

How to Measure the Diagnostic Accuracy of Noninvasive Liver Fibrosis Indices: The Area Under the ROC Curve Revisited

Jerome Lambert,¹ Philippe Halfon,² Guillaume Penaranda,³ Pierre Bedossa,⁴ Patrice Cacoub,⁵ and Fabrice Carrat^{1*}

BACKGROUND: The area under the ROC curve (AUC) is widely used to measure the diagnostic accuracy of noninvasive fibrosis indices. However, use of the AUC assumes a binary gold standard, whereas fibrosis staging is based on an ordinal scale and also depends on the distribution of fibrosis stages in the study sample. We explored other fibrosis staging accuracy measures designed for ordinal gold standards, the C-statistic and the Obuchowski measure.

METHODS: We performed a simulation study to assess the bias in estimating the accuracy measures when the distribution of fibrosis stages in the study sample do not fit the reference distribution in the population to which the indices are applied. We also estimated the type I error of the tests comparing these measures in 2 samples with different distributions of fibrosis stages. We illustrated the practical use of these measures by reanalyzing real data.

RESULTS: Compared with the AUC or the C-statistic, the Obuchowski measure showed limited bias when the distribution of fibrosis stages in the study sample differed from the reference distribution. The type I error was strongly inflated with the AUC or the C-statistic but was preserved in the Obuchowski measure. When we compared noninvasive indices on real data, AUC analysis led to discordant results depending on how the fibrosis stages were grouped together. One single conclusion was drawn from the analysis based on the Obuchowski measure.

CONCLUSIONS: We recommend using the Obuchowski measure for assessing the diagnostic accuracy of noninvasive indices of fibrosis.

© 2008 American Association for Clinical Chemistry

Clinical management of chronic viral hepatitis is based on the assessment of liver fibrosis. The stage of fibrosis is a key factor for prognosis and decision-making regarding antiviral treatment (1–3). Liver biopsy is currently the gold standard for fibrosis staging in chronic viral hepatitis and other chronic liver diseases. Several semiquantitative staging systems have been validated, such as the METAVIR score, which evaluates fibrosis on a 5-point scale (F0 = no fibrosis, F1 = portal fibrosis without septa, F2 = portal fibrosis and few septa, F3 = septal fibrosis without cirrhosis, F4 = cirrhosis) (4). However, liver biopsy is an invasive procedure, with complications occurring in 1%–3% of patients and death in 1 per 10 000–12 000 biopsies (5, 6). Performing a liver biopsy is also costly and requires hospitalization for 6–18 h (7, 8).

During the last 10 years, because of the difficulties associated with liver biopsy, there has been a growing interest in noninvasive methods for assessing liver fibrosis (7). The method used to estimate the diagnostic accuracy of noninvasive methods is remarkably constant across the different studies and is based on the area under the ROC curve (AUC)⁶ (7, 9).

In this context, the AUC represents the probability that a noninvasive index will correctly rank 2 randomly chosen patients, 1 with a liver biopsy considered “diseased” and the other with a liver biopsy considered “normal” (10). The diagnostic accuracies of 2 noninvasive indices are compared using the 2 AUCs and an appropriate statistical test (11, 12).

The use of the AUC raises 2 methodological issues. First, its use is based on the assumption that the gold standard is binary, whereas fibrosis staging uses an ordinal scale. This difference implies that fibrosis stages in the study sample have to be aggregated into 2 groups,

¹ Université Pierre et Marie Curie-Paris; INSERM, UMR-S 707; unité de santé publique, Assistance Publique Hôpitaux de Paris, Hôpital Saint-Antoine, Paris, France; ² Laboratoire Alphabio, Hôpital Ambroise Paré, Marseille, France; ³ Département de Biostatistiques, CDL Pharma, Marseille, France; ⁴ Service d'Anatomie Pathologique, Assistance Publique Hôpitaux de Paris, Hôpital Beaujon, Clichy, France; ⁵ Service de Médecine Interne, Hôpital Pitié-Salpêtrière, Assistance Publique Hôpitaux de Paris, Paris, France.

* Address correspondence to this author at: UMR-S 707, Faculté de Médecine

Saint Antoine, 27, rue Chaligny, 75571 Paris cedex 12. Fax +33 1 44 73 84 53; e-mail carrat@u707.jussieu.fr.

Received September 18, 2007; accepted May 6, 2008.

Previously published online at DOI: 10.1373/clinchem.2007.097923

⁶ Nonstandard abbreviations: AUC, area under the ROC curve; AF, advanced fibrosis; adjAUC, adjusted AUC; NAF, nonadvanced fibrosis; C-statistic, concordance statistic; FT, Fibrotest; APRI, aspartate aminotransferase-to-platelet ratio index.

a process that can lead to discordant conclusions, depending on how the groups are aggregated. The C-statistic, which was introduced to estimate diagnostic accuracy for outcomes with more than 2 categories (13), can overcome this limitation, but has never been used to estimate the diagnostic accuracy of noninvasive indices.

Analysis based on the AUC can also be biased by the way in which the proportion of each stage of fibrosis in the sample fits the distribution in the reference population to which the indices are applied. As a result, the comparison of different AUCs based on samples with different stage distributions may be flawed (14). A recent report advocated standardizing the AUC for the distribution of fibrosis stages to deal with this source of variability, but the method is not straightforward and has not yet been validated from a statistical standpoint (15, 16).

To overcome these 2 methodological issues, Obuchowski recently proposed a measure that can be interpreted similarly to the AUC and can be used in situations in which the gold standard is not binary (17–19).

The aim of this study was to compare the AUC, C-statistic, and Obuchowski measures to assess the diagnostic accuracy of noninvasive fibrosis indices.

Materials and Methods

BRIEF OVERVIEW OF ACCURACY MEASURES

AUC

The AUC of noninvasive indices is generally used to differentiate between patients with advanced fibrosis and patients with non advanced fibrosis. Several methods for estimating and comparing the AUCs have been described (20, 21). In this study we used a nonparametric estimate of the AUC, equivalent to the Mann–Whitney statistic (10, 11). (More details can be found in the Data Supplement that accompanies the online version of this article at <http://www.clinchem.org/content/vol54/issue8>.)

Briefly, the calculation relies on selecting every possible pair of patients, one with advanced fibrosis (AF, i.e., stages F2, F3 and F4 of the METAVIR score) and one with nonadvanced fibrosis (NAF, i.e., stages F0 and F1), and then evaluating if the noninvasive index correctly ranks the two patients. The estimated AUC is the proportion of all pairs in which the patient with advanced fibrosis has the higher value of the noninvasive index. It can be interpreted as the probability that the noninvasive index will correctly rank 2 randomly chosen patients, one with AF and one with NAF.

C-STATISTIC (13)

The concordance (C)-statistic is an accuracy measure that can be used for ordinal or nominal outcomes. If we assume that there are N categories of the gold standard outcome (in this case, 5 fibrosis stages), calculation of the C-statistic requires selection of every possible pair of patients having different categories of the outcome and evaluation of the proportion of all pairs in which the noninvasive index correctly ranks the 2 patients. The C-statistic has the same interpretation as the AUC, i.e., the probability of correctly ranking 2 randomly chosen patients in 2 different categories. Like the AUC, the C-statistic depends on the distribution of fibrosis stages in the study sample.

OBUCHOWSKI MEASURE (19)

This measure is a multinomial version of the AUC. With N ($= 5$) categories of the gold standard outcome and AUC_{st} , the estimate of the AUC of diagnostic tests for differentiating between categories s and t , the Obuchowski measure, is a weighted average of the $N(N - 1)/2$ ($= 10$) different AUC_{st} corresponding to all the pairwise comparisons between 2 of the N categories. Weighting can be based on the relative proportion of the 5 fibrosis stages in the study sample, or, as in this case, on a reference distribution of fibrosis stages similar to that in the population.

Each pairwise comparison can also be weighted to take into account the distance between fibrosis stages (i.e., the number of units on the ordinal scale). We thus defined a penalty function proportional to the difference in METAVIR units between stages (the penalty function was 0.25 when the difference between stages was 1, 0.5 when the difference was 2, 0.75 when the difference was 3, and 1 when the difference was 4).

With a weighting scheme based on the relative proportion of fibrosis stages in the study sample and no penalty function, the Obuchowski measure is equivalent to the C-statistic. Note also that the AUC can be seen as a particular value of the Obuchowski measure, for which AUC_{st} corresponding to pairwise comparisons of stages s and t belonging to the same aggregated category (i.e., AF or NAF) are not calculated, with a weighting scheme based on the relative proportion of stages in the study sample and no penalty function. In this latter case, if the weighting scheme is based on a reference distribution of stages in the population, an adjusted-to-the-stages distribution AUC (adjAUC) is estimated.

The Obuchowski measure can be interpreted as the probability that the noninvasive index will correctly rank 2 randomly chosen patient samples from different fibrosis stages according to the weighting scheme, with a penalty for misclassifying patients (see above).

COMPARISON OF NONINVASIVE INDICES

A general method for comparing 2 or more AUCs derived from the same patient population has been published elsewhere (11). This method has also been extended to the Obuchowski measure (19). Assuming that $\hat{\theta}_1$ and $\hat{\theta}_2$ are the respective measures of diagnostic accuracy (the AUC, the C-statistic, or the Obuchowski measure) of 2 noninvasive indices, the value of the test statistic for assessing the null hypothesis (no difference in accuracy between the 2 indices) is:

$$z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\text{var}(\hat{\theta}_1) + \text{var}(\hat{\theta}_2) - 2\text{cov}(\hat{\theta}_1, \hat{\theta}_2)}}$$

which follows a standard gaussian distribution. Detailed calculation of the variances ($\text{var}(\hat{\theta}_1)$ and $\text{var}(\hat{\theta}_2)$) and covariance ($\text{cov}(\hat{\theta}_1, \hat{\theta}_2)$) are described elsewhere (11, 19).

All statistical tests were 2-tailed, with a type I error of 5%.

DATA SET

The data on noninvasive indices used here come from a previously published report (22), derived from the Fibropaca study. Fibropaca was a French multicenter prospective cross-sectional study involving 519 patients that was performed in hepatogastroenterology units or internal medicine units of 5 centers in the southeast region, known for their expertise in hepatitis C (23). All the patients had chronic hepatitis C virus infection without liver complications. Liver biopsies were analyzed for the fibrosis stage in each center by the local pathologist, using the METAVIR scoring system. On the same day as the biopsy specimens were obtained, biochemical parameters were collected to assess several noninvasive markers in a subgroup of 235 patients.

Our analysis focused on APRI (aspartate aminotransferase-to-platelet ratio index) and Fibrotest (FT). FT is calculated from the patient's age, sex, and 5 biochemical parameters: α 2-macroglobulin, haptoglobin, γ -glutamyl transpeptidase (GGT), total bilirubin, and apolipoprotein A1 (24). APRI is the ratio of the aspartate aminotransferase concentration to the platelet count (25).

SIMULATION

We illustrated the variability and bias related to the discrepancy between the distributions of fibrosis stages in the study sample compared to its reference distribution in the population. We estimated the nonadjusted (AUC, C-statistic) and adjusted (adjAUC, Obuchowski measure) accuracy measures of FT in 1000 samples of size 235, sampled from the Fibropaca study with different distributions of fibrosis stages, namely a

predominance of extreme stages [proportion of stage F0 (P_{F0}) = 30%, P_{F1} = 10%, P_{F2} = 10%, P_{F3} = 20%, P_{F4} = 30%], and a predominance of intermediate stages (P_{F0} = 10%, P_{F1} = 30%, P_{F2} = 30%, P_{F3} = 20%, P_{F4} = 10%). To describe the population, a reference distribution of fibrosis stages in this setting was chosen (P_{F0} = 6%, P_{F1} = 39%, P_{F2} = 28%, P_{F3} = 14%, P_{F4} = 13%) (26). The true values of the AUC, the adjAUC, the C-Statistic, or the Obuchowski measure were empirically calculated from 1000 samples of size 235 under the reference distribution, and the bias was calculated by averaging the differences between each estimated measure and their corresponding true value. We also calculated the nominal coverage of 95% CIs, i.e., how often the true value was included in the 95% CI.

Finally, we evaluated the type I error of tests comparing the FT in 2 samples with different distribution of stages (extreme vs intermediate), based either on an adjusted accuracy measure (adjAUC and Obuchowski measure) or a nonadjusted accuracy measure (AUC or C-statistic). All calculations were performed using R.

FIBROPACA ANALYSIS

We used the Obuchowski measure to assess the accuracy of FT and APRI for diagnosing the stage of liver fibrosis. We compared these results to those obtained by using AUC analysis, and the AUC_{st} values corresponding to each pairwise comparison were plotted to explain discrepancies and to illustrate the consequences of grouping fibrosis stages.

Results

VARIABILITY AND BIAS RELATED TO DIFFERENCES IN THE DISTRIBUTION OF FIBROSIS STAGES IN THE STUDY SAMPLE AND THE REFERENCE POPULATION

Fig. 1 shows the AUC, adjAUC, C-statistic, and Obuchowski measure of FT calculated in 1000 simulated samples with different distributions of fibrosis stages. The true values in the population were 0.808 for the AUC, 0.756 for the C-statistic, and 0.796 for the Obuchowski measure. It is clear that the AUC and the C-statistic depended strongly on the distribution of fibrosis stages in the sample, whereas the adjusted measures did not. Considering the extreme distribution, which differed strongly from the reference distribution, the bias was 0.063 (95% CI, 0.016–0.105) for the AUC and 0.031 (95% CI, –0.007–0.067) for the C-statistic. For the adjusted measures, the bias was 0 (95% CI, –0.094–0.084) for adjAUC and 0 (95% CI, –0.056–0.046) for the Obuchowski measure. The bias was lower for the intermediate distribution, which was closer to the reference distribution, with values of

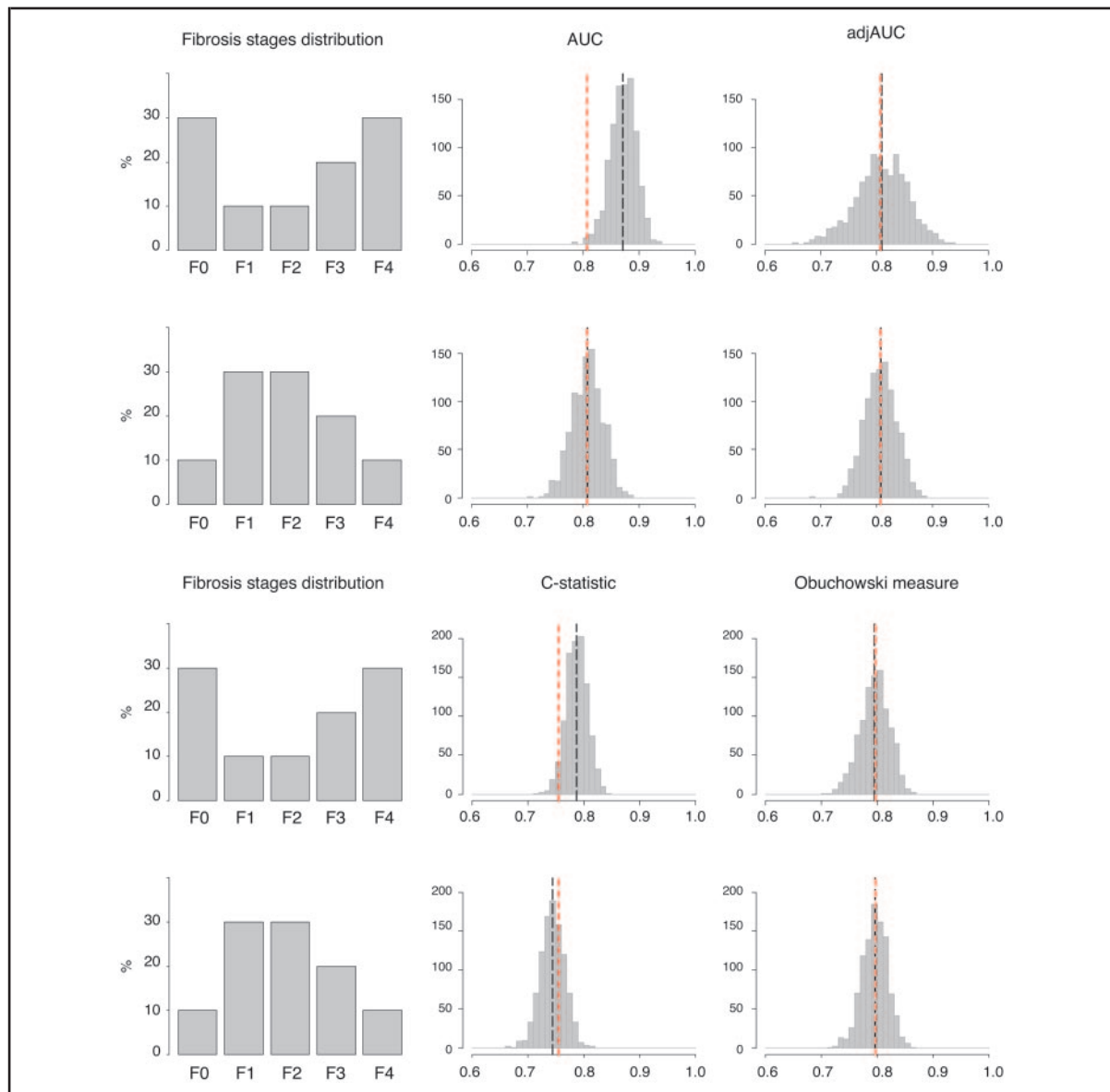


Fig. 1. AUC, adjAUC, C-statistic, and Obuchowski measure for the FT index according to the distribution of fibrosis stages in the sample.

Each accuracy measure was calculated for 1000 sets of data of 235 patients sampled from the Fibropaca study. The black dashed lines indicate the mean accuracy measures over the 1000 samples. The red dashed lines indicate the true accuracy measures in the population.

−0.002 (95% CI, −0.064–0.053) for the AUC, and −0.012 (95% CI, −0.053–0.028) for the C-statistic.

With the extreme distribution, the nominal coverage of 95% CIs was 28% for the AUC, meaning that only 28% the 95% CI contained the true AUC value. The nominal coverage of 95% CI was 60% for the C-

statistic, whereas the adjusted measures had 95% coverage as expected.

TYPE I ERROR

When comparing the same index (FT) between samples with intermediate vs extreme stage distributions,

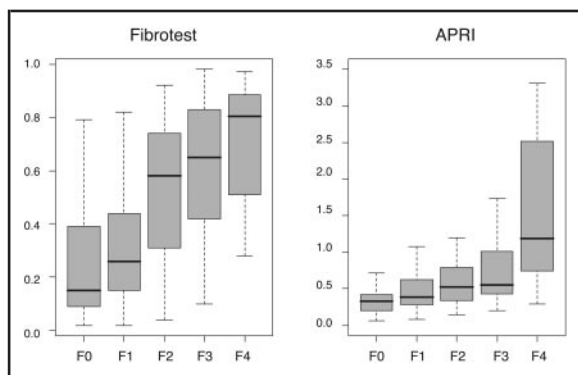


Fig. 2. FT and APRI indices for each fibrosis stage in the Fibropaca study.

The top and bottom of the box are the 25th and 75th percentiles. The line across the box represents the median. The whiskers extend to 1.5 times the box width from the top and bottom of the box.

the type I error was 42% with the AUC and 33% with the C-statistic, while the corresponding values were 6% with adjAUC and 5% with the Obuchowski measure.

ANALYSIS OF FIBROPACA DATA

With the METAVIR system, the fibrosis stages were 14% F0, 43% F1, 18% F2, 18% F3, and 7% F4. Fig. 2 shows the distribution of FT and APRI according to the fibrosis stage. The AUCs of FT and APRI for the diagnosis of advanced fibrosis ($\geq F2$) were, respectively, 0.81 (95% CI, 0.76–0.87) and 0.74 (95% CI, 0.67–0.80). The difference was statistically significant ($P = 0.02$), leading to the conclusion that FT has greater accuracy than APRI. For the diagnosis of cirrhosis (F4), there was no significant difference ($P = 0.82$) between FT [AUC = 0.82 (95% CI, 0.73–0.92)] and APRI [AUC = 0.83 (95% CI, 0.72–0.95)]. To understand why we found discordant results, we plotted the AUCs of all pairwise comparisons of fibrosis stages (Fig. 3). When calculating the AUC for the diagnosis of significant fibrosis, we averaged 6 pairwise comparisons, namely F0 or F1 with F2, F3, or F4. Among these, FT was more accurate in 5 comparisons. For the diagnosis of cirrhosis, 4 pairwise comparisons were averaged, namely F0, F1, F2, and F3 with F4. APRI was more accurate in 3 comparisons, but the averaged difference was not significant because the F1 vs F4 comparison favored FT and included a larger number of patients. When the analysis used adjAUC, similar conclusions were drawn in all comparisons.

We then reanalyzed the diagnostic accuracy of these 2 noninvasive markers with the measures designed for ordinal gold standards. The C-statistic values of FT and APRI were, respectively, 0.75 (95% CI,

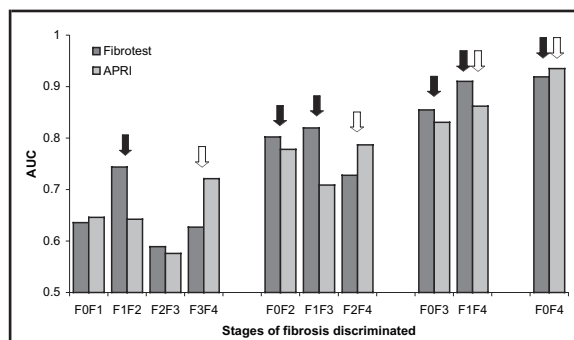


Fig. 3. AUC_{st}s of 10 pairwise comparisons of FT and APRI in the Fibropaca study.

Black arrows indicate the pairwise comparisons taken into account when calculating the AUC for the diagnosis of advanced fibrosis ($\geq F2$). White arrows indicate the pairwise comparisons taken into account when calculating the AUC for the diagnosis of cirrhosis (F4).

0.71–0.80), and 0.71 (95% CI, 0.66–0.75) ($P = 0.053$). The Obuchowski measures of FT and APRI were, respectively, 0.80 (95% CI, 0.75–0.84) and 0.75 (95% CI, 0.69–0.81), $P = 0.09$. In the Obuchowski measures, 10 pairwise comparisons were averaged, among which FT was more accurate in 6 comparisons. A single conclusion was drawn: in the population, and considering the penalty function, FT would not be more accurate than APRI for predicting the fibrosis stage.

Discussion

Noninvasive indices of liver fibrosis are a subject of growing interest. Studies designed to validate such methods generally use the AUC as a summary measure of diagnostic accuracy.

Here we present a new measure, initially developed by Obuchowski for nonbinary gold standards, and show how it can be used to evaluate the accuracy of noninvasive indices. This measure summarizes all pairwise comparisons of fibrosis stages defined by liver biopsy, with a weighting scheme and a penalty function.

The Obuchowski measure has several advantages over the AUC. By using a weighting scheme based on a reference distribution, we eliminated the bias related to the distribution of fibrosis stages and corrected the inflated type I error. This bias is the consequence of a spectrum effect, which has been widely discussed in the literature since the introductory paper by Ransohoff and Feinstein (27). By using Obuchowski measure with the same weighting scheme, results from different studies could easily be compared or combined in a meta-analysis, and the spectrum effect is controlled.

Moreover, the Obuchowski measure can be used to estimate an adjusted-on-fibrosis-stages distribution AUC, by omitting pairwise comparisons of stages belonging to the same aggregated category and with no penalty function.

Another advantage of the Obuchowski measure is that AUC analyses require the results of liver biopsy to be aggregated into 2 outcomes. Numerous studies have shown different AUCs for the same noninvasive index, owing to different ways of grouping fibrosis stages; this procedure can be interpreted as subgroup analysis. When comparing 2 noninvasive indices, this approach would imply multiple testing of several AUCs, which would require appropriate correction for the type I error. It can also lead to discrepancies in the results, which complicate their interpretation, as seen in our reanalysis of the Fibropaca data. In contrast, the Obuchowski measure allows 2 noninvasive indices to be compared with a single test. However, the study of statistical power is not straightforward and will depend on the weighting scheme and on how the penalty function is parameterized. Most notably the power will also depend on the homogeneity of the difference between 2 indices of the AUC_{st} for discriminating between 2 categories in the ordinal outcome.

Third, the use of a weighting scheme and a penalty function increase the clinical relevance of the Obuchowski measure. Measures of diagnostic accuracy should ideally reflect real-life conditions. Clearly, the medical consequences of misclassifying an F0 patient as F1 are less serious than if the same patient is misclassified as F4.

The choice of a linear penalty function to quantify the difference between observed and predicted fibrosis stages is open to discussion (28, 29). Other penalty functions might be used, more closely related to the true difference in fibrosis between different stages, or based on the clinical consequences of misclassifying a patient. This deserves further studies.

Here we analyzed the results obtained with the FT and APRI indices in the Fibropaca study, but the Obuchowski measure has far wider potential applications in the more general field of diagnostic tests with nominal or ordinal outcomes. The method has been successfully applied to assess the accuracy of magnetic resonance imagery for diagnosing damage to heart tissue after myocardial infarction (17) and to assess physician accuracy in diagnosing the cause of abdominal pain in children (19), yet the Obuchowski measure is still not widely used. It could also be used to assess the usefulness of ordinal or polytomous regression model for differentiating between more than 2 outcomes. To our knowledge, ordinal or polytomous regressions have never been considered for noninvasive indices and are still rarely used in diagnostic research (30). In a recent comparison of dichotomous and polytomous regression analyses for diagnosing serious bacterial infections (31), 3 outcomes were studied and 3 AUCs (presence of 1 outcome vs absence) were calculated for each estimated model. The Obuchowski measure could have been used instead and would have permitted the comparison of the discriminating performance of these models with a single metric.

To conclude, we recommend that future studies of noninvasive methods for assessing fibrosis use the Obuchowski measure instead of the AUC to assess diagnostic accuracy. For greater clinical relevance, we recommend a weighting scheme based on a fibrosis stage distribution as close as possible to that in the reference population, and a penalty function proportional to the difference between fibrosis stages.

Grant/Funding Support: This work received financial support from Fondation pour la Recherche Médicale.

Financial Disclosures: None declared.

Acknowledgments: We thank David Young and Anders Boyd for their help in editing the manuscript.

References

- Alberti A, Clumeck N, Collins S, Gerlich W, Lundgren J, Palu G, et al. Short statement of the first European Consensus Conference on the treatment of chronic hepatitis B and C in HIV co-infected patients. *J Hepatol* 2005;42: 615–24.
- Dhumeaux D, Marcellin P, Lerebours E. Treatment of hepatitis C. The 2002 French consensus. *Gut* 2003;52:1784–7.
- National Institutes of Health Consensus Development Conference Statement: Management of Hepatitis C: 2002—June 10–12, 2002. *Hepatology* 2002;36:S3–20.
- Bedossa P. Intraobserver and interobserver variations in liver biopsy interpretation in patients with chronic hepatitis C. *Hepatology* 1994;20: 15–20.
- Bravo AA, Sheth SG, Chopra S. Liver biopsy. *N Engl J Med* 2001;344:495–500.
- van der Poorten D, Kwok A, Lam T, Ridley L, Jones DB, Ngu MC, Lee AU. Twenty-year audit of percutaneous liver biopsy in a major Australian teaching hospital. *Intern Med J* 2006;36:692–9.
- Sebastiani G, Alberti A. Non invasive fibrosis biomarkers reduce but not substitute the need for liver biopsy. *World J Gastroenterol* 2006;12: 3682–94.
- Wong JB, Koff RS. Watchful waiting with periodic liver biopsy versus immediate empirical therapy for histologically mild chronic hepatitis C: a cost-effectiveness analysis. *Ann Intern Med* 2000;133: 665–75.
- Gebo KA, Herlong HF, Torbenson MS, Jenckes MW, Chander G, Ghanem KG, et al. Role of liver biopsy in management of chronic hepatitis C: a systematic review. *Hepatology* 2002;36:S161–72.
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44: 837–45.
- Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983;148:839–43.
- Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982;247:2543–6.

-
14. Poynard T, Halfon P, Castera L. La prévalence de chaque stade de fibrose est un facteur majeur de variabilité de l'estimation de la valeur diagnostique des marqueurs de fibrose par les courbes ROC. *Gastroenterol Clin Biol* 2006;30:1039.
 15. Poynard T, Halfon P, Castera L, Charlotte F, Le Bail B, Munteanu M, et al. Variability of the area under the receiver operating characteristic curves in the diagnostic evaluation of liver fibrosis markers: impact of biopsy length and fragmentation. *Aliment Pharmacol Ther* 2007;25:733–9.
 16. Poynard T, Halfon P, Castera L, Munteanu M, Imbert-Bismut F, Ratziu V, et al. Standardization of ROC curve areas for diagnostic evaluation of liver fibrosis markers based on prevalences of fibrosis stages. *Clin Chem* 2007;53:1615–22.
 17. Obuchowski NA. Estimating and comparing diagnostic tests' accuracy when the gold standard is not binary. *Acad Radiol* 2005;12:1198–204.
 18. Obuchowski NA. An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale. *Stat Med* 2006;25:481–93.
 19. Obuchowski NA, Goske MJ, Applegate KE. Assessing physicians' accuracy in diagnosing paediatric patients with acute abdominal pain: measuring accuracy for multiple diseases. *Stat Med* 2001;20:3261–78.
 20. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;39:561–77.
 21. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;8:283–98.
 22. Bourliere M, Penaranda G, Renou C, Botta-Fridlund D, Tran A, Portal I, et al. Validation and comparison of indexes for fibrosis and cirrhosis prediction in chronic hepatitis C patients: proposal for a pragmatic approach classification without liver biopsies. *J Viral Hepat* 2006;13:659–70.
 23. Halfon P, Bourliere M, Deydier R, Botta-Fridlund D, Renou C, Tran A, et al. Independent prospective multicenter validation of biochemical markers (fibrotest-actitest) for the prediction of liver fibrosis and activity in patients with chronic hepatitis C: the Fibropaca study. *Am J Gastroenterol* 2006;101:547–55.
 24. Imbert-Bismut F, Ratziu V, Pieroni L, Charlotte F, Benhamou Y, Poynard T. Biochemical markers of liver fibrosis in patients with hepatitis C virus infection: a prospective study. *Lancet* 2001;357:1069–75.
 25. Wai CT, Greenson JK, Fontana RJ, Kalbfleisch JD, Marrero JA, Conjeevaram HS, Lok AS. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. *Hepatology* 2003;38:518–26.
 26. Payan C, Roudot-Thoraval F, Marcellin P, Bled N, Duverlie G, Fouchard-Hubert I, et al. Changing of hepatitis C virus genotype patterns in France at the beginning of the third millennium: The GEMHEP GenoCII Study. *J Viral Hepat* 2005;12:405–13.
 27. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299:926–30.
 28. Bedossa P, Dargere D, Paradis V. Sampling variability of liver fibrosis in chronic hepatitis C. *Hepatology* 2003;38:1449–57.
 29. Hubscher SG. Histological grading and staging in chronic hepatitis: clinical applications and problems. *J Hepatol* 1998;29:1015–22.
 30. Biesheuvel CJ, Vergouwe Y, Steyerberg EW, Grobbee DE, Moons KG. Polytomous logistic regression analysis could be applied more often in diagnostic research. *J Clin Epidemiol* 2008;61:125–34.
 31. Roukema J, van Loenhout RB, Steyerberg EW, Moons KG, Bleeker SE, Moll HA. Polytomous regression did not outperform dichotomous logistic regression in diagnosing serious bacterial infections in febrile children. *J Clin Epidemiol* 2008;61:135–41.