Clinical Chemistry 58:10
1408–1417 (2012)

Review

Quantifying the Added Value of a Diagnostic Test or Marker

Karel G.M. Moons,[1*†] Joris A.H. de Groot,[1†] Kristian Linnet,[2] Johannes B. Reitsma,[1] and Patrick M.M. Bossuyt[3]

In practice, the diagnostic workup usually starts with a patient with particular symptoms or signs, who is suspected of having a particular target disease. In a sequence of steps, an array of diagnostic information is commonly documented. The diagnostic information conveyed by different results from patient history, physical examination, and subsequent testing is to varying extents overlapping and thus mutually dependent. This implies that the diagnostic potential of a test or biomarker is conditional on the information obtained from previous tests. A key question about the accuracy of a diagnostic test/biomarker is whether that test improves the diagnostic workup beyond already available diagnostic test results. This second report in a series of 4 gives an overview of several methods to quantify the added value of a new diagnostic test or biomarker, including the area under the ROC curve, net reclassification improvement, integrated discrimination improvement, predictiveness curve, and decision curve analysis. Each of these methods is illustrated with the use of empirical data. We reiterate that reporting on the relative increase in discrimination and disease classification is relevant to obtain insight into the incremental value of a diagnostic test or biomarker. We also recommend the use of decision-analytic measures to express the accuracy of an entire diagnostic workup in an informative way.

The diagnostic workup in practice starts with a patient presenting with particular symptoms or signs. Although these symptoms and signs usually raise the suspicion of several underlying disorders (the differential diagnosis), the diagnostic workup is often initially targeted to include or exclude a particular disease or disorder on this list of differential diagnoses, the so-called target disease *(1–5)*. For example, a man with a red, swollen leg may be suspected of having deep vein thrombosis, a woman with a palpable breast node can be suspected of having breast cancer, or a child with neck stiffness and fever suspected of having bacterial meningitis. This target disorder can be the most severe disorder of the differential diagnoses ('the one not to miss') but also the most likely one.

The diagnostic workup usually comprises a series of sequential steps in which an array of diagnostic information (test results) is obtained. In principle, after each step the physician intuitively integrates the information into a judgment regarding the probability of the target disease being present, and perhaps even of the other conditions on the differential diagnosis list. The initial pieces of information always include patient history and physical examination results, to varying extents. If uncertainty about disease presence remains, as is commonly the case, subsequent tests are performed, often in a stepwise fashion. These additional tests can range from simple blood or urine tests to imaging, electrophysiology, and genetic tests, to eventually more invasive testing such as biopsy, angiography, or arthroscopy. The information of each subsequent test is implicitly added to the previously obtained information, and the target disease probability is adjusted. This process is continued until a final diagnosis can be set and a treatment choice can be made. Because information from history taking and physical examination is almost always obtained in diagnostic practice, hardly any clinical diagnosis is exclusively based on a single test or technology result; rather, diagnostic practice seems to involve a multivariable (multiple-test) and phased process per se, in which care providers have to decide whether the next test will add information to what is already known *(6–8)*.

Studies of diagnostic tests or biomarkers should reflect this phased diagnostic process and be aimed at understanding the added value of subsequent diagnostic tests beyond the information that is already available. It may turn out that the diagnostic information of a subsequent test or biomarker is already conveyed by the simpler previous test results. When considered by

[1] Julius Center for Health Sciences and Primary Care, UMC Utrecht, The Netherlands; [2] Section of Forensic Chemistry, Dept. of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Denmark; [3] Department Clinical Epidemiology, Biostatistics & Bioinformatics, Academic Medical Center, University of Amsterdam, the Netherlands.
* Address correspondence to this author at: Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, P.O. Box 85500, 3508 GA Utrecht, the Netherlands. Fax +31-88-7555485; e-mail k.g.m.moons@umcutrecht.nl.
† Karel Moons and Joris de Groot contributed equally to the work, and both should be considered as first authors.
Received January 17, 2012; accepted July 11, 2012.
Previously published online at DOI: 10.1373/clinchem.2012.182550

itself the subsequent test/biomarker may show diagnostic value, but when evaluated in the context of the overall diagnostic workup it does not. Such a situation can occur because different tests often measure the same underlying pathologic process to varying extents and thus provide similar diagnostic information. In statistical terms, these different test results, whether obtained from patient history, physical examination, or subsequent tests, are to varying extents mutually dependent *(9–11)*. The main question in diagnostic accuracy research seems, therefore, to be not what the diagnostic accuracy of a particular (new) test or biomarker is, but rather, whether it improves the diagnostic accuracy of the existing workup beyond that available from the already documented and standard available diagnostic information.

This second report of the series gives an overview of available methods to quantify what value a specific (new) diagnostic test or biomarker adds to that available from current biomarkers, for which the aim is not necessarily to replace the previous tests or biomarkers but rather to complement them. We focus on assessing whether a certain test adds information to preceding test results, in terms of improved discrimination between disease presence vs absence, but still in comparison to a clinical reference standard.

Empirical Example: Diagnosis of Suspected Deep Venous Thrombosis

To illustrate the methods to assess the added value of a subsequent diagnostic test we use the same data as used in the first report of this series, i.e., the data from a previously published study in primary care patients suspected for deep venous thrombosis (DVT)[4] *(12, 13)*. In brief, there were 2086 patients suspected of DVT, defined as having at least one of the following symptoms: presence of swelling, redness, and/or pain in the leg. All patients underwent a standardized diagnostic workup including index tests from medical history taking, physical examination, and D-dimer testing.

The reference standard consisted of repeated compression ultrasonography, in line with current clinical practice, where this test is used to decide on further management. This reference test was performed in all patients independent of the results of the index tests and blinded to these index test results. In total, 416 of the 2068 included patients (20%) had DVT present on ultrasonography. We would point out that for this report these data are used for illustration purposes only, and by no means to define the best diagnostic workup for the clinical problem at hand or to compare our results with existing reports on the topic. Our aim here is to illustrate how to quantify the extent to which D-dimer biomarker values provide added value to the diagnostic test results (variables) from history taking and physical examination in the correct discrimination between DVT presence (requiring further workup including reference test) or absence (no further workup needed). We have used only a subset of the originally documented diagnostic variables (Table 1). In addition, although the reference test is applied in current routine practice, it may not be perfect. The potential consequences of using imperfect reference tests and how to handle this problem are highly important, but these topics are beyond the scope of this report and have been covered elsewhere *(14–18)*.

Table 1 also shows the association between each individual diagnostic test result and the presence or absence of DVT. These are single-test diagnostic accuracy measures. If one had to choose the most promising tests on the basis of these single-test accuracy results, it would be difficult if not impossible. Clearly, none of the history and physical examination tests was pathognomonic. Some variables had a high sensitivity but a low specificity (such as absence of leg trauma and pain on walking), whereas other tests showed a high specificity and low sensitivity (e.g., presence of malignancy or recent surgery). Some tests were better for exclusion, others for inclusion. The ROC area of the continuous tests, age and calf difference (but also the D-dimer test), was below 1. So the question can be raised whether the combination of the history and physical examination variables has improved accuracy compared to their individual accuracy estimates, and whether the D-dimer biomarker could provide even further incremental improvement.

To formally quantify the diagnostic value of these patient history and physical examination results combined, and the added value of the D-dimer biomarker, a multivariable statistical analytical approach is needed. As the outcome is dichotomous (DVT presence yes/no), we used multivariable logistic regression modelling. Such models express the probability of DVT (on the logit scale) as a linear function of the respective test results. Table 2 (model 1) presents the results from history and physical examination that were significantly associated with DVT in the multivariable analysis (here defined by a multivariable odds ratio significantly different from 1, with a P value of <0.05). To quantify whether the D-dimer biomarker added diagnostic value beyond these history and physical examination results combined, the basic model 1 was simply extended by

[4] Nonstandard abbreviations: DVT, deep venous thrombosis; AUC, area under the ROC curve; NRI, net reclassification improvement; IDI, integrated discrimination improvement, TN, true negative; FP, false positive; FN, false negative; TP, true positive; CT, computed tomography.

Table 1. Distribution and accuracy of each diagnostic variable compared to the reference standard outcome.[a]

	DVT						
	Yes (n = 416)			No (n = 1670)			
	n	Sensitivity,[b] % (95% CI)	PPV,[c] % (95% CI)	n	Specificity, % (95% CI)	NPV, % (95% CI)	ROC area[d] (95% CI)
Male sex	194	47 (42–51)	25 (22–29)	569	66 (64–68)	83 (81–85)	—
Age, mean (SD), years	62 (17)	—	—	59 (18)	—	—	0.53 (0.50–0.56)
Presence of malignancy	40	10 (7–13)	35 (27–44)	75	96 (94–96)	81 (79–83)	—
Recent surgery	76	18 (15–22)	27 (22–33)	202	88 (86–89)	81 (79–83)	—
Absence of recent leg trauma	47	89 (85–91)	21 (19–23)	297	18 (16–20)	86 (82–90)	—
Vein distension	115	28 (24–32)	28 (24–32)	302	82 (80–84)	82 (80–84)	—
Pain on walking	344	83 (79–86)	21 (19–23)	1325	21 (19–23)	83 (79–86)	—
swelling whole leg	247	59 (55–64)	26 (23–29)	699	58 (56–60)	85 (83–87)	—
Difference in calf circumference, mean (SD), cm	3 (2)	—	—	2 (2)	—	—	0.69 (0.67–0.72)
D-dimer, mean (SD), ng/mL	4549 (2665)	—	—	1424 (1791)	—	—	0.86 (0.84–0.88)

[a] Reference standard outcome: DVT present or absent on the basis of repeated compression ultrasonography.
[b] Sensitivity, the proportion of patients with true DVT determined by the diagnostic test; specificity, the proportion of patients with true absence of DVT determined by the diagnostic test.
[c] PPV, positive predictive value, i.e. the proportion of patients labeled DVT by the diagnostic test with true DVT; NPV, negative predictive value, i.e. the proportion of patients labeled no DVT by the diagnostic test with true absence of DVT.
[d] An ROC area lower than 0.5 means that overall this test result was better for excluding than including DVT presence.

including the D-dimer result, yielding model 2 (Table 2).

After the addition of the D-dimer assay, the regression coefficients of most history and physical tests in model 2 differ from those in model 1: they now reflect the contribution of the corresponding variables, conditional on a specific D-dimer result. This change indicates that the history and physical and the D-dimer results are correlated, and partly conveys the same information regarding the inclusion or exclusion of DVT. The fact that the regression coefficients of most findings are lower can be interpreted as follows: a portion of the information supplied by the history and physical items is now "taken over" by the D-dimer assay. Note that the effect of the variable recent surgery has completely disappeared owing to the addition of the D-dimer biomarker.

Overall Discriminative or Diagnostic Accuracy of a Model: The Area under the ROC Curve

The multivariable diagnostic model, in which the different diagnostic test results are combined, as in models 1 and 2 in Table 2, can be considered as a single test, consisting of a composite of the series of individual tests. A patient's test result by this model is simply the model's calculated probability of DVT presence given the observed test results in that patient (see the footnote to Table 2 on how to calculate this probability of disease presence).

We can use the area under the ROC curve (AUC), or so-called *c*-statistic, to assess the ability of the combination of these test results (i.e., the diagnostic model) to discriminate between patients with and without DVT, similar to the approach explained in report 1 of this series for a single continuous or quantitative diagnostic test or biomarker *(19, 20)*. Here the ROC area represents the proportion out of all possible pairs of a patient with and without DVT in which the patient with DVT has a higher calculated probability than the patient without DVT *(6)*. Fig. 1 gives the ROC curves and areas for models 1 and 2, similar to those presented in report 1 of this series for 2 different quantitative diagnostic tests.

Fig. 1 shows that adding the D-dimer biomarker to model 1 led to an increase in the ROC area from 0.72 to 0.87, a substantial and statistically significant gain (P value <0.01) *(20)*. This means the overall diagnostic accuracy of the information from patient history and physical exam can be substantially and significantly increased by addition of the D-dimer test.

The use of the difference in ROC area to express the added value of a new test/biomarker has been criticized *(21–23)*. First, the AUC is clearly an overall measure of discrimination and has no direct clinical interpretation in terms of correct or incorrect diagnostic

Table 2. The basic and extended multivariable diagnostic model to discriminate between DVT presence vs absence.[a,b]

	Model 1 (basic model)			Model 2 (basic model + D-dimer)		
	Regression coefficient (SE)	OR (95% CI)	*P*	Regression coefficient (SE)	OR (95% CI)	*P*
Intercept	−3.70 (0.26)	—	<0.01	−4.94 (0.32)	—	<0.01
Presence of malignancy	0.62 (0.22)	1.9 (1.2–2.9)	<0.01	0.22 (0.26)	1.2 (0.7–2.1)	0.41
Recent surgery	0.44 (0.16)	1.6 (1.1–2.1)	<0.01	0.003 (0.19)	1.0 (0.7–1.5)	0.99
Absence of leg trauma	0.75 (0.18)	2.1 (1.5–3.0)	<0.01	0.67 (0.20)	2.0 (1.3–2.9)	<0.01
Vein distension	0.48 (0.13)	1.6 (1.1–2.1)	<0.01	0.25 (0.16)	1.3 (0.9–1.8)	0.12
Pain on walking	0.41 (0.15)	1.5 (1.1–2.0)	<0.01	0.46 (0.18)	1.6 (1.1–2.3)	0.01
Swelling whole leg	0.36 (0.12)	1.4 (1.1–1.8)	<0.01	0.47 (0.14)	1.6 (1.2–2.1)	<0.01
Difference in calf circumference, cm	0.36 (0.04)	1.4 (1.3–1.5)	<0.01	0.29 (0.04)	1.3 (1.2–1.4)	<0.01
D-dimer, per 500 ng/mL	NA	NA	NA	0.29 (0.02)	1.3 (1.3–1.4)	<0.01

[a] The exponential of the regression coefficient is known as the odds ratio (OR) of a diagnostic test result. For example, an OR of 2 for absence of leg trauma (model 2) means that a suspected patient without a recent leg trauma has a 2 times higher chance of having DVT than a patient with a recent leg trauma (because in the latter the leg trauma would more likely be the cause of the presenting symptoms and signs). Similarly, an OR of 1.3 for calf difference in centimeters (model 2) means that for every centimeter increase in calf circumference difference, a patient has a 1.3 times (or 30%) higher chance of having DVT.

[b] A diagnostic model can be considered as a single overall or combined test consisting of different test results, with the probability of DVT presence as its test result. For example, for a male patient without malignancy, recent surgery or leg trauma, but with vein distension and a painful not swollen leg when walking, with a calf difference of 6 cm the formula is (model1): $-3.70 + (0.62 \times 0) + (0.44 \times 0) + (0.75 \times 0) + (0.48 \times 1) + (0.41 \times 1) + (0.36 \times 0) + (0.36 \times 6) = -0.65$.

The probability for this patient of the presence of DVT based on the basic model then is $\exp[-0.65)/(1 + \exp(-0.65)] = 34\%$.

classifications or absolute patient numbers, because a specific diagnostic algorithm uses a specific diagnostic cutpoint. Second, on the basis of empirical applications, researchers have observed that the increase in AUC is often very small in an absolute sense, certainly when the AUC of the baseline model is large *(24)*. On one hand this observation is not surprising: good models are harder to improve upon. However, it may not be wise to place too much emphasis on the extent of AUC improvement, because this measure depends on baseline *c*, rather than on the effect size or odds ratio of the new test or marker with the outcome at interest *(22, 23, 25)*. Several alternative measures have been proposed to quantify the added value of a novel test/biomarker to address these limitations.

Reclassification Measures

To overcome the problems of the difference in ROC area, the so-called reclassification table has been proposed *(26)*. This reclassification table shows how many patients are reclassified by adding a new test/biomarker to existing tests (in this case summarized or combined into a multivariable model) after introducing a particular probability of disease presence threshold. The reclassification table for our DVT example is shown in Table 3, with a threshold at 25%. This means that patients with a calculated diagnostic probability of 25% or higher are referred for further workup and perhaps treatment initiation, whereas the others are not referred and treated for DVT.

Table 3 shows the reclassification of patients that occurs when using model 2 instead of model 1 at a DVT probability threshold of 25%. For example, in patients with DVT, 36% (123/416 + 26/416) were reclassified by model 2 compared to model 1. For patients without DVT this percentage was 21% (227/1670 + 116/1670).

Simply indicating the change in classification of individuals to different probability categories of DVT presence, however, is insufficient to properly evaluate improvement in diagnostic accuracy by a new test/biomarker; the changes must also be appropriate. Otherwise, an "upward" movement to higher probability categories for individuals with the DVT present implies improved diagnostic classification, and any downward movement indicates worse diagnostic classification. The interpretation is opposite for individuals without the diagnosis *(27)*. The overall improvement in diagnostic reclassification can be quantified in various ways depending on the chosen denominators, but commonly it is calculated as the difference between 2 differences: First computing the difference between the proportions of individuals moving up and the proportion of individuals moving down for those with DVT,

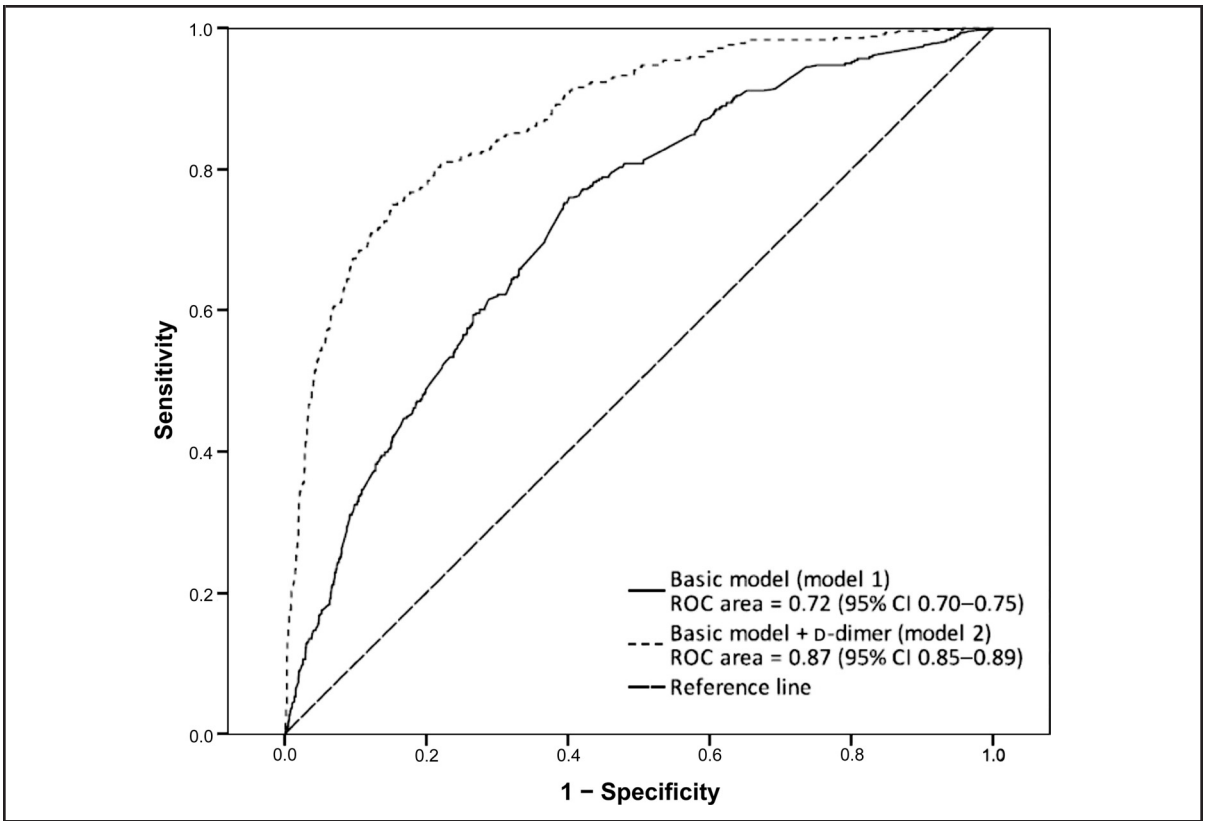

Fig. 1. ROC curves for the basic model with and without the D-dimer biomarker.

Table 3. Reclassification table from the basic and extended (with D-dimer) model, at an arbitrary cut-off value of 25%.[a]

	Model 2 with D-dimer		
DVT yes (n = 416)	**≤25**	**>25**	**Total**
Model 1 without D-dimer			
≤25	92	123	215
>25	26	175	201
Total	118	298	416
DVT no (n = 1670)			
	Model 2 with D-dimer		
	≤25	**>25**	**Total**
Model 1 without D-dimer			
≤25	1223	116	1339
>25	227	104	331
Total	1450	220	1670

[a] Patient with a model's probability of higher than 25% is considered high probability of having DVT and is further worked up or managed for DVT.

then computing the corresponding difference in proportions for those without DVT, and taking the difference of these 2 differences. This measure has been introduced as the net reclassification improvement (NRI) *(28)*. The NRI is thus estimated as follows:

$$\mathrm{NRI} = [\mathrm{P}(\mathrm{up} \mid \mathrm{D} = 1) - \mathrm{P}(\mathrm{down} \mid \mathrm{D} = 1)] - [\mathrm{P}(\mathrm{up} \mid \mathrm{D} = 0) - \mathrm{P}(\mathrm{down} \mid \mathrm{D} = 0)],$$

where P is the proportion of patients, upward movement (up) is defined as a change into a higher probability of disease presence category based on model 2, and downward movement (down) as a change in the opposite direction. D denotes the disease classification, in this DVT, present (1) or absent (0).

The NRI for addition of D-dimer assay to the combination of history and physical exam results with the use of the numbers shown in Table 3 was: (0.30 − 0.06) − (0.07 − 0.14) = 0.31 (95% CI, 0.24–0.36). For 123/416 (i.e., 0.30) patients who experienced DVT events, classification improved with the model with D-dimer, and for 26/416 (0.06) people it became worse, with the net gain in reclassification proportion of 0.24.

In patients who did not experience an event 116/1670 (0.07) individuals were reclassified worse by the

model with the D-dimer and 227/1670 (0.14) were reclassified better, resulting in a net gain in reclassification proportion of 0.07. The total net gain in reclassification proportion therefore was 0.24 + 0.07 = 0.31.

This estimate was significantly different from 0 (*P* value <0.001). The 95% CI around the NRI estimate was calculated with the formula proposed by Pencina et al. *(28)*.

The NRI is very dependent on categorization of the probability threshold(s). Most people use 3–4 categories. Different thresholds may result in very different NRIs for the same added test result. To overcome this problem of arbitrary cutoff choices, another option is to calculate the so-called integrated discrimination improvement (IDI), which considers the magnitude of the reclassification probability improvements or worsenings by a new test/biomarker over all possible categorizations or probability thresholds *(27–29)*.

The IDI is calculated as follows:

$$IDI = [(P_{extended} \mid D = 1) - (P_{basic} \mid D = 1)] - [(P_{extended} \mid D = 0) - (P_{basic} \mid D = 0)].$$

In this equation $P_{extended} \mid D = 1$ and $P_{extended} \mid D = 0$ are the means of the predicted DVT probability by the extended model 2 (Table 2) for, respectively, the patients with DVT and the patients without DVT, whereas $P_{basic} \mid D = 1$ and $P_{basic} \mid D = 0$ are the means of the predicted DVT probability by model 1 (Table 2) for, respectively, the patients with DVT and the patients without DVT. Here also the 95% CI around the NRI estimate was calculated with the formula proposed by Pencina et al. *(28)*.

The IDI for our DVT example was: (0.49 − 0.13) − (0.28 − 0.18) = 0.26 (95% CI, 0.23–0.28).

This means that the addition of D-dimer to history and physical examination increased the difference in mean predicted probability between patients with DVT and patients without DVT with 0.26. This can also be interpreted as equivalent to the increase in mean sensitivity given no changes in specificity *(28)*.

Although very popular and increasingly requested in reports on added value estimations, the NRI and IDI are only measures of discrimination between disease and nondiseased, as is also the case for ROC area. They give no information about whether the diagnostic probabilities calculated with a diagnostic model are in agreement with the observed disease prevalence, i.e., whether the models' DVT probabilities are over- or underestimated compared to the observed DVT prevalence, nor do they account in any way for the consequences of diagnostic misclassifications when a diagnostic biomarker or test is added *(30, 31)*. The following methods better address these issues.

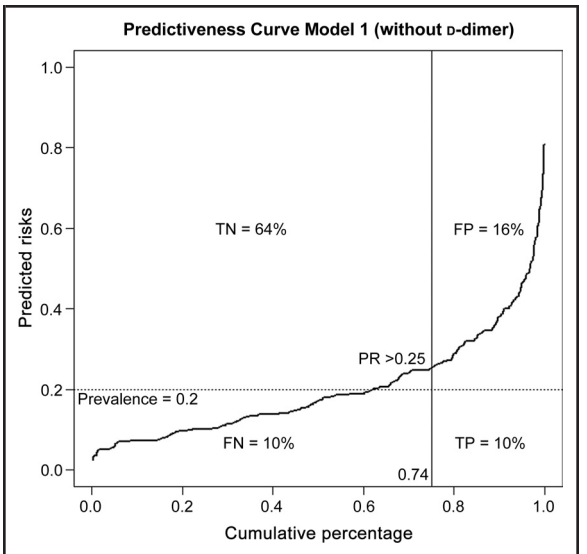

Fig. 2. Predictiveness curve for model 1 (without D-dimer) with the distribution of positive and negative patients at a predictive risk (PR) threshold of 0.25.

Predictiveness Curve

The predictiveness curve *(32, 33)* is a graphic display of the distribution of the predictive disease probabilities. For example, the predictive probabilities of model 1 (without D-dimer added) are ordered from lowest to highest and then plotted (Fig. 2).

The *x* axis depicts the cumulative percentage over all individuals in the study; the *y* axis shows the probabilities calculated with model 1. Focusing first on this single model 1 we see for our example that if those who have a posterior risk (after history and physical examination) of >25% are selected for further workup (regarded as positive), then 74% of patients will actually be negative and 26% will be positive (vertical dividing line in Fig. 2).

The 4 areas defined by the vertical dividing line represent, respectively, the true negatives (TN) (64%), false positives (FP) (16%), false negatives (FN) (10%), and true positives (TP) (10%).

In this particular example (threshold of >25%) the sensitivity becomes: TP/prevalence × 100 = 0.10/0.2 × 100 = 50%.

The specificity becomes:

TN/(1 − prevalence) × 100 = 0.64/0.8 × 100 = 80%.

The graph thus shows the range and distribution of estimated probabilities associated with the history and physical exam mode when applied to the source

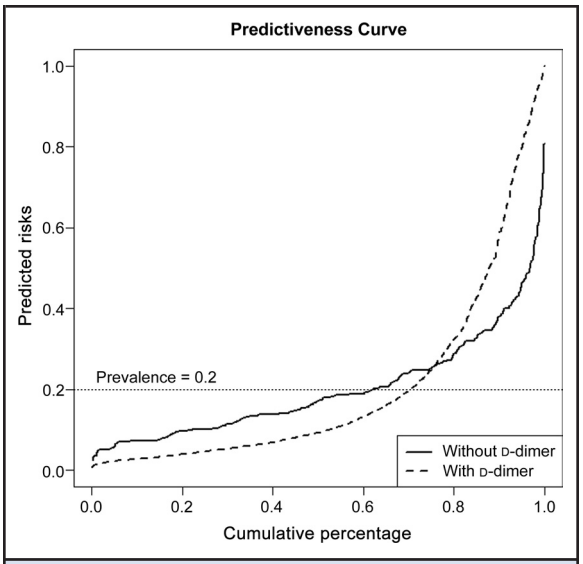

Fig. 3. Predictiveness curve for the 2 models of Table 2 with and without D-dimer.

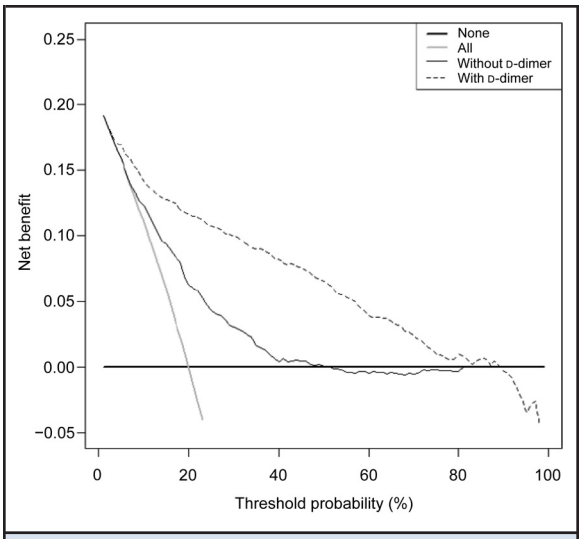

Fig. 4. Decision curve analysis, with the net benefit of referring none of the patients for reference testing, referring all patients for reference testing, the basic prediction model, and the extended prediction model, depending on the choice of probability threshold for referral.

population from which the study patients theoretically were sampled *(33)*.

But the graph can also be used to compare the different diagnostic models, and thus the added value of the D-dimer assay in terms of correct estimation of the probability of DVT presence. The predictiveness curve for model 1 is significantly less accurate than the more comprehensive model that includes the D-dimer results (Fig. 3). For example, if we choose <0.1 as a threshold for low risk and >0.4 as a threshold for high risk (at the *y* axis) we see that $90 - 20 = 70\%$ of the predictions of model 1 fall in the equivocal range between these thresholds, whereas for the predictions of model 2, only $85 - 50 = 35\%$ fall between these thresholds. This means that model 2 has much better performance in categorizing or discriminating patients into low (<0.1) vs high (>0.4) risk, as can be directly inferred from the difference in steepness or slope of the predictiveness curve of the model with D-dimer (steeper) compared to the model without *(33)*.

Decision Curve Analysis

Finally, decision curve analysis, as proposed by Vickers and Elkin *(34)*, is an approach that even more explicitly quantifies the clinical usefulness of a new test or biomarker when added to existing ones. In contrast to the NRI, in which a single predefined probability threshold is chosen, this analysis allows each physician or patient to determine his or her own desired threshold for further actions (i.e., referral for further diagnostic testing or for deciding on treatment initiation, depending on the intended use of the test or model), and to judge the corresponding net benefits without explicitly assigning weights or utilities to the false classifications.

As shown in Fig. 4, a probability threshold of 50% would imply that an incorrect referral (FP) is just as bad as a missed thrombosis (FN). A risk-averse physician/patient might opt for referral for reference testing or further management using a low threshold (e.g., if the subsequent test is relatively nonburdensome or the treatment of choice has relatively low risk of adverse reactions), e.g., if the risk of DVT is 20%. Such a lower threshold for referral means that one accepts a larger percentage of incorrect referrals (FPs) rather than missing a diseased case (FN). Otherwise, in this example, one implicitly assigns more weight to a missed DVT case (FN) compared to unnecessary referrals or workup of non-DVT patients (FPs). Alternatively, a physician/patient who is more concerned about the costs or burden of further workup (e.g., if the subsequent test is very invasive or subsequent therapy entails high risks of adverse reactions) might use a high threshold of, e.g., 70%. This means that one assigns a higher weight to incorrect referral of patients without the disease (FPs) and implicitly less weight to a missed diseased (DVT) case (FNs).

The graph shows the entire range of probability thresholds for further management on the *x* axis and the net benefit of the exhibited diagnostic strategies or models on the *y* axis *(35)*. To calculate the net benefit,

Table 4. Relationship between true DVT status and the result of the basic and extended prediction model with thresholds of 20% and 70% predicted probability.

	DVT (N = 2086)	
	Present	Absent
Basic model (model 1): probability of DVT ≥20%		
Yes	263	528
No	153	1142
Extended model (model 2): probability of DVT ≥20%		
Yes	319	301
No	97	1369
Basic model (model 1): probability of DVT ≥70%		
Yes	3	6
No	413	1664
Extended model (model 2): probability of DVT ≥70%		
Yes	123	31
No	293	1639

the proportion of all patients who are FP are subtracted from the proportion of all patients who are TP, weighting by the relative harm of an FP and an FN classification *(34)*. We illustrate this with a numerical example.

Table 4, which displays results for this empirical study, shows that when we used the above-mentioned threshold of 20%, the TP count for model 1 was 263 and the FP count 528. The total number of patients (N) was 2086. The net benefit for model 1 at the threshold of 20% was $(263/2086) - (528/2086) \times (0.2/0.8) = 0.06$. The net benefit ratio 0.2/0.8 directly indicates that less weight is now assigned to the FPs compared to the FNs, as described above. For model 2 the net benefit at the threshold of 20% was $(319/2086) - (301/2086) \times (0.2/0.8) = 0.12$, 2 times higher. Although model 2 performs clearly better than model 1, it should be noted that when the extended model is applied at this threshold, 97 of 416 known cases would not be treated or referred for further testing. This makes the overall diagnostic performance of our illustrative model 2 relatively poor at this theoretically chosen threshold.

The net benefit for model 1 at the threshold of 70% was $(3/2086) - (6/2086) \times (0.7/0.3) = -0.01$ and for model 2 the net benefit at the threshold of 70% was $(123/2086) - (31/2086) \times (0.7/0.3) = 0.02$.

The net benefit of model 1 of 0.06 at a threshold of 20% can be interpreted as: "Compared to no one referring, referral by model 1 is the equivalent of a strategy that correctly refers 6 patients with DVT of 100 suspected patients without having an unnecessary (i.e., FP) referral."

The key aspect of the decision curve (Fig. 4) is to see which diagnostic strategy leads to the highest net benefit given the doctor's or patient's individual choice for a probability threshold. The horizontal black line along the *x* axis in Fig. 4 assumes that no patients will be referred to reference testing. Because this strategy refers 0 patients, the net benefit of this strategy is set at 0, i.e., all patients with DVT are incorrectly managed. The grey steep declining line in Fig. 4 shows the net benefit for the strategy in which every individual simply undergoes reference testing. This line crosses the *x* axis at the threshold probability of 20%, i.e., prevalence of the study. Thus, when the probability threshold for further management is the same as that obtained when assigning the prevalence as the predicted risk, the net benefit of referring and not referring is the same (i.e., net benefit = 0). Furthermore it can be seen that model 2 has the greatest net benefit (i.e., it is the highest line) at all threshold probabilities. Therefore we can say that, irrespective of the applied probability thresholds, the extended model with D-dimer added is superior to the basic model.

Concluding Remarks

In this report we reiterate the limitation of single-test accuracy studies and review several traditional and relatively novel measures for quantification of the added value of a new diagnostic test or biomarker beyond available or existing diagnostic test results. To illustrate the various methods to assess the added value of a new test or biomarker, we used D-dimer testing in patients with suspicion of DVT as an example. Interestingly, in this particular example, the ROC area of the D-dimer assay result in isolation was 0.86, whereas the ROC area of the full model including D-dimer (model 2) was 0.87. From a purely quantitative and scientific perspective one therefore might argue that the D-dimer biomarker could be used in isolation without first acquiring information from a patient history and physical examination. This would seem to favor the use of a single-test evaluation approach over a multivariable or added value approach. However, we would note that to quantify whether a test does or does not have added value, one must first conduct a multivariable diagnostic study to compare the difference in diagnostic accuracy of the test in isolation vs its use in the context of other clinical information. Second, in many instances the diagnostic accuracy of a single test or biomarker will not be as high as that found in our example, and a multivariable study

would still be required to quantify the added value of the biomarker beyond information that is standardly available *(6, 36)*. One could even argue that accuracy measures of a single diagnostic test are redundant. Third, it is quite commonly observed that tests applied at a later stage in the diagnostic workup, closer to the time of reference testing, yield high diagnostic accuracy in isolation. Although some might conclude, therefore, that the advanced, often more burdening and costly test should be performed in each suspected patient, clinically such an approach would be counterintuitive. In practice, advanced tests will always be conducted after simple tests, such as history and physical examination. Thus their accuracy beyond the simple tests needs to be quantified, and not vice versa.

There are obviously circumstances in which a single-test or single-biomarker approach is indicated. For example, there are situations in which a (diagnostic) decision is actually made on the basis of 1 test/biomarker. This notably, and perhaps only, applies to screening tests *(6, 36)*. Second, single-test evaluations are recommended, often for efficiency reasons, in the initial phase of developing a new test or biomarker or evaluating an existing test or biomarker in a new context *(6, 37–40)*. If the test/biomarker cannot discriminate between individuals with proven (extreme) disease vs those without this disease, e.g., with a type of case control design (see report 1 of this series), quantifying of the added value of the test/biomarker in the indicated population may not even be reasonable. If the test yields satisfactory diagnostic accuracy in such a situation, its contribution to extant diagnostic information in its indicated context must still be evaluated. Third, in specific situations, e.g., at the end of a diagnostic workup, the question may be whether a certain test/biomarker is better or as good as another test to replace it, e.g., computed tomography (CT) vs MRI scanning in patients suspected of lung cancer *(5, 39)*. A study in which the accuracy measures of the CT and MRI are compared may suffice, although one could argue that there might be a difference in added value, e.g., the results of the MRI scan may be more similar to preceding test results than the results of the CT scan such that the latter has more added value than the former. In most other instances, a biomarker or test is commonly part of a diagnostic workup that is performed after obtaining previous test results, requiring physicians to know whether this next test or biomarker result adds substantively to the diagnosis of the target disease.

For clinical practice, providing insight beyond the ROC area has been a motivation for some recent measures, especially in the context of extending a diagnostic model with additional information from a novel biomarker *(8, 21, 28, 33, 41)*. Researchers and physicians should recognize, however, that a single summary measure cannot give full insight in all relevant aspects of the added, clinical value of a new test or biomarker. Reporting on the increase in discrimination and classification is relevant to obtain insight into the incremental value of a biomarker. These measures should therefore ideally be reported in studies evaluating the incremental value of a novel test or biomarker. However, we also recommend the use of decision-curve analysis, because this method implicitly accounts for the consequences of the FP and negative classifications, in contrast to the other measures.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: K.G.M. Moons, the Netherlands Organization for Scientific Research (project 9120.8004 and 918.10.615).
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

References

1. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press, 2003.
2. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology. 2nd ed. Boston (MA): Little, Brown and Company; 1991.
3. Hoes AW, Grobbee DE. Chapter 3 Diagnostic Research. Clinical epidemiology. Sudbury (MA): Jones & Bartlett Publishers; 2008.
4. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138:W1–12.
5. Lord SJ, Staub LP, Bossuyt PM, Irwig LM. Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. BMJ 2011;343:d4684.
6. Moons KG, van Es GA, Michel BC, Buller HR, Habbema JD, Grobbee DE. Redundancy of single diagnostic test evaluation. Epidemiology 1999; 10:276–81.
7. Moons KG, Grobbee DE. Diagnostic studies as multivariable, prediction research. J Epidemiol Community Health 2002;56:337–8.
8. Moons KG. Criteria for scientific evaluation of novel markers: a perspective. Clin Chem 2010;56: 537–41.
9. Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity,

likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997;8:12–7.
10. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol 2009;62:5–12.
11. Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med 1984;77:64–71.
12. Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost 2005;94:200–5.
13. Toll DB, Oudega R, Bulten RJ, Hoes AW, Moons KG. Excluding deep vein thrombosis safely in primary care. J Fam Pract 2006;55:613–8.
14. Rutjes AW, Reitsma JB, Di NM, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174:469–76.
15. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess 2007;11:iii, ix–51.
16. de Groot JA, Bossuyt PM, Reitsma JB, Rutjes AW, Dendukuri N, Janssen KJ, Moons KG. Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ 2011;343:d4770.
17. de Groot JA, Janssen KJ, Zwinderman AH, Bossuyt PM, Reitsma JB, Moons KG. Correcting for partial verification bias: a comparison of methods. Ann Epidemiol 2011;21:139–48.
18. de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Bossuyt PM, Moons KG. Adjusting for differential-verification bias in diagnostic-accuracy studies: a Bayesian approach. Epidemiology 2011;22:234–41.
19. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.
20. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
21. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928–35.
22. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 2004;159:882–90.
23. Ware JH. The limitations of risk factors as prognostic tools. N Engl J Med 2006;355:2615–7.
24. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. JAMA 2009;302:2345–52.
25. Pencina MJ, D'Agostino RB, Vasan RS. Statistical methods for assessment of added usefulness of new biomarkers. Clin Chem Lab Med 2010;48:1703–11.
26. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Ann Intern Med 2009;150:795–802.
27. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van CB. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. Eur J Clin Invest 2011;42:216–28.
28. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27:157–72.
29. Pepe MS, Feng Z, Gu JW. Comments on 'Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in medicine (DOI: 10.1002/sim. 2929). Stat Med 2008;27:173–81.
30. Greenland S. The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond' by M. J. Pencina et al. Statistics in Medicine (DOI: 10.1002/sim. 2929). Stat Med 2008;27:199–206.
31. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128–38.
32. Huang Y, Sullivan PM, Feng Z. Evaluating the predictiveness of a continuous marker. Biometrics 2007;63:1181–8.
33. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol 2008;167:362–8.
34. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74.
35. Steyerberg EW, Vickers AJ. Decision curve analysis: a discussion. Med Decis Making 2008;28:146–9.
36. Moons KG, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. Clin Chem 2004;50:473–6.
37. van der Schouw YT, Verbeek AL, Ruijs JH. ROC curves for the initial assessment of new diagnostic tests. Fam Pract 1992;9:506–11.
38. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem 2005;51:1335–41.
39. Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. J Clin Epidemiol 2009;62:364–73.
40. Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG. Advantages of the nested case-control design in diagnostic research. BMC Med Res Methodol 2008;8:48.
41. Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation 2009;119:2408–16.