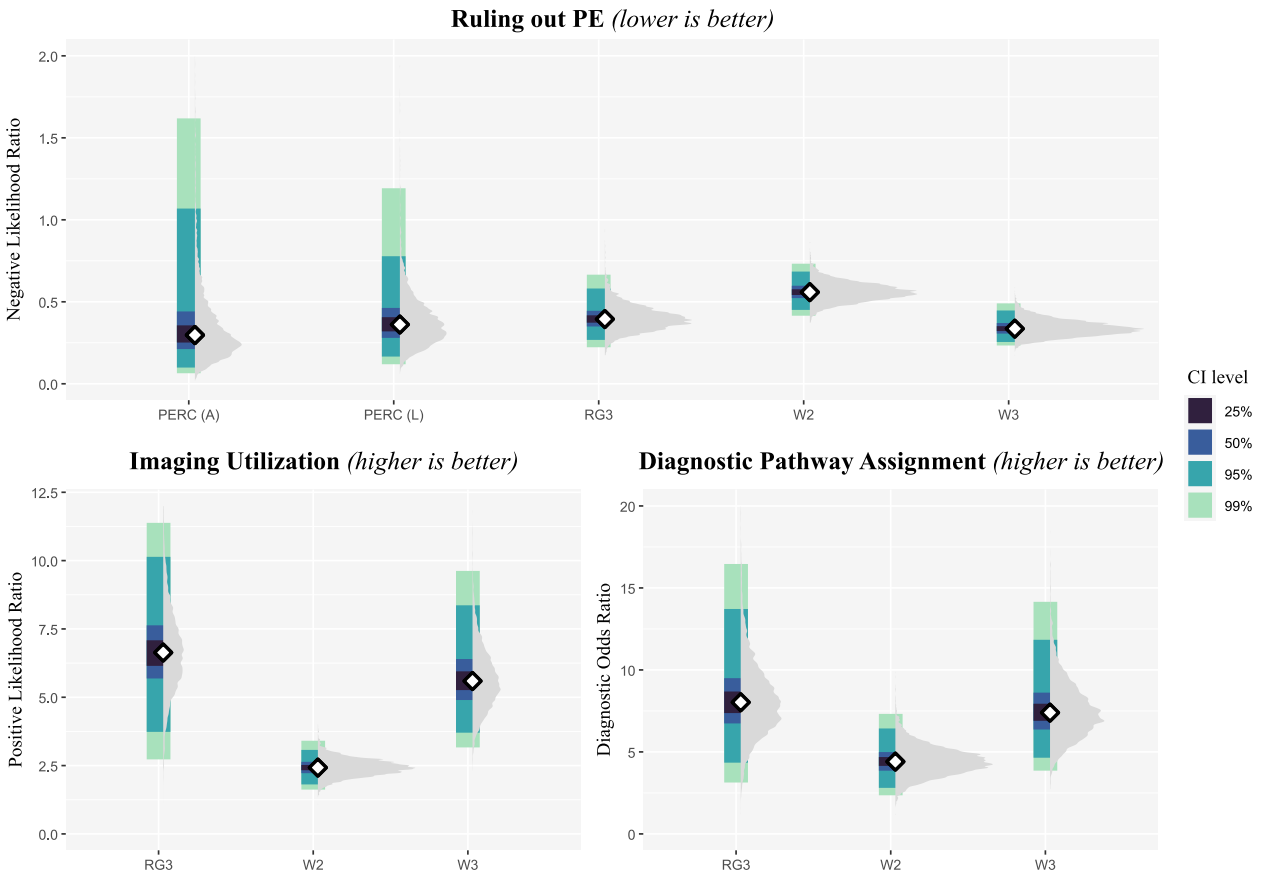

Fig. 1 Comparing the diagnostic accuracy of pre-test clinical probability scores in ruling out PE, imaging utilization and diagnostic pathway assignment. The distribution and four levels of credible intervals for each value are shown. PE: Pulmonary Embolism, PERC(**A**): PERC applied to the general population; PERC(L): PERC applied to low-probability population; RG3: Three-tier revised Geneva; W2: Two-tier Wells; W3: Three-tier Wells; CI: Credible interval

Table 1 Detailed findings of PE-unlikely and likely models based on the Bayesian beta-binomial analysis of variance analysis

Test[a]	Sensitivity: Median (95% CI)	Specificity: Median (95% CI)	LR-: Median (95% CI)	LR+: Median (95% CI)	DOR: Median (95% CI)
PE-unlikely model					
PERC (All)	0·94 (0·81, 0·98)	0·2 (0·09, 0·41)	0·3 (0·1, 1·15)	1·17 (0·98, 1·55)	4·02 (0·85, 14·15)
PERC (Low-proba bility)	0·88 (0·75, 0·94)	0·34 (0·23, 0·48)	0·36 (0·17, 0·78)	1·32 (1·08, 1·68)	3·71 (1·4, 9·45)
Revised Geneva (3 tier)	0·85 (0·76, 0·91)	0·38 (0·3, 0·46)	0·39 (0·27, 0·58)	1·37 (1·24, 1·51)	3·47 (2·19, 5·31)
Wells (2 tier)	0·58 (0·49, 0·68)	0·75 (0·64, 0·82)	0·56 (0·45, 0·68)	2·3 (1·68, 3·02)	4·15 (2·56, 6·16)
Wells (3 tier)	0·82 (0·75, 0·87)	0·55 (0·46, 0·63)	0·34 (0·25, 0·45)	1·81 (1·53, 2·17)	5·41 (3·59, 7·92)
PE-likely model					
PERC (All)	0·94 (0·83, 0·98)	0·19 (0·1, 0·32)	0·3 (0·12, 1·06)	1·16 (0·99, 1·37)	3·91 (0·93, 11·44)
PERC (Low-probability)	0·88 (0·76, 0·94)	0·33 (0·24, 0·45)	0·36 (0·16, 0·74)	1·31 (1·1, 1·6)	3·73 (1·49, 9·35)
Revised Geneva (3 tier)	0·2 (0·13, 0·3)	0·97 (0·95, 0·98)	0·83 (0·73, 0·89)	6·65 (3·75, 10·56)	8·03 (4·35, 14·1)
Wells (2 tier)	0·58 (0·48, 0·67)	0·76 (0·68, 0·82)	0·55 (0·45, 0·67)	2·43 (1·81, 3·07)	4·41 (2·81, 6·43)
Wells (3 tier)	0·28 (0·22, 0·35)	0·95 (0·92, 0·97)	0·76 (0·68, 0·82)	5·59 (3·7, 8·37)	7·4 (4·65, 11·84)

[a] It should be noted that only the three-tier tests have different values between the two models. As PERC and Wells (2-tier) scores are characterized by only two defined categories (low- and high-probability), their indices remain relatively consistent across models. The slight variations arise from the inherent sampling variability in the Bayesian models

PE Pulmonary embolism, *LR* Likelihood ratio, *DOR* Diagnostic odds ratio, *CI* Credible interval

Synthesis of results

A comparison of the summary diagnostic accuracy indices for each pCPS and their distribution is presented in Fig. 1, with the full report available in Table 1. With an R hat statistic below 1·01 and an effective sample size (ESS) > 680 for the relevant variables, both models demonstrated robust convergence and mixing.

In the PE-unlikely model, the three-tier Wells and RGS had the best overall performance in sensitively ruling-out PE, which was demonstrated by lower LR- and higher sensitivity estimates with narrow 95% CIs. PERC, whether applied to the low-probability or the entire population, had a higher point estimate. However, their large CIs demonstrated potential uncertainty in this estimate. The two-tier Wells score performed considerably worse in comparison to the rest of the scores. In the PE-likely model, the RGS outperformed the three-tier Wells and had a significantly higher specificity, LR + and DOR. The two-tier Wells score performed considerably worse in comparison to the rest of the scores in this model as well. PERC was not considered in this comparison as it is not designed for ruling-in PE. All comparisons were statistically significant ($p < 0.001$).

In the simulated population of 100 patients with a true PE prevalence of 20%, PERC (low-probability), three-tier Wells and RGS missed 3 patients with PE by classifying them as low-probability. In contrast, the two-tier Wells missed more than 8 patients with PE. In the same simulated population, the number of patients without PE assigned to imaging (falsely classified as high-probability) was around 2 for RGS, 4 for three-tier Wells, and 19 for two-tier Wells. For their performance in populations across the entire range of plausible PE prevalences see Appendix 14 (similar trends were observed).

Additional analyses

Bivariate hierarchical model (sensitivity analysis)

The results of the bivariate hierarchical models for both the PE-unlikely and likely models are reported in Appendix 15, with highly comparable results to the Bayesian models. However, the bivariate models yielded slightly more optimistic estimates with significantly narrower confidence intervals compared to Bayesian models.

Heterogeneity, meta-regression, and subgroup analyses

In the PE-unlikely model, the three-tier Wells score displayed a high level of heterogeneity in both sensitivity and specificity (VoR ~ 1), followed by a moderate-to-high level of heterogeneity in the two-tier Wells score in both indices (VoR ~ 0·5). The rest of the indices reported a low-to-moderate level of heterogeneity. These findings can be visibly observed in the forest and sROC plots as well (Appendices 10 and 11).

To address heterogeneity in the PE-unlikely model, we evaluated pre-specified subgroup analyses and meta-regressions. In the meta-regression, only PE prevalence (0·43 [0·40 to 0·46], $p < 0·001$) and setting (0·59 [0·53 to 0·66], $p = 0·006$) of the primary studies significantly correlated with correctly identifying those with and without PE. Furthermore, the primary studies with lower sample sizes had a significantly higher prevalence of PE (Pearson's correlation coefficient: -0·40, p-value < 0·001), due to a higher possibility of sampling errors. Consequently, we limited the model to primary studies with sample sizes of > 250 and > 500, as higher sample sizes are more likely to provide a stable sample, reducing the impact of random errors and improving the validity of the findings. We observed a dose–response relationship between increasing the minimum sample size of included primary studies and a reduction in heterogeneity. In the subgroup analysis of studies with > 500 sample size or studies in the ED setting with > 500 sample size, the amount of VoR for most indices was roughly halved, resulting in low-to-moderate levels of heterogeneity for all pCPS. Details of our step-by-step approach to addressing sources of heterogeneity are presented in Appendix 16.

The PE-likely model demonstrated low-to-moderate levels of heterogeneity in the baseline model, (maximum VoR ~ 0·6) as it is apparent in the forest and sROC plots as well. Even so, limiting the model to primary studies with > 500 sample size resulted in lower levels of heterogeneity (maximum VoR ~ 0·35).

The bivariate hierarchical model findings, limited to studies with sample sizes > 500 or studies in the ED setting with sample sizes > 500, are reported in Appendices 17 and 18. Notably, these findings were similar and now even closer to the baseline Bayesian model (Table 1). Additionally, given our strict inclusion criteria for retrospective studies (Appendix 1), excluding those that met these criteria did not alter the findings, which remained highly comparable to the rest of the models (Appendix 19).

Publication bias

The funnel plot test for DORs (Deeks' test) indicated the absence of any significant publication bias in all included studies ($p = 0·78$ and 0·40 for the PE-unlikely and likely models, respectively). Furthermore, no evidence for publication bias was found when each pre-test clinical probability score was evaluated separately.

Discussion

We introduced a comprehensive framework to compare the effectiveness of p-CPS in a general population of patients with suspected PE. This approach allowed us to evaluate the reliability of pCPS in ruling out PE and

imaging utilization. Additionally, we evaluated how effectively they integrate into current diagnostic pathways by assessing their ability to differentiate between patients who require d-dimer vs. CTPA. The RGS and three-tier Wells scores performed similarly in ruling out PE, but the RGS was more effective in imaging utilization and assigning patients to the right diagnostic tests. These findings were reflected in the simulated populations of patients with suspected PE, in which the RGS was associated with fewer unnecessary imaging assignments across the full range of reported PE prevalences, and even slightly fewer missed PE cases. The two-tier Wells score performed worse in all areas. While PERC showed potential as the most reliable p-CPS for ruling out PE, its wide CI suggest that the available data do not provide enough certainty in the derived summary estimates.

Prior systematic reviews and landmark studies have mostly focused on comparing the pCPS in ruling out PE. A rigorous systematic review and meta-analysis, [8] reported the two-tier Wells being significantly worse at ruling out PE, while the three-tier Wells and RGS demonstrated similar performances with values comparable to the present review. Furthermore, a landmark prospective cohort compared two-tier pCPS and reported sensitivity and specificity values for the two-tier Wells consistent with our findings [49]. These findings are in line with our conclusion of 2-tier Wells not being on par with other scores and three-tier Wells and RGS performing similarly in ruling out PE. While we recognize that 2-tier pCPS are more feasible to implement compared to their 3-tier counterparts, there may be a considerable tradeoff in the general accuracy of the approach, especially in the case of the Wells score. However, it can be argued that the 2-tier approach may encourage more widespread utilization of pCPS. Nevertheless, as automation increasingly shapes these diagnostic processes, the compromise in precision may outweigh the benefits of ease of use. Given the lower popularity of the 2-tier Geneva scores, there were insufficient studies to provide a meaningful comparison.

There is a gap in the literature on comparing pCPS in imaging utilization and their ability in differentiating between low- and high-probability patients, particularly in studies focused on pCPS performance alone (i.e., independent of subsequent diagnostic tests like d-dimers). By considering these aspects, the underperformance of the two-tier Wells and the advantages of RGS over the three-tier Wells become readily observable, especially in the clinical context. Additionally, a number of prior reviews have conducted meta-analyses using methods not specifically tailored for the synthesis of DT accuracy indices [7–12]. While more straightforward to conduct and interpret, these methods may lead to bias, [67] and this variability in methodologies limits cross-study comparisons. Therefore, employing comprehensive frameworks with specialized meta-analysis methods could facilitate a more effective translation of findings into clinical practice.

Applying PERC to rule out PE in low-probability patients without the need for further DTs is highly appealing, especially given its operational simplicity, which has been recommended in a number of guidelines with varying levels of support [6]. However, because PERC may lead to discharging its designated low-probability patients without further testing, it lacks the safety net found in other pCPS, where false-negatives are often caught with d-dimer, a highly sensitive test. This increases the risk that a false-negative on PERC could result in premature discharge and a potentially fatal missed diagnosis. Furthermore, there are two important sources of heterogeneity in the available primary studies reporting on the diagnostic accuracy of PERC. First, some previous primary studies have reported on the diagnostic accuracy of PERC in all patients with a clinical suspicion of PE, in contrast to limiting it to pre-determined low-probability patients. Furthermore, the pre-determination of low-probability patients has been based on various pCPS (including Wells, RGS, and clinical gestalt) [23, 31, 64]. While we analyzed the studies reported on PERC in the low-probability group separately, a further subgroup on different types of pre-determining low-probability would have made the meta-analysis unfeasible due to the low number of homogenous studies. Nevertheless, some previous meta-analyses have pooled all PERC studies together, disregarding the mentioned sources of heterogeneity, which may explain the differences in our findings [68]. Additionally, a number of European studies, which generally report a higher PE prevalence, have shown a considerably worse performance of PERC [64, 69]. We acknowledge that much of the uncertainty surrounding the PERC score stems from its relative novelty. However, the mentioned factors, added to the unique position of the PERC score in the current diagnostic pathways, and the wide 95% CI of our synthesized LR-, merits a cautious stance regarding the definitive recommendation of the PERC score, given the currently available evidence. Our reservations about recommending PERC for low-probability PE patients align with recent concerns raised by the European Society of Cardiology, highlighting the issues surrounding its generalizability [70].

The limitations of the present review are as follows. First, despite extensive sensitivity analyses, a low-to-moderate level of heterogeneity remains for some of the reported outcomes, which may affect our findings. Second, to minimize statistical dependency from studies that reused participants, we included only one record per cohort, which may have led to bias by excluding some relevant studies. By blindly designing a decision rule for

record selection before data extraction, we minimized this source of bias. Third, due to practical and ethical concerns, using CTPA to finalize diagnosis in all patients is not feasible in most studies, which can lead to an overestimation of accuracy. However, adjusting for different reference standards didn't have any significant impact on our model outcomes. Fourth, direct comparisons between pCPS were rarely reported in the included studies, which may affect the robustness of our network meta-analysis. However, our use of a Bayesian beta-binomial analysis of variance model for the network meta-analysis, designed to combine direct and indirect comparisons, has minimized this bias. That being said, the sophistication of our analyses required at least four studies for a valid synthesis, which led to the exclusion of a few pCPS, notably including YEARS, which limits the broader applicability of the present review. Furthermore, to ensure proper fitting of the bivariate hierarchical model, certain parameters were fixed to address high correlation (e.g., specificity for RGS), which, although justified, may have influenced this model's findings. However, since the beta-binomial model converged without the need to fix any parameters and produced highly comparable results, this likely had minimal impact on our findings. Finally, we limited the included studies to English-language publications, but given the large body of available evidence, it likely did not significantly affect our outcomes.

Conclusion

In conclusion, the RGS integrates better into the PE diagnostic pathways and outperforms three-tier Wells in the clinical setting. Although the difference isn't too large, the independence of RGS from subjective clinical judgment adds to the argument for its recommendation. On the other hand, the two-tier Wells score generally underperforms in comparison to the other pCPS. PERC shows promise in reducing unnecessary testing in low-probability patients, and with further evidence, particularly in regions with varying PE prevalence, it could potentially serve as a useful tool for clinical decision-making.

Abbreviations

pCPS	Pre-Test Clinical Probability Score
PE	Pulmonary Embolism
DT	Diagnostic Test
CTPA	Computed Tomography Pulmonary Angiography
DOR	Diagnostic Odds Ratio
RGS	(Three-Tier) Revised Geneva Score
VoR	Variance of Random Effects

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12890-025-03637-6.

Supplementary Material 1.

Acknowledgements

Not applicable.

Authors' contributions

A.E.: Conceptualization, Formal Analysis, Methodology, Project Administration, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing. M.H. , H.R. , A.M. and T.M.: Data Curation, Writing – Original Draft Preparation, Writing – Review & Editing. J.G. : Conceptualization, Methodology, Writing – Review & Editing C.R. , T.K. , Y.J and W. A.: Conceptualization, Writing – Review & Editing K.H. : Conceptualization, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing.

Funding

The authors state that no specific financial or non-financial support was received for the present review.

Data availability

All meta-analytic data and analysis codes and scripts (for both R and Python) are publicly available at the study's associated repository on the Open Science Framework (https://osf.io/pbfta/?view_only=ad5a72f9606741749488bf3578dbf07b).

Declarations

Ethics approval and consent to participate

This systematic review and meta-analysis did not require ethical approval or informed consent as it involves the analysis of publicly available data from previously published studies.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Tehran Heart Center, Cardiovascular Diseases Research Institute, Tehran University of Medical Sciences, Tehran, Iran. [2]Cardiology Division, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [3]Department of Cardiovascular Medicine, Mayo Clinic College of Medicine, Rochester, MN, USA. [4]Department of Cardiology, Westchester Medical Center, New York Medical College, Valhalla, NY, USA. [5]Perelman School of Medicine, Cardiovascular Medicine Division, University of Pennsylvania, Philadelphia, PA, USA.

Received: 7 January 2025 Accepted: 28 March 2025
Published online: 08 April 2025

References

1. Interventions to Reduce the Overuse of Imaging for Pulmonary Embolism: A Systematic Review - Deblois - 2018 - Journal of Hospital Medicine - Wiley Online Library. Available from: https://shmpublications.onlinelibrary.wiley.com/doi/https://doi.org/10.12788/jhm.2902. Cited 2025 Mar 26.
2. Raja AS, Greenberg JO, Qaseem A, Denberg TD, Fitterman N, Schuur JD, et al. Evaluation of patients with suspected acute pulmonary embolism: best practice advice from the clinical guidelines committee of the american college of physicians. Ann Intern Med. 2015;163(9):701–11.
3. Crichlow A, Cuker A, Mills AM. Overuse of Computed Tomography Pulmonary Angiography in the Evaluation of Patients with Suspected Pulmonary Embolism in the Emergency Department. Acad Emerg Med. 2012;19(11):1219–26.
4. Perera M, Aggarwal L, Scott IA, Cocks N. Underuse of risk assessment and overuse of computed tomography pulmonary angiography in patients with suspected pulmonary thromboembolism. Intern Med J. 2017;47(10):1154–60.
5. Venkatesh AK, Kline JA, Courtney DM, Camargo CA, Plewa MC, Nordenholz KE, et al. Evaluation of Pulmonary Embolism in the Emergency Department and Consistency With a National Quality Measure:

Quantifying the Opportunity for Improvement. Arch Intern Med. 2012;172(13). Available from: http://archinte.jamanetwork.com/article.aspx?doi=https://doi.org/10.1001/archinternmed.2012.1804. Cited 2024 Jun 19.
6. Falster C, Hellfritzsch M, Gaist TA, Brabrand M, Bhatnagar R, Nybo M, et al. Comparison of international guideline recommendations for the diagnosis of pulmonary embolism. Lancet Haematol. 2023;10(11):e922–35.
7. Van Maanen R, Martens ESL, Takada T, Roy PM, De Wit K, Parpia S, et al. Accuracy of physicians' intuitive risk estimation in the diagnostic management of pulmonary embolism: an individual patient data meta-analysis. J Thromb Haemost. 2023;21(10):2873–83.
8. Lucassen W, Geersing GJ, Erkens PMG, Reitsma JB, Moons KGM, Büller H, et al. Clinical decision rules for excluding pulmonary embolism: A meta-analysis. Ann Intern Med. 2011;155(7):448.
9. Bass A, Fields K, Goto R, Turissini G, Dey S, Russell L. Clinical decision rules for pulmonary embolism in hospitalized patients: a systematic literature review and meta-analysis. Thromb Haemost. 2017;117(11):2176–85.
10. Ceriani E, Combescure C, Le Gal G, Nendaz M, Perneger T, Bounameaux H, et al. Clinical prediction rules for pulmonary embolism: a systematic review and meta-analysis. J Thromb Haemost. 2010;8(5):957–70.
11. Shen JH, Chen HL, Chen JR, Xing JL, Gu P, Zhu BF. Comparison of the Wells score with the revised Geneva score for assessing suspected pulmonary embolism: a systematic review and meta-analysis. J Thromb Thrombolysis. 2016;41(3):482–92.
12. Ma M, Li Y, Xu X, Ji C. Early diagnosis for pulmonary embolism: A systematic review and meta-analysis. Medicine (Baltimore). 2023;102(28):e34352.
13. Drevon D, Fursa SR, Malcolm AL. Intercoder Reliability and Validity of WebPlotDigitizer in Extracting Graphed Data. Behav Modif. 2017;41(2):323–39.
14. López-López JA, Page MJ, Lipsey MW, Higgins JPT. Dealing with effect size multiplicity in systematic reviews and meta-analyses. Res Synth Methods. 2018;9(3):336–51.
15. Whiting PF. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med. 2011;155(8):529.
16. Marill KA. Diagnostic and prognostic test assessment in emergency medicine: likelihood and diagnostic odds ratios. Emerg Med J. 2022;39(8):635–42.
17. Phillips MR, Steel DH, Wykoff CC, Busse JW, Bannuru RR, Thabane L, et al. A clinician's guide to network meta-analysis. Eye. 2022;36(8):1523–6.
18. Germini F, Zarabi S, Eventov M, Turcotte M, Li M, de Wit K. Pulmonary embolism prevalence among emergency department cohorts: A systematic review and meta-analysis by country of study. J Thromb Haemost. 2021;19(1):173–85.
19. N Nyaga V, Arbyn M, Aerts M. Beta-binomial analysis of variance model for network meta-analysis of diagnostic test accuracy data. Stat Methods Med Res. 2018 Aug;27(8):2554–66.
20. Veroniki AA, Tsokani S, Agarwal R, Pagkalidou E, Rücker G, Mavridis D, et al. Diagnostic test accuracy network meta-analysis methods: A scoping review and empirical assessment. J Clin Epidemiol. 2022;146:86–96.
21. Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y. Chapter 9: Understanding meta-analysis. In: Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy [Internet]. Version 2.0. Cochrane; 2023. (July 2023). Available from: https://training.cochrane.org/handbook-diagnostic-test-accuracy/current
22. Esiéné A, Tochie JN, Metogo JAM, Etoundi PO, Minkande JZ. A comparative analysis of the diagnostic performances of four clinical probability models for acute pulmonary embolism in a sub-Saharan African population: a cross-sectional study. BMC Pulm Med. 2019;19(1):263.
23. Theunissen J, Scholing C, Van Hasselt W, Van Der Maten J, Ter Avest E. A retrospective analysis of the combined use of PERC rule and Wells score to exclude pulmonary embolism in the Emergency Department. Emerg Med J. 2016;33(10):696–701.
24. Kucher N, Kohler HP, Dornhöfer T, Wallmann D, Lämmle B. Accuracy of d-dimer/fibrinogen ratio to predict pulmonary embolism: a prospective diagnostic study. J Thromb Haemost. 2003;1(4):708–13.
25. Kearon C, Ginsberg JS, Douketis J, Turpie AG, Bates SM, Lee AY, et al. An Evaluation of D -dimer in the diagnosis of pulmonary embolism: a randomized Trial. Ann Intern Med. 2006;144(11):812.
26. 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(1):92.
27. Robert-Ebadi H, Mostaguir K, Hovens MM, Kare M, Verschuren F, Girard P, et al. Assessing clinical probability of pulmonary embolism: prospective validation of the simplified Geneva score. J Thromb Haemost. 2017;15(9):1764–9.
28. Penaloza A, Mélot C, Dochy E, Blocklet D, Gevenois PA, Wautrecht JC, et al. Assessment of pretest probability of pulmonary embolism in the emergency department by physicians in training using the Wells model. Thromb Res. 2007;120(2):173–9.
29. Hasanoğlu C, Argüder E, Kılıç H, Parlak ES, Karalezli A. Atrial fibrillation, an obscured cause of pulmonary embolism can be revealed by adding to Wells criteria. J Investig Med. 2019;67(7):1042–7.
30. Steeghs N, Goekoop RJ, Niessen RWLM, Jonkers GJPM, Dik H, Huisman MV. C-reactive protein and D-dimer with clinical probability score in the exclusion of pulmonary embolism. Br J Haematol. 2005;130(4):614–9.
31. 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(8):1247–55.
32. Miniati M, Bottai M, Monti S. Comparison of 3 clinical models for predicting the probability of pulmonary embolism. Medicine (Baltimore). 2005;84(2):107–14.
33. Sanson BJ, Lijmer J, Mac Gillavry M, Turkstra F, Prins M, Büller H, et al. Comparison of a clinical probability estimate and two clinical models in patients with suspected pulmonary embolism. Thromb Haemost. 2000;83(02):199–203.
34. Gökharman FD. Comparison of multidetector computed tomography findings with clinical and laboratory data in pulmonary thromboembolism. Pol J Radiol. 2015;80:252–8.
35. Penaloza A, Verschuren F, Meyer G, Quentin-Georget S, Soulie C, Thys F, et al. Comparison of the unstructured clinician gestalt, the wells score, and the revised geneva score to estimate pretest probability for suspected pulmonary embolism. Ann Emerg Med. 2013;62(2):117-124.e2.
36. De Wit K, Al-Haimus F, Hu Y, Ikesaka R, Chan N, Ibrahim Q, et al. Comparison of YEARS and Adjust-Unlikely D-dimer testing for pulmonary embolism in the emergency department. Ann Emerg Med. 2023;81(5):558–65.
37. Obradović D, Joveš B, Pena Karan S, Stefanović S, Ivanov I, Vukoja M. Correlation between the W ells score and the Q uanadli index in patients with pulmonary embolism. Clin Respir J. 2016;10(6):784–90.
38. Gupta RT, Kakarla RK, Kirshenbaum KJ, Tapson VF. D -Dimers and Efficacy of Clinical Risk Estimation Algorithms: Sensitivity in Evaluation of Acute Pulmonary Embolism. Am J Roentgenol. 2009;193(2):425–30.
39. Wells P, Anderson D, Rodger M, Ginsberg J, Kearon C, Gent 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(03):416–20.
40. Hogg K, Thomas D, Mackway-Jones K, Lecky F, Cruickshank K. Diagnosing pulmonary embolism: a comparison of clinical probability scores. Br J Haematol. 2011;153(2):253–8.
41. Kearon C, De Wit K, Parpia S, Schulman S, Afilalo M, Hirsch A, et al. Diagnosis of pulmonary embolism with D-dimer adjusted to clinical probability. N Engl J Med. 2019;381(22):2125–34.
42. Freund Y, Cachanado M, Aubry A, Orsini C, Raynal PA, Féral-Pierssens AL, et al. effect of the pulmonary embolism rule-out criteria on subsequent thromboembolic events among low-risk emergency department Patients: The PROPER randomized clinical trial. JAMA. 2018;319(6):559.
43. Galipienzo J, Garcia de Tena J, Flores J, Alvarez C, Garcia-Avello A, Arribas I. Effectiveness of a diagnostic algorithm combining clinical probability, D-dimer testing, and computed tomography in patients with suspected pulmonary embolism in an emergency department. Romanian J Intern Med Rev Roum Med Interne. 2012;50(3):195-202.
44. Effectiveness of Managing Suspected Pulmonary Embolism Using an Algorithm Combining Clinical Probability, D-Dimer Testing, and Computed Tomography. JAMA. 2006;295(2):172.
45. Golshani K, Sharafsaleh M. Evaluation of the Diagnostic Value of Bedside Transthoracic Ultrasonography (TTUS) and Lower Extremity Three-Points Compression Duplex in the Diagnosis of the Pulmonary Embolism. J Diagn Med Sonogr. 2020;36(5):423–30.
46. Wells PS, Anderson DR, Rodger M, Stiell I, Dreyer JF, Barnes D, et al. Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and D -dimer. Ann Intern Med. 2001;135(2):98.

47. Aujesky D, Hayoz D, Yersin B, Perrier A, Barghouth G, Schnyder P, et al. Exclusion of pulmonary embolism using C-reactive protein and D-dimer: A prospective comparison. Thromb Haemost. 2003;90(12):1198–203.
48. Kline JA, Hogg M. Measurement of expired carbon dioxide, oxygen and volume in conjunction with pretest probability estimation as a method to diagnose and exclude pulmonary venous thromboembolism. Clin Physiol Funct Imaging. 2006;26(4):212–9.
49. Douma RA. performance of 4 clinical decision rules in the diagnostic management of acute pulmonary embolism: a prospective cohort study. Ann Intern Med. 2011;154(11):709.
50. Calisir C, Yavas US, Ozkan IR, Alatas F, Cevik A, Ergun N, et al. Performance of the wells and revised geneva scores for predicting pulmonary embolism. Eur J Emerg Med. 2009;16(1):49–52.
51. Posadas-Martínez ML, Vázquez FJ, Giunta DH, Waisman GD, De Quirós FGB, Gándara E. Performance of the Wells score in patients with suspected pulmonary embolism during hospitalization: A delayed-type cross sectional study in a community hospital. Thromb Res. 2014;133(2):177–81.
52. Le Gal G, Righini M, Roy PM, Sanchez O, Aujesky D, Bounameaux H, et al. Prediction of pulmonary embolism in the emergency department: the revised Geneva score. Ann Intern Med. 2006;144(3):165.
53. Sanjuán P, Rodríguez-Núñez N, Rábade C, Lama A, Ferreiro L, González-Barcala FJ, et al. Probability scores and diagnostic algorithms in pulmonary embolism: are they followed in clinical practice? Arch Bronconeumol Engl Ed. 2014;50(5):172–8.
54. Kline JA, Courtney DM, Kabrhel C, Moore CL, Smithline HA, Plewa MC, et al. Prospective multicenter evaluation of the pulmonary embolism rule-out criteria. J Thromb Haemost. 2008;6(5):772–80.
55. Kline JA, Runyon MS, Webb WB, Jones AE, Mitchell AM. prospective study of the diagnostic accuracy of the simplify D-dimer assay for pulmonary embolism in emergency department patients. Chest. 2006;129(6):1417–23.
56. Penaloza A, Soulié C, Moumneh T, Delmez Q, Ghuysen A, El Kouri D, et al. Pulmonary embolism rule-out criteria (PERC) rule in European patients with low implicit clinical probability (PERCEPIC): a multicentre, prospective, observational study. Lancet Haematol. 2017;4(12):e615–21.
57. Leclercq M, Lutisan J, Marwijk Kooy M, Kuipers B, Oostdijk A, Van Der Leur J, et al. Ruling out clinically suspected pulmonary embolism by assessment of clinical probability and D-dimer levels: a management study. Thromb Haemost. 2003;89(01):97–103.
58. Geersing GJ, Erkens PMG, Lucassen WAM, Buller HR, Cate HT, Hoes AW, 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(oct04 2):e6564–e6564.
59. Goekoop RJ, Steeghs N, Niessen RWLM, Jonkers GJPM, Dik H, Castel A, et al. Simple and safe exclusion of pulmonary embolism in outpatients using quantitative D-dimer and Wells' simplified decision rule. Thromb Haemost. 2007;97(01):146–50.
60. Van Der Hulle T, Cheung WY, Kooij S, Beenen LFM, Van Bemmel T, Van Es J, et al. Simplified diagnostic management of suspected pulmonary embolism (the YEARS study): a prospective, multicentre, cohort study. The Lancet. 2017;390(10091):289–97.
61. Kabrhel C. The contribution of the subjective component of the canadian pulmonary embolism score to the overall score in emergency department patients. Acad Emerg Med. 2005;12(10):915–20.
62. Kurt OK, Alpar S, Sipit T, Guven SF, Erturk H, Demirel MK, et al. The diagnostic role of capnography in pulmonary embolism. Am J Emerg Med. 2010;28(4):460–5.
63. Castelli R, Bergamaschini L, Sailis P, Pantaleo G, Porro F. The impact of an aging population on the diagnosis of pulmonary embolism: comparison of young and elderly patients. Clin Appl Thromb. 2009;15(1):65–72.
64. Hugli O, Righini M, Le Gal G, Roy P-M, Sanchez O, Verschuren F, et al. The pulmonary embolism rule-out criteria (PERC) rule does not safely exclude pulmonary embolism. J Thromb Haemost. 2011;9(2):300–4.
65. Anderson DR, Kovacs MJ, Dennie C, Kovacs G, Stiell I, Dreyer J, et al. Use of spiral computed tomography contrast angiography and ultrasonography to exclude the diagnosis of pulmonary embolism in the emergency department. J Emerg Med. 2005;29(4):399–404.
66. Aydoğdu M, Topbaşi SiNanoğlu N, Doğan NÖ, Oğuzülgen İK, DemïRcan A, BïLdïK F, et al. Wells Score and Pulmonary Embolism Rule Out Criteria in Preventing Over Investigation of Pulmonary Embolism in Emergency Departments. Tuberk Ve Toraks. 2014 Apr 19;12–21.
67. Salameh JP, Bossuyt PM, McGrath TA, Thombs BD, Hyde CJ, Macaskill P, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ. 2020;14:m2632.
68. Singh B, Parsaik AK, Agarwal D, Surana A, Mascarenhas SS, Chandra S. Diagnostic accuracy of pulmonary embolism rule-out criteria: a systematic review and meta-analysis. Ann Emerg Med. 2012;59(6):517-520.e4.
69. Righini M, Le gal G, Perrier A, Bounameaux H. More on: clinical criteria to prevent unnecessary diagnostic testing in emergency department patients with suspected pulmonary embolism. J Thromb Haemost. 2005;3(1):188–9.
70. Konstantinides SV, Meyer G, Becattini C, Bueno H, Geersing GJ, Harjola VP, et al. 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS). Eur Heart J. 2020;41(4):543–603.

Publisher's Note