Covariate-specific ROC curve analysis can accommodate differences between covariate subgroups in the evaluation of diagnostic accuracy

Lee, J.Y.; Es, N. van; Takada, T.; Klok, F.A.; Geersing, G.J.; Blume, J.; ... ; IPD study team

Journal of Clinical Epidemiology 160 (2023) 14−23

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Covariate-specific ROC curve analysis can accommodate differences between covariate subgroups in the evaluation of diagnostic accuracy

Jenny Lee[a,*], Nick van Es[b,c], Toshihiko Takada[d,e], Frederikus A. Klok[f], Geert-Jan Geersing[d], Jeffrey Blume[e,g], Patrick M. Bossuyt[a], the IPD study team[1]

[a]*Epidemiology and Data Science, Amsterdam UMC location University of Amsterdam, Amsterdam, The Netherlands*
[b]*Department of Vascular Medicine, Amsterdam UMC, location University of Amsterdam, Amsterdam, The Netherlands*
[c]*Amsterdam Cardiovascular Sciences, Pulmonary Hypertension & Thrombosis, Amsterdam, The Netherlands*
[d]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[e]*Department of General Medicine, Shirakawa Satellite for Teaching And Research (STAR), Fukushima Medical University, 2-1 Toyochi Kamiyajiro, Shirakawa, Fukushima, 961-0005, Japan*
[f]*Department of Medicine - Thrombosis and Hemostasis, Leiden University Medical Center, Leiden University, Leiden, The Netherlands*
[g]*Department of Data Science, University of Virginia, Charlottesville, VA, USA*

Accepted 1 June 2023; Published online 8 June 2023

Abstract

Objectives: We present an illustrative application of methods that account for covariates in receiver operating characteristic (ROC) curve analysis, using individual patient data on D-dimer testing for excluding pulmonary embolism.

Study Design and Setting: Bayesian nonparametric covariate-specific ROC curves were constructed to examine the performance/positivity thresholds in covariate subgroups. Standard ROC curves were constructed. Three scenarios were outlined based on comparison between subgroups and standard ROC curve conclusion: (1) identical distribution/identical performance, (2) different distribution/identical performance, and (3) different distribution/different performance. Scenarios were illustrated using clinical covariates. Covariate-adjusted ROC curves were also constructed.

Results: Age groups had prominent differences in D-dimer concentration, paired with differences in performance (Scenario 3). Different positivity thresholds were required to achieve the same level of sensitivity. D-dimer had identical performance, but different distributions for YEARS algorithm items (Scenario 2), and similar distributions for sex (Scenario 1). For the later covariates, comparable positivity thresholds achieved the same sensitivity. All covariate-adjusted models had AUCs comparable to the standard approach.

Conclusion: Subgroup differences in performance and distribution of results can indicate that the conventional ROC curve is not a fair representation of test performance. Estimating conditional ROC curves can improve the ability to select thresholds with greater applicability.

Keywords: Diagnostic accuracy study; ROC curve; Covariate-adjustment; Subgroup analysis; D-dimer; Pulmonary embolism

1. Introduction

Biomarkers are regularly investigated for their ability to classify subjects as diseased or nondiseased. Receiver operating characteristic (ROC) curves are, unarguably, the most widely used tool for evaluating the discriminatory capacity, initially popular with the evaluation of imaging modalities. Their use has now spread to all tests that deliver results on an ordinal, interval or ratio scale [1]. The overall diagnostic accuracy of a medical test is then expressed as the corresponding area under the ROC curve (AUC). The shape of a ROC curve illustrates the trade-off between the sensitivity and specificity of a test at various positivity thresholds,

Funding: The LITMUS project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 777377. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA.

[1] Study team members can be found in the Supplement.

* Corresponding author. Epidemiology and Data Science, Amsterdam UMC, Location AMC, Meibergdreef 9, 1105AZ Amsterdam, The Netherlands. Tel.: +31-0-205668520.

E-mail address: j.a.lee@amsterdamumc.nl (J. Lee).

https://doi.org/10.1016/j.jclinepi.2023.06.001

What is new?

Key findings

- Receiver operating characteristics (ROC) curves can fluctuate between covariate subgroups and in some cases result in differences in the overall performance, expressed as the area under the ROC curve (AUC).
- Prominent subgroup differences in the distribution of index test results and performance may indicate that different positivity thresholds are required to achieve the same level of sensitivity or specificity.

What this adds to what was known?

- Differences in test performance and distribution of index test results between covariate subgroups can indicate that the standard ROC curve is not a fair representation of test performance.
- It is possible that covariate-specific ROC curves are completely identical between subgroups, when paired with identical distribution of index test results, we can assume the standard AUC expresses performance well.

What is the implication and what should change now?

- Several methods for conditional ROC curve analysis have been proposed to accommodate covariate information when evaluating test accuracy, including covariate-adjusted and covariate-specific performance estimates.
- Diagnostic accuracy studies should account for important interactions between test performance and covariates.

which converts a continuous classifier into a dichotomous one. Oftentimes, a desired level of classification is specified, to maximize the true positive or true negative results, and identify the corresponding threshold [2].

The result of a test can be associated with other factors than the presence or absence of the target condition. For instance, older patients tend to have higher D-dimer values than younger ones, while males have higher hemoglobin levels than females. In the presence of such associations, there may also be covariate-specific (such as age or sex) differences in test performance. Moreover, selecting thresholds from a standard ROC curve can be misleading, as compared to subgroup specific ROC curves, when strong associations between the marker and covariate are present, and result in differences in sensitivity and specificity between covariate subgroups. Thus, when covariate information is available, it should be considered, as neglecting such information may inflate our estimates of the relative proportion of false negative or false positive test results for certain subgroups [3].

In light of this, several methods that account for covariates in ROC curve analysis have been proposed [4,5]. They allow assessment of covariate-specific and covariate-adjusted ROC curves; the former models ROC curves for each stratum of a given covariate (e.g., men and women), while the other models a single ROC curve that can be interpreted as the weighted average of covariate-specific curves [6]. Despite the widespread consideration of covariate effects in randomized trials of interventions, it is not yet standard practice in diagnostic accuracy studies [7,8].

Associations between covariates and the positivity threshold are even less considered, when, in fact, it has direct implications for how the test will be implemented for practical use. Understanding the magnitude of potential covariate effects and applying appropriate techniques are therefore fundamental to produce robust and reliable results that can be translated into clinical practice.

We here present an illustrative application of the use of covariate-specific and covariate-adjusted ROC analyses, to encourage a more widespread application of such methods in evaluations of diagnostic accuracy. The following sections are structured as follows: the motivating example, an outline of conventional ROC curve analysis and possible scenarios when considering covariates, the application, and concluding remarks.

2. Motivating example

Pulmonary embolism (PE) is a common venous thromboembolic disease that can cause significant morbidity and mortality [9,10]. Patients with suspected venous thromboembolism (VTE), comprised of PE and deep vein thrombosis, usually undergo imaging testing, such as compression ultrasonography or computed tomography pulmonary angiography (CTPA), for a confirmation or exclusion of diagnosis. However, signs and symptoms indicating PE are nonspecific, and therefore PE is not confirmed in many patients with the suspected disease. Considering the additional risks and costs of performing CTPA, scoring systems and tests have been proposed to indicate those at greater risk.

Diagnostic clinical scores comprised of clinical characteristics, such as the Wells score, have been developed to classify patients with suspected PE into pretest probability, and ultimately minimize the number of patients subjected to CTPA testing [11,12]. More recently, the YEARS algorithm was proposed, consisting of only three components, offering a more simplified decision rule [11].

D-dimer is a sensitive plasma marker of endogenous fibrinolysis that appears following blood clot degradation [13]. Measuring levels of this degradation product is commonly used as a diagnostic test in patients with signs

Table 1. Covariate-specific performance of D-dimer and prevalence in corresponding subgroups

Covariate	Proportion with VTE	AUC (95% CI)	Sensitivity = 0.98	
			Threshold[a] (95% CI)	Specificity (95% CI)
Standard	15%	0.87 (0.86, 0.88)	470	0.47
Age				
<50 yr	9%	0.88 (0.87, 0.89)	114 (0, 205)	0.13 (0.03, 0.26)
≥50 yr	18%	0.84 (0.84, 0.85)	401 (346, 452)	0.28 (0.23, 0.33)
Sex				
Male	18%	0.86 (0.85, 0.87)	351 (286, 413)	0.35 (0.27, 0.41)
Female	13%	0.87 (0.86, 0.87)	297 (235, 357)	0.29 (0.21, 0.37)
YEARS				
YEARS = 0	8%	0.87 (0.86, 0.88)	311 (228, 392)	0.31 (0.18, 0.43)
YEARS ≥ 1	21%	0.85 (0.85, 0.86)	327 (269, 385)	0.32 (0.27, 0.36)

Venous thromboembolism (VTE), area under the receiver operating characteristic curve (AUC) and 95% confidence interval (95% CI).
YEARS algorithm components: clinical signs or symptoms of deep vein thrombosis, haemoptysis, pulmonary embolism likely diagnosis.
[a] D-dimer thresholds expressed in ng/mL.

and symptoms suggestive of venous thromboembolism. A threshold of 500 ng/mL was initially proposed for D-dimer to rule out VTE in patients with non–high pretest probability. A more recent study factored the patient's pretest probability and proposed an additional upper threshold of 1,000 ng/mL for those without any YEARS items to increase the proportion of patients in whom imaging can be withheld [14]. Yet the optimal approach for adjusting d-dimer thresholds has still to be determined [15].

Other factors have shown to influence D-dimer concentration. Age, for example, is associated with D-dimer positivity [16]. The D-dimer concentration naturally increases with age, leading to many older patients without PE presenting with D-dimer levels above the conventional threshold of 500 ng/mL [17]. When D-dimer testing is performed among elderly, the proportion of false-positive results is higher leading to unnecessary imaging [18,19]. Age-dependent threshold values for D-dimer were proposed and its diagnostic performance has been compared to the conventional threshold [20,21]. The age-adjusted D-dimer threshold was defined as age(years)*10 ng/mL for patients aged over 50 years, based on evaluating optimal values for 10-year interval age groups. Studies have also shown the influence of other factors, such as setting (inpatient/outpatient) and cancer status [22,23].

2.1. Individual patient data cohort

We consider data from a large individual patient data (IPD) meta-analysis of studies assessing the accuracy of clinical decision rules and D-dimer testing for detection of VTE among patients with suspected PE [24]. In the IPD cohort, data from 21,621 patients, from 16 studies recruited between 1990 and 2020, were included in the analysis. In this cohort, 15% was diagnosed with PE. PE diagnosis was objectively confirmed with either CTPA or clinical follow-up of at least 1 month in those without initial anticoagulation treatment upon initial testing. The characteristics of the IPD cohort are described in Supplementary Table 2.

3. Receiver operating characteristic (ROC) curve analysis

3.1. Conventional ROC curve analysis

In a diagnostic accuracy study, ROC curves can be constructed where the results of one or more index tests are compared against the results of the clinical reference standard, the best available test to evaluate the presence or absence of the target condition [25].

3.2. Positivity threshold

If a positivity threshold is defined, the diagnostic accuracy of an index test can be expressed by estimates of its sensitivity and specificity. If higher index test results make the target condition more likely, sensitivity corresponds to the proportion of those with a target condition whose test result exceeds the positivity threshold. Analogously, the specificity refers to the proportion of those without the target condition whose test result does not exceed the positivity threshold.

If no positivity threshold can be defined, or none was defined a priori, one can consider the full ROC curve. The y-axis of the ROC curve displays all possible values of the sensitivity (or true positive fraction, TPF). The x-axis displays all possible values of the specificity, from right to left, or of the false positive fraction (FPF, one minus specificity), from left to right.

The ROC curve links the TPF and FPF; it is based on the survival function (one minus the cumulative distribution function) of the test results in the subgroup with the target condition, as indicated by the reference standard, and links this to the survival function of the test results in the

Sensitivity = 0.95		Sensitivity = 0.90	
Threshold[a] (95% CI)	Specificity (95% CI)	Threshold (95% CI)	Specificity (95% CI)
686	0.61	910	0.70
402 (325, 473)	0.57 (0.47, 0.65)	662 (596, 725)	0.75 (0.72, 0.78)
655 (610, 698)	0.47 (0.44, 0.50)	893 (852, 934)	0.60 (0.58, 0.62)
609 (555, 661)	0.54 (0.51, 0.56)	850 (799, 901)	0.65 (0.62, 0.67)
560 (507, 610)	0.55 (0.51, 0.58)	803 (755, 850)	0.67 (0.65, 0.69)
575 (510, 638)	0.58 (0.54, 0.62)	818 (757, 877)	0.69 (0.66, 0.72)
582 (536, 627)	0.50 (0.47, 0.53)	825 (782, 868)	0.64 (0.61, 0.66)

subgroup without the target condition. The area under the ROC curve (AUC, also known as AUROC) takes values between zero and one, where one indicates perfect performance and 0.5 refers to performance no better than flipping a coin.

3.3. ROC curve analysis incorporating covariates

In most diagnostic accuracy studies, all available index test and reference standard results are used to construct ROC curves. No other patient or study characteristics are considered as covariates. By now it is well known that diagnostic accuracy is not a fixed property of a test and that it can vary between population subgroups, test types, settings, and depending on the position of the test in the clinical pathway [26,27].

3.3.1. Covariate-specific ROC curve scenarios and implications

If the covariate can be indicated by one dichotomous variable, the investigators can create two subgroups and correspondingly create two different ROC curves. In line with terminology in the statistical literature, we will refer to these as covariate-specific ROC curves.

Three scenarios can be drawn based on a comparison of these two ROC curves, as well as conclusions regarding the standard ROC curve. We illustrate the scenarios using sex as the covariate of interest.

3.3.1.1. Scenario 1. Identical distribution, identical performance. It is possible that the covariate-specific ROC curves are completely identical. That would be the case, for example, if the underlying distributions of test results in those with and those without the target condition are identical in men and women. Each positivity threshold would then yield the same sensitivity and specificity in women as in men. The AUC would be the same in men and in women.

3.3.1.2. Implication. If no difference exists and the covariate-specific ROC curves are identical, the standard AUC expresses performance well, since the covariate-specific AUC are one and the same.

3.3.1.3. Scenario 2. Different distribution, identical performance. In a different scenario, the distribution of test results differs between men and women. Again, as an example, men may have higher values, on average, than women, both in those with and in those without the target condition. In that case, a single positivity threshold would yield a different sensitivity in men compared to women, and a different specificity. Sensitivity will be higher in men but specificity lower.

It is still possible that the two covariate-specific ROC curves are identical. For example, if the distributions of those with and without the target condition have the same difference in means between men and women, without any differences in variance, then the two covariate-specific curves will be the same, as well as the AUC. Overall performance, as expressed by the AUC, will be the same in men and women, but different positivity thresholds must be selected to yield the same sensitivity and specificity.

3.3.1.4. Implication. As demonstrated by Janes and Pepe (2008), if a difference in distributions exists but the covariate-specific ROC curves are identical, the standard AUC can present a biased upward estimate of test performance [3]. This will be the case if one subgroup, say men, is more likely to have the target condition. The standard ROC curve will also capture that additional difference between men and women and will lie above the covariate-specific ROC curve. The standard AUC, though correctly

estimated, will then also show upward bias, since it does not only express the performance of the test but is also based on the pre-existing difference in prevalence between men and women.

If, in an alternative scenario, the prevalence between men and women is the same, the standard ROC curve will be attenuated: it will lie below the covariate-specific ROC curves. The standard AUC will not express performance well. The identical covariate-specific AUC, based on thresholds that differ between men and women, will be higher: it reflects the gain in performance that is possible from using such stratified positivity thresholds.

3.3.1.5. Scenario 3. Different distribution, different performance. In a third scenario, the distributions of the test results in those with and without the target condition differ in such a way that the covariate-specific ROC curves are no longer identical. That can happen in diverse ways. It is possible that men without the target condition have the same distribution as women without the target condition but, with the target condition, men have much higher values than women. If so, overall performance will also be different. Depending on the distributions, a single positivity threshold may also lead to different values for sensitivity and specificity in the subgroups.

3.3.1.6. Implication. If the covariate-specific ROC curves and AUC differ, the standard AUC is not a fair representation of performance, as it ignores the potentially meaningful differences between the subgroups. Presenting the significantly different covariate-specific ROC curves and the corresponding AUC may be more informative for clinical decision-making.

3.4. Bayesian nonparametric model

All of the performed analyses were based on the Bayesian nonparametric approach proposed by In acio de Carvalho et al. [28]. This approach incorporates covariate information by using a single-weights dependent Dirichlet process mixture of normal distributions. Specifically, the model includes a mixture of normal distributions with means that follow a regression model, which may be linear or nonlinear, dependent on the covariate(s) [29]. This allows for the construction of covariate-specific ROC curves, specified for the conditional CDF which changes as a function of the covariate, as opposed to just considering the mean or variance of the distribution, as in other semiparametric approaches [30].

3.5. Statistical analysis

Cumulative distribution function (CDF) plots and histograms were created for covariate subgroups to explore the distribution and density of index test results among the diseased and nondiseased.

The standard empirical ROC curve was constructed without incorporating covariate information [7]. We utilized the Bayesian nonparametric approach to construct covariate-specific ROC curves [28], including ordinal and continuous covariates which were dichotomized, where necessary, into clinically relevant categories. For each Bayesian nonparametric model, we estimated the densities and distribution by disease status. In addition, we also constructed covariate-adjusted ROC curves, initially developed by Janes and Pepe [31], but adapted to the Bayesian nonparametric approach. Here we included covariates without categorization. Diagnostic accuracy was expressed as the AUC with its 95% confidence interval (95% CI). For each individual ROC curve, positivity thresholds corresponding to a sensitivity of 0.98, 0.95, 0.90 were identified.

All statistical analyses were performed using R software version 4.0.3, using the ROCnReg package [32]. For detailed introduction and illustration of various frameworks for covariate consideration in ROC curve analysis, we refer to the ROCnReg guidance document [32].

4. Application

4.1. Subgroup differences

We conducted a series of exploratory analyses to landscape the distribution of index test values across covariate subgroups and in the diseased and nondiseased subgroups. There were differences in D-dimer concentration between age groups, more prominent in the nondiseased group, with much wider dispersion among the diseased. Differences were less pronounced for other covariates (Supplementary Fig. 1).

Overall, there was unanimous right-skewed distribution of test results. We further visually confirmed largely overlapping distributions of test results between sex subgroups, with differences in frequency of lower test results among the nondiseased for some covariates (Supplementary Fig. 1). The PE prevalence varied between some of the covariate subgroups, for example those based on age and on the presence of YEARS items (Table 1).

4.2. Performance estimates with conventional ROC approach

We first constructed the standard ROC curve to evaluate performance without incorporating any covariate information. Using an empirical estimator, we found D-dimer had an AUC of 0.87 (95% CI: 0.86, 0.88) in detecting VTE. This was considered as the benchmark performance indicator.

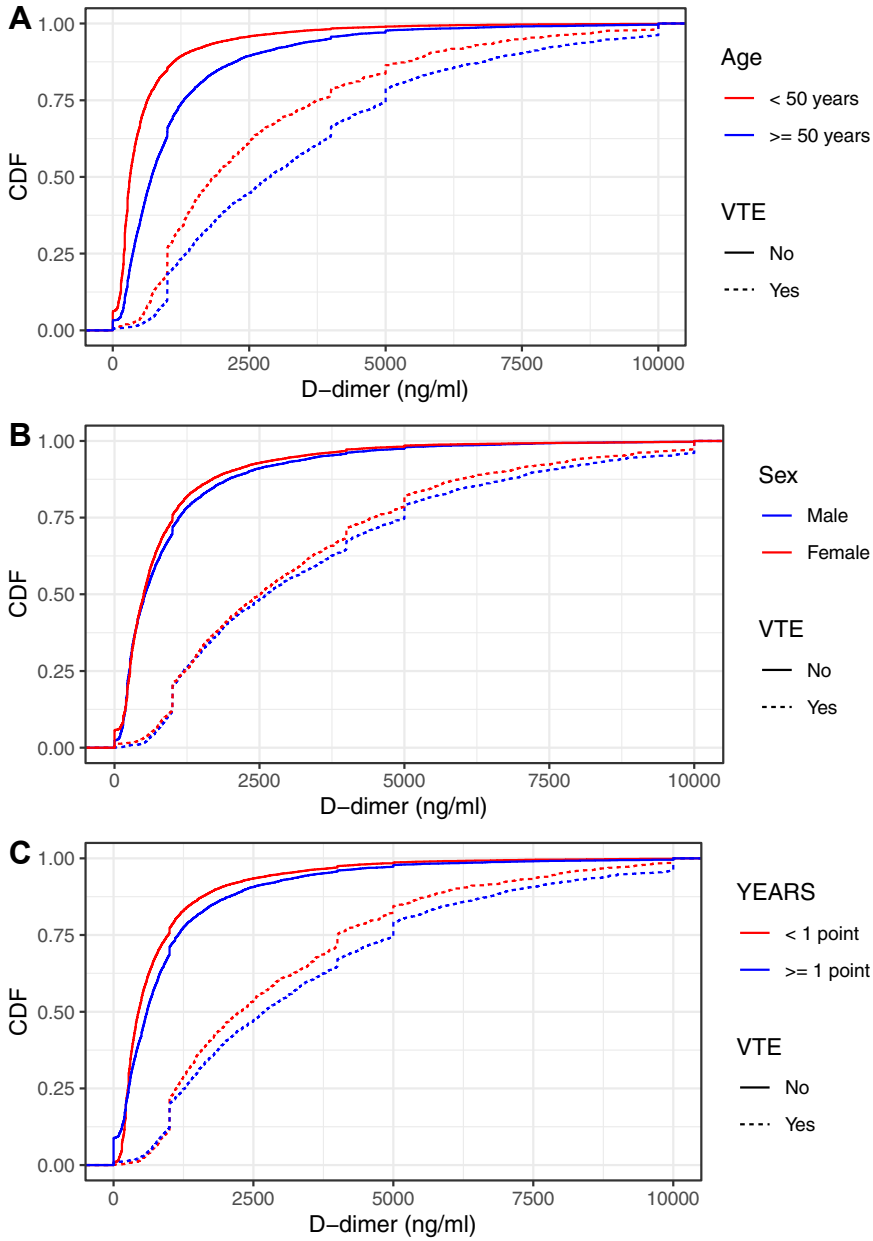

Fig. 1. Cumulative distribution function (CDF) plots by covariate [(A). Age, (B). Sex, and (C). YEARS algorithm] and venous thromboembolism (VTE) status. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article).

4.3. Performance estimates with covariate-specific ROC curve analysis

When constructing covariate-specific ROC curves, our interest was in evaluating whether the discriminatory capacity of the index test varies between covariate subgroups. In our case, we were interested in the possible effect of age, sex, and pretest probability by use of the YEARS algorithm on the performance of the index test, D-dimer.

Performance of the index test varied significantly between age groups (Table 1). In older patients the AUC was lower (0.84 [0.84, 0.85]) than in the younger group (AUC of 0.88 [0.87, 0.89]). We saw a noticeable gap in the CDF between the age groups (Fig. 1). Younger patients tend to have lower index test results (Supplementary Fig. 1). There are also differences in the proportion with and without VTE in the two subgroups. This is an example of Scenario 3 (different distribution, different performance). Providing only the standard ROC curve analysis is not a fair representation of performance, as it ignores meaningful differences between age subgroups, in this case compromised performance in older patients.

In contrast, the covariate-specific ROC curves and AUCs were nearly identical between men and women, consistent with the similar distribution of test results in subgroups by sex (Supplementary Fig. 1), and nearly overlapping CDF

(Fig. 1). We can assume that there are no meaningful differences based on sex, an example of Scenario 1 (identical distribution, identical performance). In this case, we can conclude the standard AUC fairly expresses the performance.

Patients with and without items of the YEARS had similar performance (Table 1), with some differences in the distribution of index test results (Fig. 1). However, differences in the distribution are less noticeable in the lower ranges of the index test values (Supplementary Fig. 1, F). This resembles Scenario 2 (different distribution, identical performance). The prevalence of the target condition differs drastically between the subgroups (8% vs. 21%). The standard ROC curve analysis in this case may therefore be biased, as it does not consider underlying differences."

The former analyses considered the effect of a single covariate. It is also possible that there is an interaction between a pair of covariates, such as age and sex. By specifying the parameters of the regression model to include both covariates, we can model their effect on performance. In Supplementary Figure 2 we can see that discriminatory capacity of the index test is slightly lower in younger men and younger women, with no pronounced differences by sex. In cases where the effect is unclear, further significance testing should be performed [33].

4.4. Selection of covariate-specific positivity thresholds

We also compared standard vs. covariate-specific threshold values for desired performance levels (Table 1). We found different thresholds were necessary to achieve the same level of performance. Taking age as an example, there is a difference of nearly 250 ng/mL in thresholds for younger vs. older patients to achieve a sensitivity of 0.95. Thus, higher positivity thresholds for D-dimer have to be selected for elderly patients, visually illustrated in Figure 2. This is consistent with the understanding that D-dimer levels increase with age. In such settings, recall Scenario 3, the covariate-specific ROC curves are different, and covariate-specific positivity thresholds should, ideally, be used.

Differences were less prominent for other covariates. Looking at each point of the YEARS algorithm (scale of 0 to 3), the positivity thresholds are nearly identical, meaning the standard threshold would achieve the same sensitivity in both groups (Table 1). As mentioned previously, this may be explained by the similarities in distribution in the lower ranges, which corresponds to high sensitivity. In Figure 2 we can see that the same threshold would apply to any point on the YEARS algorithm to meet the desired sensitivity level. As no meaningful differences are observed, we can conclude that the ROC curves are identical.

4.5. Performance estimates with covariate-adjusted ROC curve analysis

In some cases, it may be informative to also present a covariate-adjusted ROC curve, one that takes covariate information into account: a weighted average of the covariate-specific ROC curves, with weights corresponding to the proportion of those with the target condition in the

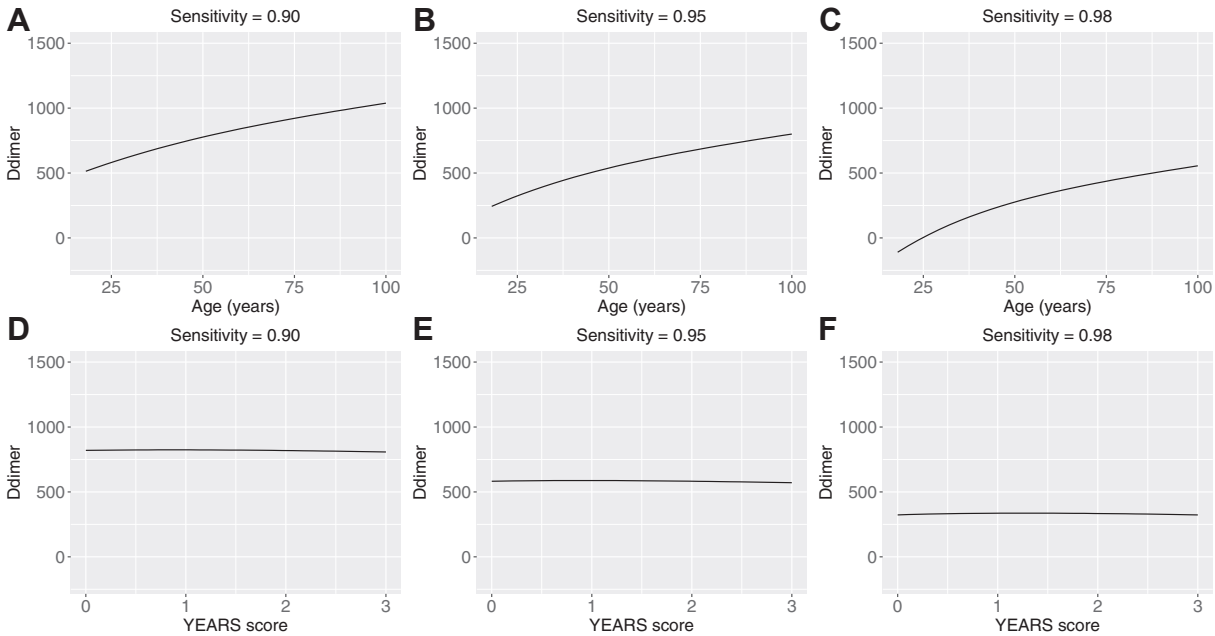

Fig. 2. Threshold values for D-dimer, along age (A, B, C) and the YEARS score (D, E, F), modeling using the Bayesian nonparametric approach. Posterior mean (solid black line) and 95% pointwise credible band for D-dimer thresholds, corresponding to sensitivity of 0.98, 0.95 and 0.90.

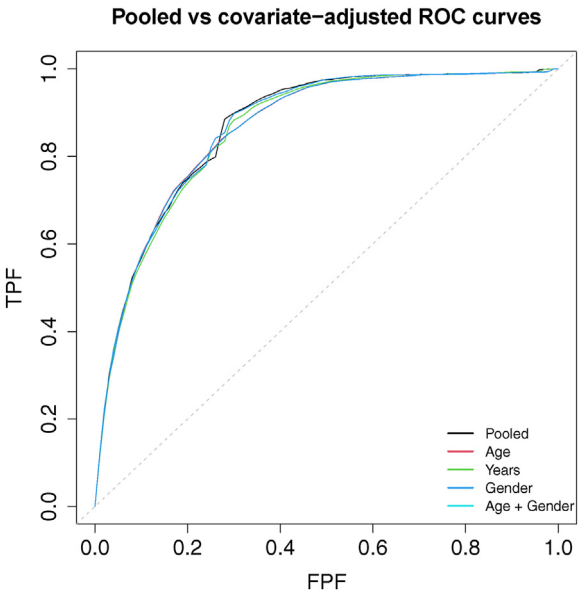

Fig. 3. Standard (pooled) vs. covariate-adjusted ROC curves. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article).

two subgroups. We constructed ROC curves that were adjusted for age, sex, the YEARS algorithm, and the combination of two covariates. The covariate-adjusted AUCs were almost identical to each other and reflected the standard pooled AUC of 0.87 (see Fig. 3).

5. Discussion

Most diagnostic accuracy studies that present ROC curves to summarize test performance ignore covariates. Yet constructing covariate-specific ROC curves can be informative for understanding the relationship between covariates and a test's performance and positivity threshold. There may be stratum-specific differences in performance that can influence further clinical decision making or, in the absence of any meaningful differences, we may conclude that the standard ROC curve produces fair estimates of performance.

Adjusting for covariate effects would be most necessary when comparing the performance of different tests, to alleviate any bias that may arise from unfair representation of patient characteristics where the performance may vary, as illustrated by Pepe et al. [3]. We can also assume multicenter studies, with intrinsically different test settings, may benefit from adjusting for covariates. Yet comparisons between other techniques such as multilevel analysis or other prosed mixed methods for handling issues related to multicenter data are less established. Importantly, this is an area that deserves more attention, particularly in the presence of clustered data. Covariate adjustment may be preferred with smaller sample sizes, which may be problematic for covariate-specific analysis. Covariate-adjusted ROC analyses can also consider continuous variables, in addition to categorical and binary ones.

Our study presents some limitations. In our IPD cohort, verification of the outcome was not the same for all patients in most studies and we relied on multiple reference standards. Imaging was performed for those with high clinical suspicion and/or high D-dimer, and clinical follow-up for those with low D-dimer levels.

In meta-analyses of diagnostic accuracy studies, heterogeneity in test performance is common across the primary studies. This may be due to patient or test characteristics and thus present a genuine difference in performance based on biology, but it may also be artifactual and due to study design flaws. Such artifactual factors can be identified with the QUADAS-2 tool [34]. In the motivating example, we selected covariates based on a biologic basis, however, we can also utilize study design characteristics as covariates. Various approaches for including such covariates in meta-analysis of ROC curves have been proposed [35]. Inclusion of artifactual covariates in conditional ROC curve analysis can also be incorporated to reflect, for example, center differences in multicenter diagnostic accuracy studies. This application warrants further exploration, as adjustment for center differences is another element of diagnostic accuracy research that remains less established.

We here presented results using a Bayesian nonparametric approach. Other methods, such as a semiparametric approach and a nonparametric kernel-based regression model, have been proposed [33]. The Bayesian nonparametric approach is flexible to various distribution features, as it can adapt to skewness, nonlinearities, or data with higher variability. This makes it a practical choice for use with many different diseases and populations. The computational demand is however greater with this approach compared to some other models. We further note the limitations of methods such as the kernel-based regression models, which have long computation times and more limitations regarding number and type of covariates that can be included in the model. For an in-depth review, including an overview of various proposed statistical concepts and their application can be found in the work by In acio and Rodríguez-Alvarez et al. and related publications [4,5].

Incorporating covariate information into ROC curve analysis is not yet common practice, despite methods that have been proposed decades ago. We hope that the analysis presented here will lead to a more widespread application of such conditional ROC curves, which can provide more robust information on test performance and may improve our ability to select thresholds catered for specific subgroups.

Declaration of Competing Interest

F.A.K. reports grants or contracts from Bayer, BMS, BSCI, MSD, Leo Pharma, Actelion, Pharm-X, The Netherlands Organisation for Health Research and Development, The Dutch Thrombosis Association, The Dutch Heart Foundation and the Horizon Europe Program, all unrelated to this work and paid to his institution.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2023.06.001.

References

[1] Obuchowski NA, Bullen JA. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. Phys Med Biol 2018;63(7):07TR1.

[2] Pepe MS, Janes H, Li CI, Bossuyt PM, Feng Z, Hilden J. Early-phase studies of biomarkers: what target sensitivity and specificity values might confer clinical utility? Clin Chem 2016;62:737−42.

[3] Janes H, Pepe MS. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. Am J Epidemiol 2008;168:89−97.

[4] Inácio V, Rodríguez-Álvarez MX. The covariate-adjusted ROC curve: the concept and its importance, review of inferential methods, and a new Bayesian estimator. Stat Sci 2022;37(4):541−61.

[5] Inácio V, Rodríguez-Álvarez MX, Gayoso-Diz P. Statistical evaluation of medical tests. Ann Rev Stati Appl 2021;8:41−67.

[6] Rodríguez-Álvarez MX, Inacio V. ROCnReg: An R Package for Receiver Operating Characteristic Curve Inference with and without Covariate Information 2020: arXiv preprint arXiv:200313111.

[7] Metz CE. Basic principles of ROC analysis. Sem Nucl Med 1978; 8(9):283−98.

[8] Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561−77.

[9] Martin KA, Molsberry R, Cuttica MJ, Desai KR, Schimmel DR, Khan SS. Time trends in pulmonary embolism mortality rates in the United States, 1999 to 2018. J Am Heart Assoc 2020;9(17): e016784.

[10] Barco S, Mahmoudpour SH, Valerio L, Klok FA, Münzel T, Middeldorp S, et al. Trends in mortality related to pulmonary embolism in the European Region, 2000−15: analysis of vital registration data from the WHO Mortality Database. Lancet Respir Med 2020; 8(3):277−87.

[11] 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. Lancet 2017;390:289−97.

[12] Wells PS, Ihaddadene R, Reilly A, Forgie MA. Diagnosis of venous thromboembolism: 20 years of progress. Ann Intern Med 2018;168: 131−40.

[13] Weitz JI, Fredenburgh JC, Eikelboom JW. A test in context: D-dimer. J Am Coll Cardiol 2017;70:2411−20.

[14] Van Es J, Beenen L, Douma R, Den Exter P, Mos I, Kaasjager H, et al. A simple decision rule including D-dimer to reduce the need for computed tomography scanning in patients with suspected pulmonary embolism. J Thromb Haemost 2015;13:1428−35.

[15] Konstantinides SV, Meyer G, Becattini C, Bueno H, Geersing G-J, Harjola V-P, et al. 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS) the Task Force for the diagnosis and management of acute pulmonary embolism of the European Society of Cardiology (ESC). Eur Heart J 2020;41: 543−603.

[16] Kabrhel C, Mark Courtney D, Camargo CA Jr, Plewa MC, Nordenholz KE, Moore CL, et al. Factors associated with positive D-dimer results in patients evaluated for pulmonary embolism. Acad Emerg Med 2010;17(6):589−97.

[17] Harper P, Theakston E, Ahmed J, Ockelford P. D-dimer concentration increases with age reducing the clinical value of the D-dimer assay in the elderly. Intern Med J 2007;37(9):607−13.

[18] Righini M, Le Gal G, Perrier A, Bounameaux H. The challenge of diagnosing pulmonary embolism in elderly patients: influence of age on commonly used diagnostic tests and strategies. J Am Geriatr Soc 2005;53(6):1039−45.

[19] Schutgens RE, Haas FJ, Biesma DH. Reduced efficacy of clinical probability score and D-dimer assay in elderly subjects suspected of having deep vein thrombosis. Br J Haematol 2005;129(5): 653−7.

[20] Schouten HJ, Geersing GJ, Koek HL, Zuithoff NPA, Janssen KJM, Douma RA, et al. Diagnostic accuracy of conventional or age adjusted D-dimer cut-off values in older patients with suspected venous thromboembolism: systematic review and meta-analysis. BMJ 2013;346:f2492.

[21] Douma RA, Le Gal G, Söhne M, Righini M, Kamphuisen PW, Perrier A, et al. Potential of an age adjusted D-dimer cut-off value to improve the exclusion of pulmonary embolism in older patients: a retrospective analysis of three large cohorts. BMJ 2010;340: c1475.

[22] Den Ouden M, Ubachs JH, Stoot J, Van Wersch J. Thrombin-antithrombin III and D-dimer plasma levels in patients with benign or malignant ovarian tumours. Scand J Clin Lab Invest 1998;58(7): 555−60.

[23] Douma RA, van Sluis GL, Kamphuisen PW, Söhne M, Leebeek FW, Bossuyt PM, et al. Clinical decision rule and D-dimer have lower clinical utility to exclude pulmonary embolism in cancer patients. Explanations and potential ameliorations. Thromb Haemost 2010; 104(4):831−6.

[24] Geersing GJ, Kraaijpoel N, Büller HR, van Doorn S, van Es N, Le Gal G, et al. Ruling out pulmonary embolism across different subgroups of patients and healthcare settings: protocol for a systematic review and individual patient data meta-analysis (IPDMA). Diagn Progn Res 2018;2:10.

[25] Bossuyt PM, Olsen M, Hyde C, Cohen JF. An analysis reveals differences between pragmatic and explanatory diagnostic accuracy studies. J Clin Epidemiol 2020;117:29−35.

[26] Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ 2002;324:669−71.

[27] Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6(11): e012799.

[28] de Carvalho VI, Jara A, Hanson TE, de Carvalho M. Bayesian nonparametric ROC regression modeling. Bayesian Anal 2013;8(3): 623−46.

[29] De Iorio M, Johnson WO, Müller P, Rosner GL. Bayesian nonparametric nonproportional hazards survival modeling. Biometrics 2009;65:762−71.

[30] Janes H, Pepe MS. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. Biometrika 2009;96(2):371−82.

[31] Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. Biometrics 1998;54:124−35.

[32] Rodríguez-Álvarez MX, Vanda I. ROCnReg: an R package for receiver operating characteristic curve inference with and without covariate information. 2020. arXiv preprint arXiv:2003.13111.

[33] Rodríguez-Álvarez MX, Roca-Pardiñas J, Cadarso-Suárez C. ROC curve and covariates: extending induced methodology to the non-parametric framework. Stat Comput 2011;21(4):483−99.

[34] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529−36.

[35] Doebler P, Holling H. Meta-analysis of diagnostic accuracy and ROC curves with covariate adjusted semiparametric mixtures. Psychometrika 2015;80(4):1084−104.