Published in final edited form as:

J Hepatol. 2009 January ; 50(1): 36–41. doi:10.1016/j.jhep.2008.07.039.

Exceeding the limits of liver histology markers

Shruti H. Mehta, PhD MPH¹, Bryan Lau, PhD^{1,2}, Nezam H. Afdhal, MD³, and David L. Thomas, MD MPH^{1,2}

¹Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

²Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD

³Liver Center, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA

Abstract

Background/Aims—Alternatives to liver biopsy for staging liver disease caused by hepatitis C virus (HCV) have not appeared accurate enough for widespread clinical use. We characterized the magnitude of the impact of error in the “gold standard” on the observed diagnostic accuracy of surrogate markers.

Methods—We calculated the area under the receiver operating characteristic curve (AUROC) for a surrogate marker against the gold standard (biopsy) for a range of possible performances of each test (biopsy and marker) against truth and a gradient of clinically significant disease prevalence.

Results—In the ‘best’ scenario where liver biopsy accuracy is highest (sensitivity and specificity of biopsy are 90%) and the prevalence of significant disease 40%, the calculated AUROC would be 0.90 for a perfect marker (99% actual accuracy) which is within the range of what has already been observed. With lower biopsy sensitivity and specificity, AUROC determinations > 0.90 could not be achieved even for a marker that perfectly measured disease.

Conclusions—We demonstrate that error in the liver biopsy result itself makes it impossible to distinguish a perfect surrogate from ones that are now judged by some as clinically unacceptable. An alternative gold standard is needed to assess the accuracy of tests used to stage HCV-related liver disease.

Keywords

liver disease; biopsy; hepatitis C virus; validity; surrogate markers

BACKGROUND

Liver biopsy is widely considered as the gold standard for assessment of treatment urgency in persons with hepatitis C virus (HCV)-related liver disease (1-3). Because of biopsy expense and medical risk, there is a widespread effort to develop a safer, less expensive surrogate (4, 5). Candidate surrogates have included blood tests, algorithms based on the results of multiple serum markers (6-12), liver elastography (13), and others. However, in scores of studies of different surrogates, the diagnostic accuracy of candidate tests (compared to biopsy) has failed

Corresponding Author: David L. Thomas, MD PhD, Chief, Infectious Diseases, Professor of Medicine, 1830 E Monument St, Room 455-ID, Baltimore, MD 21287, 410.955.0349 (phone), 410.614.7564 (fax), dthomas@jhmi.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

to exceed 0.88 of the area under the receiver operating characteristic curve (AUROC) (6-12, 14). A recent review of studies of the most widely validated surrogate markers, FibroTest and Fibroscan reinforced that surrogate markers have not been widely adopted in clinical practice primarily because of these perceived limitations in diagnostic accuracy (15).

It is widely appreciated that there is error in the liver biopsy measurement itself. Marked reductions in the sensitivity for detection of significant fibrosis have been demonstrated with biopsies of less than 3 cm in length (16,17), fragmentation (18) and steatosis (19) which, together with regional differences in fibrosis (e.g., left versus right lobe) and lack of agreement among those examining slides, comprise error in this gold standard (20). Even among biopsies up to 4 cm in length, substantial error has been observed when biopsy specimens have been compared to the full liver (16). Thus, an alternative interpretation of the limited diagnostic accuracy of surrogate markers is that it is due to error of the biopsy measurement itself (6,19, 21,22).

When errors in a diagnostic test and the gold standard are independent, the observed sensitivity and specificity of the diagnostic test will be underestimated (23-25). However, the degree to which measurement error in the biopsy may impact the observed diagnostic accuracy of fibrosis marker panels has not been estimated. This is a major limitation since, depending on the magnitude of effect, it is possible that a valid surrogate might already exist and could not be differentiated from an inadequate test as long as the liver biopsy result is the comparator. In other words, biopsy error could make it impossible to distinguish a perfect and clinically inadequate surrogate. To estimate the magnitude of the bias, we characterized the optimum performance of surrogate markers based on a range of conservative estimates of biopsy error.

METHODS

Because the formulae that characterize the diagnostic validity of a surrogate marker and the gold standard contain common terms, the degree to which error in the biopsy affects the expected performance of a marker (when the biopsy is used as a gold standard) can be directly calculated. We quantified the expected performance of a surrogate marker compared to liver biopsy as the area under the ROC curve (AUROC). The AUROC is simply a plot of the conditional probability of the sensitivity of the marker vs. biopsy at a specific marker cut-off, c , versus 1-specificity of marker vs. biopsy at that same cut-off.

The formulae in Table 1 illustrate how the expected sensitivity and specificity of a marker compared to liver biopsy can be calculated when three components are known: i) the values of sensitivity and specificity of the biopsy vs. true disease, ii) sensitivity and specificity of the marker vs. true disease, and iii) the prevalence of true disease are 'known'. Given that the distribution of values obtained from all surrogate marker panels take on a continuous distribution, one value for sensitivity and specificity with respect to either the true state of disease or the biopsy does not take into account the full range of variability. The continuous distribution requires that a range of disease cut-offs be defined, rather than a single result. Sensitivity and specificity can then be calculated at each of these cut-offs. These formulae then can be used to calculate sensitivity and specificity of the marker panel vs. biopsy at a given cutoff c using Bayes' rule, assuming that the value of the surrogate marker panel and the result of the biopsy are independent of the true stage of liver disease. For example, one could calculate sensitivity and specificity of an alanine aminotransferase (ALT) level of 40 IU/L for detection of significant liver disease in a setting where it is found in 40% of patients, and then repeat that for all other possible ALT cut-offs.

These formulae allowed us to consider hypothetical or 'known' values of sensitivity and specificity of a marker (vs. true disease) for a particular cut-off. For example, we could assume

for illustration that the ALT of 40 IU/ml actually has a sensitivity of 0.95 and a specificity of 0.7 and then calculate how accurately it would appear to be in a population with a specified disease prevalence (e.g., 30%) when compared to biopsy based on the accuracy of the biopsy itself (e.g., sensitivity and specificity of biopsy vs. true disease of 0.85 and 0.90 respectively). In this illustration, the marker would appear to have 81% sensitivity and 66% specificity.

In our calculations, we used a full range of cut-off values from negative infinity to positive infinity. For simplicity, we represent this full range of sensitivity and specificity (for all values of c) of the marker vs true disease by plotting the AUROC of the marker panel vs. true disease (instead of all of the values for the different combinations of sensitivity and specificity, figure 1). Similarly, we have represented the expected sensitivity and specificity of the marker panel vs. liver biopsy at each cut-off through the expected AUROC of the marker panel vs. biopsy (figure 2).

The values chosen for the components of the formulae: sensitivity and specificity of biopsy vs. true disease, AUROC for marker vs. true disease and prevalence of true disease were chosen to represent reasonable estimates determined by literature review and interviews with expert clinicians. Given its importance in clinical practice, the focus was on measurement of 'significant' or portal fibrosis (metavir 2-4) (1,3,26). As prevalence of significant liver fibrosis varies in each population, we represented a full range (10-50%) to correspond with the exiting literature.(6-12,14) It is noteworthy, that cirrhosis prevalence is an alternate measure that could similarly be developed. Since this is meant to be a surrogate medical test, we considered high degrees of actual marker AUROC values up to 1 (a perfect test). Because there is no true gold standard against which to compare biopsy, we represented biopsy validity by its sensitivity and specificity, imputed from sources of error, which were represented across 'highest' and generally-achievable ranges (27).

Two categories of error were considered for the liver biopsy: sampling and observer. Since liver fibrosis is not necessarily uniform, sampling error depends on the location and size of the biopsy. In one study in which 124 persons had simultaneous, laparoscopic needle biopsies of the right and left lobes of the liver, discordant classification of significant fibrosis occurred in 12 (10%) of patients (28). A number of other studies suggest that biopsy samples with more visible portal tracts yield more accurate and repeatable fibrosis readings (16,17). In one study, >3 cm biopsy sections were read and then reduced in size and re-read by the same pathologist. Overall, 19 (12%) of 161 biopsies were discordant in detecting significant fibrosis. Observer error is influenced by the skill and experience of the pathologist. Studies have suggested inter-observer agreement of ~85% and intraobserver agreement ~90% for the classification of significant fibrosis versus no fibrosis (27,29).

Because no study directly measures the sum of measurement and observer error and because both naturally vary from study to study, we performed our calculations across a range of biopsy sensitivity and specificity (versus truth) taking into account all components of error. Shown in this paper are high biopsy sensitivities and specificities of 80% to 90%.

RESULTS

The results of this investigation confirm the hypothesis that biopsy error causes the true validity of surrogate tests to be underestimated by an amount that would make a clinician falsely misperceive the test as inaccurate. Even with conservative estimates of biopsy error such as sensitivity and specificity of biopsy of 80%, true liver disease prevalence of 40%, and marker vs. true disease AUROC of 0.80, the calculated AUROC of the marker vs. biopsy would be 0.70 (Figure 2). For the same assumptions of disease prevalence and biopsy sensitivity and specificity, a perfect test (AUROC of marker vs. true disease of 0.99) would have an expected

validity (AUROC of marker vs. biopsy) of 0.76. If the biopsy sensitivity and specificity were 90% and disease prevalence remained 40%, a perfect marker would have an expected AUROC of 0.90. Interestingly, observed AUROC values of the marker vs. biopsy for many published studies fall within the range of 0.76 to 0.88 (6-12,14).

These data also imply that a marker panel with an observed AUROC as compared with the liver biopsy at the lower bound of 0.76 may truly have an AUROC (vs. true disease) between 0.93 and 0.99 under a sensitivity and specificity of biopsy of 80% and prevalence between 0.3 and 0.5. When the sensitivity and specificity of biopsy are 90%, the marker vs. true disease AUROC would be 0.83, thus still exceeding the observed AUROC of 0.76 (when prevalence is 0.5).

DISCUSSION

The results of this investigation demonstrate that even a perfect non-invasive marker could not be distinguished from less reliable assays with most tenable assumptions of biopsy sensitivity and specificity. In addition, our findings explain why existing published marker validity estimates cluster in an AUROC range of 0.76-0.88 (6-12,14). Moreover, the maximal expected real world performance of the surrogate marker occurred when the disease prevalence exceeded 40% and the sensitivity and specificity of the biopsy exceeded 90%, which is not feasible in most settings.

These calculations have implications for the interpretation of the performance of surrogate markers as well as their application in clinical practice. A perfect surrogate marker of liver fibrosis could already exist but not be recognized. Alternatively, correlated error (identifying the same false-positive and negative results using the biopsy and marker) could be misinterpreted as an improvement in observed validity of the marker. Since markers are developed by using biopsy data, the latter consideration is especially germane and probably already occurs.

Accumulating evidence regarding the limitations of biopsy have led some to suggest that non-invasive markers should replace biopsy as the initial method for disease staging (30-33). However, guidelines and practice patterns differ between countries and even within a given country. Further research is needed to evaluate the long-term effectiveness of these strategies before a global recommendation can be made. Others have considered alternate strategies where both non-invasive markers and biopsy are used in combination since complementary information can be obtained (33). Further research is needed to evaluate the long-term effectiveness of these strategies before a global recommendation can be made.

In this study, we considered measurement of significant liver fibrosis in our calculations. Other thresholds exist, such as detection of cirrhosis or 'no' versus 'some' fibrosis. We chose significant fibrosis to correspond with treatment guidelines and many published studies (1, 26). Most studies suggest that the measurement and observer error for detection of cirrhosis is lower (16,28). This may explain why markers often appear to be more valid representations of this stage (6). Further, our calculations did not consider the full range of fibrosis stage. As described previously, the underlying spectrum of disease represented by a dichotomization into significant liver fibrosis vs. not can be quite broad (18,34). It is likely that surrogate markers would perform better against a liver biopsy when the extremes are overrepresented (e.g., high representation of F0 and F4). Though we did not address this issue specifically, our calculations can be extended to comparisons of adjacent (e.g., F1 vs. F2) or nonadjacent stages of fibrosis (e.g., F1 vs. F4) to address this concern.

The calculations presented within this paper further rely on the assumption of conditional independence of the surrogate marker and biopsy results. We recognize that there have been

several recent demonstrations of non-parametric approaches to estimate ROC curves (35,36) as well as a latent class model approach (37). However, our goal was to illustrate why previous results for the AUROC that have not utilized specialized methods to correct for imperfect gold standards find limited AUROC estimates. Furthermore, the discrepant resolution method requires an imperfect standard test plus an additional method to resolve discrepancies and the composite reference standards method requires several imperfect reference tests that may be combined together to which the surrogate markers may be compared against (35,36,38). These methods may be useful in future studies that consider samples where biopsy measurements, elastography data and serum marker data are available.

Finally, we have not addressed the issue of discordance between biopsy results and surrogate markers. Even studies that observe high AUROC values have a large number of patients with discrepant biopsy and surrogate marker results. Interestingly, these studies often suggest that when there are differences between the two methods, biopsy has underestimated disease (28). This is not surprising given that liver biopsy is more likely to miss fibrosis when it is actually present as opposed to the reader overestimating the presence of fibrosis. Further, some non-invasive marker (e.g. APRI) levels tend to be higher when the Fibroscan estimates a higher disease burden but the biopsy suggests a low disease stage (33).

Our results emphasize the importance of minimizing biopsy error in studies developing surrogate markers. Since measurement error increases markedly when biopsy size is less than 3.0 cm, one application is that only such samples be used to characterize marker validity (16, 17). Likewise, future studies should make every effort to minimize reader error. Absent another gold standard, we cannot assess with confidence whether it is even possible to increase biopsy validity sufficiently to substantively differentiate a new marker from those we already have. However, these calculations make it clear that attempts to validate markers in 'real world' settings will always be constrained since biopsy sensitivity and specificity is much lower.

Although some clinicians already use liver biopsy surrogate markers in their practices, others are waiting for more valid tests. Our results strongly suggest that major improvements in surrogate markers are unlikely when evaluated against liver biopsy. Thus, novel strategies are needed to move the field forward. In particular, long-term prospective studies of markers against clinical gold standards, such as development of end-stage liver disease are needed to assess the best measures of intermediate disease stages. Likewise, the validity of all outcome measures must be carefully considered when assessing the validity of surrogate markers in biomedical research or clinical practice.

Acknowledgements

The authors acknowledge Maria Guido and John McHutchison for sharing relevant data.

References

1. Strader DB, Wright T, Thomas DL, Seeff LB. Diagnosis, management, and treatment of hepatitis C. *Hepatology* 2004;39:1147–1171. [PubMed: 15057920]
2. Lok AS, McMahon BJ. Chronic hepatitis B: update of recommendations. *Hepatology* 2004;39:857–861. [PubMed: 14999707]
3. *J Hepatol*; Proceedings of the European Association for the Study of the Liver (EASL) International Consensus Conference on Hepatitis B; September 14-16, 2002; Geneva, Switzerland. 2003. p. S1-S235.
4. Poynard T, Ratziu V, Bedossa P. Appropriateness of liver biopsy. *Can J Gastroenterol* 2000;14:543–548. [PubMed: 10888734]
5. Bravo AA, Sheth SG, Chopra S. Liver biopsy. *N Engl J Med* 2001;344:495–500. [PubMed: 11172192]

6. 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;357:1069–1075. [PubMed: 11297957]
7. Patel K, Gordon SC, Jacobson I, Hezode C, Oh E, Smith KM, et al. Evaluation of a panel of non-invasive serum markers to differentiate mild from moderate-to-advanced liver fibrosis in chronic hepatitis C patients. *J Hepatol* 2004;41:935–942. [PubMed: 15582126]
8. Sud A, Hui JM, Farrell GC, Bandara P, Kench JG, Fung C, et al. Improved prediction of fibrosis in chronic hepatitis C using measures of insulin resistance in a probability index. *Hepatology* 2004;39:1239–1247. [PubMed: 15122752]
9. Forns X, Ampurdanes S, Llovet JM, Aponte J, Quinto L, Martinez-Bauer E, et al. Identification of chronic hepatitis C patients without hepatic fibrosis by a simple predictive model. *Hepatology* 2002;36:986–992. [PubMed: 12297848]
10. Wai CT, Greenson JK, Fontana RJ, Kalbfleisch JD, Marrero JA, Conjeevaram HS, et al. A simple non-invasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. *Hepatology* 2003;38:518–526. [PubMed: 12883497]
11. Kelleher TB, Mehta SH, Bhaskar R, Sulkowski M, Astemborski J, Thomas DL, et al. Prediction of hepatic fibrosis in HIV/HCV co-infected patients using serum fibrosis markers: the SHASTA index. *J Hepatol* 2005;43:78–84. [PubMed: 15894397]
12. Leroy V, Monier F, Bottari S, Trocme C, Sturm N, Hilleret MN, et al. Circulating matrix metalloproteinases 1, 2, 9 and their inhibitors TIMP-1 and TIMP-2 as serum markers of liver fibrosis in patients with chronic hepatitis C: comparison with PIIINP and hyaluronic acid. *Am J Gastroenterol* 2004;99:271–279. [PubMed: 15046217]
13. Foucher J, Chanteloup E, Vergniol J, Castera L, Le Bail B, Adhoute X, et al. Diagnosis of cirrhosis by transient elastography (FibroScan): a prospective study. *Gut* 2006;55:403–408. [PubMed: 16020491]
14. Saadeh S, Cammell G, Carey WD, Younossi Z, Barnes D, Easley K. The role of liver biopsy in chronic hepatitis C. *Hepatology* 2001;33:196–200. [PubMed: 11124836]
15. Shaheen AA, Wan AF, Myers RP. FibroTest and FibroScan for the prediction of hepatitis C-related fibrosis: a systematic review of diagnostic test accuracy. *Am J Gastroenterol* 2007;102:2589–2600. [PubMed: 17850410]
16. Bedossa P, Dargere D, Paradis V. Sampling variability of liver fibrosis in chronic hepatitis C. *Hepatology* 2003;38:1449–1457. [PubMed: 14647056]
17. Colloredo G, Guido M, Sonzogni A, Leandro G. Impact of liver biopsy size on histological evaluation of chronic viral hepatitis: the smaller the sample, the milder the disease. *J Hepatol* 2003;39:239–244. [PubMed: 12873821]
18. 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;25:733–739. [PubMed: 17311607]
19. Poynard T, Munteanu M, Imbert-Bismut F, Charlotte F, Thabut D, Le Calvez S, et al. Prospective analysis of discordant results between biochemical markers and biopsy in patients with chronic hepatitis C. *Clin Chem* 2004;50:1344–1355. [PubMed: 15192028]
20. Dienstag JL. The natural history of chronic hepatitis C and what we should do about it. *Gastroenterology* 1997;112:651–655. [PubMed: 9024319]
21. Afdhal NH. Biopsy or biomarkers: is there a gold standard for diagnosis of liver fibrosis? *Clin Chem* 2004;50:1299–1300. [PubMed: 15277345]
22. Zeremski M, Talal AH. Non-invasive markers of hepatic fibrosis: are they ready for prime time in the management of HIV/HCV co-infected patients? *J Hepatol* 2005;43:2–5. [PubMed: 15922482]
23. Valenstein PN. Evaluating diagnostic tests with imperfect standards. *Am J Clin Pathol* 1990;93:252–258. [PubMed: 2405632]
24. Phelps CE, Hutson A. Estimating diagnostic test accuracy using a “fuzzy gold standard”. *Med Decis Making* 1995;15:44–57. [PubMed: 7898298]
25. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. *J Clin Epidemiol* 1999;52:943–951. [PubMed: 10513757]

26. NIH Consensus Statement on Management of Hepatitis C: 2002. NIH Consensus State Sci Statements 2002;19:1–46.
27. Intraobserver and inter-observer variations in liver biopsy interpretation in patients with chronic hepatitis C. The French METAVIR Cooperative Study Group. Hepatology 1994;20:15–20. [PubMed: 8020885]
28. Regev A, Berho M, Jeffers LJ, Milikowski C, Molina EG, Pyrsopoulos NT, et al. Sampling error and intraobserver variation in liver biopsy in patients with chronic HCV infection. Am J Gastroenterol 2002;97:2614–2618. [PubMed: 12385448]
29. Bedossa P, Poynard T. An algorithm for the grading of activity in chronic hepatitis C. The METAVIR Cooperative Study Group. Hepatology 1996;24:289–293. [PubMed: 8690394]
30. Poynard T, Ratziu V, Benhamou Y, Thabut D, Moussalli J. Biomarkers as a first-line estimate of injury in chronic liver diseases: time for a moratorium on liver biopsy? Gastroenterology 2005;128:1146–1148. [PubMed: 15825108]
31. Castera L, Denis J, Babany G, Roudot-Thoraval F. Evolving practices of non-invasive markers of liver fibrosis in patients with chronic hepatitis C in France: time for new guidelines? J Hepatol 2007;46:528–529. [PubMed: 17239479]
32. La Haute Autorite de Sante (HAS). The HAS recommendations for the management of the chronic hepatitis C using non-invasive biomarkers. 2007 [17, May 2008]. Available from: URL: http://www.has-sante.fr/portail/display.jsp?id=c_476504
33. Castera L, Vergniol J, Foucher J, Le Bail B, Chanteloup E, Haaser M, et al. Prospective comparison of transient elastography, Fibrotest, APRI, and liver biopsy for the assessment of fibrosis in chronic hepatitis C. Gastroenterology 2005;128:343–350. [PubMed: 15685546]
34. 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;53:1615–1622. [PubMed: 17634213]
35. Zhou XH, Castelluccio P, Zhou C. Nonparametric estimation of ROC curves in the absence of a gold standard. Biometrics 2005;61:600–609. [PubMed: 16011710]
36. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Stat Med 1999;18:2987–3003. [PubMed: 10544302]
37. Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J Clin Epidemiol 1988;41:923–937. [PubMed: 3054000]
38. Hall P, Zhou XH. Nonparametric estimation of component distributions in a multivariate mixture. Ann Statist 2003;31:201–224.

Figure 1.

Family of receiver operating characteristic (ROC) curves and area under the ROC curve (AUROC) values of a surrogate marker vs. true disease used in the estimation of the validity (AUROC) of the surrogate marker vs. the liver biopsy. These AUROC values represent the full range of sensitivity and specificity of a surrogate marker vs. true disease at an infinite number of cut-offs for defining true disease.

Figure 2.

The expected performance of a surrogate marker of a liver biopsy is shown as the area under the receiver operating characteristic curve (AUROC) and depicted in each graph in a third dimension (y2 axis) across a gradient of lowest (red) to highest (yellow) according to increasing prevalence of significant liver disease (metavir ≥ 2 , y axis) and increasing validity of the surrogate marker as measured by the AUROC of the marker vs. true disease (x axis). Since the actual validity of liver biopsy itself is unknown, we provide 4 representations, graphs a-d, in which the sensitivity and specificity of the biopsy vary (80-90 %). Hashed lines demark the AUROC values of markers (compared to biopsy) that have been reported in previous studies. (6-12,14) Black shading represents the set of conditions under which the AUROC values exceed what has already been observed, which by our calculations should only occur if the biopsy is $\geq 90\%$ sensitive and specific and the surrogate marker has near perfect accuracy, graph d, upper right. If biopsy sensitivity and specificity are lower, even a perfect surrogate marker will have AUROC versus liver biopsy in the range of what has already been reported.

Table 1

Formulae for calculating sensitivity and 1-specificity of a marker panel vs. the liver biopsy according to the true stage of disease and true validity of the marker panel and liver biopsy.*

Sensitivity of a marker panel vs. liver biopsy at cut-off c

$$\text{prob}[M > c \mid B+] = \frac{\text{prob}[M > c \mid D+] \mu_{B:D} p_D + \text{prob}[M > c \mid D-] (1 - \gamma_{B:D}) (1 - p_D)}{\mu_{B:D} p_D + (1 - \gamma_{B:D}) (1 - p_D)}$$

1-specificity of a marker panel vs. liver biopsy at cut-off c

$$\text{prop}[M > c \mid B-] = \frac{\text{prop}[M > c \mid D+] (1 - \mu_{B:D}) p_D + \text{prop}[M > c \mid D-] \gamma_{B:D} (1 - p_D)}{(1 - \mu_{B:D}) p_D + \gamma_{B:D} (1 - p_D)}$$

* D+ and D- denote the true stage of disease; B+ and B- indicate biopsy diagnosed disease; M denotes the value on a continuous scale of a marker; $\mu_{B:D}$ and $\gamma_{B:D}$ are the sensitivity and specificity of the biopsy relative to true disease, respectively; p_D is the prevalence of true disease; $\text{prob}[M > c \mid D+]$ and $\text{prob}[M > c \mid D-]$ are the sensitivity of marker vs. true disease at cut-off c and 1-specificity of marker vs. true disease at cut-off c ; For example, for a perfect marker panel where the sensitivity and specificity with respect to true disease are 100% (AUROC = 1), a disease prevalence of 40% and the sensitivity and specificity of the biopsy vs. true disease both equal to 90%, the observed sensitivity and specificity of the marker panel vs. biopsy will be equal to 0.86 and 0.93, respectively.