The International Journal of Biostatistics

Volume 5, *Issue* 1 2009 *Article* 5

Estimating Complex Multi-State Misclassification Rates for Biopsy-Measured Liver Fibrosis in Patients with Hepatitis C

Peter Bacchetti* Ross Boylan†

*University of California, San Francisco, peter@biostat.ucsf.edu

†University of California, San Francisco, ross@biostat.ucsf.edu

Estimating Complex Multi-State Misclassification Rates for Biopsy-Measured Liver Fibrosis in Patients with Hepatitis C*

Peter Bacchetti and Ross Boylan

Abstract

For both clinical and research purposes, biopsies are used to classify liver damage known as fibrosis on an ordinal multi-state scale ranging from no damage to cirrhosis. Misclassification can arise from reading error (misreading of a specimen) or sampling error (the specimen does not accurately represent the liver). Studies of biopsy accuracy have not attempted to synthesize these two sources of error or to estimate actual misclassification rates from either source. Using data from two studies of reading error and two of sampling error, we find surprisingly large possible misclassification rates, including a greater than 50% chance of misclassification for one intermediate stage of fibrosis. We find that some readers tend to misclassify consistently low or consistently high, and some specimens tend to be misclassified low while others tend to be misclassified high. Non-invasive measures of liver fibrosis have generally been evaluated by comparison to simultaneous biopsy results, but biopsy appears to be too unreliable to be considered a gold standard. Non-invasive measures may therefore be more useful than such comparisons suggest. Both stochastic uncertainty and uncertainty about our model assumptions appear to be substantial. Improved studies of biopsy accuracy would include large numbers of both readers and specimens, greater effort to reduce or eliminate reading error in studies of sampling error, and careful estimation of misclassification rates rather than less useful quantities such as kappa statistics.

KEYWORDS: fibrosis, hepatitis C, kappa statistic, latent variables, misclassification

---

*This work was supported by grant R01AI069952 from the United States National Institutes of Health. We thank Hagen Blaszyk and Paul Calès for providing raw data.

1 Introduction

Liver biopsies play a prominent role in the clinical care of patients with various liver diseases (Manning and Afdhal, 2008), notably hepatitis C virus (HCV) infection. Pathologists typically rate the stage of liver fibrosis in biopsy specimens on an ordinal scale that ranges from no damage to cirrhosis (Batts and Ludwig, 1995; Bedossa et al., 1994). Such ratings are also used in research concerning progression of HCV disease, and multi-state modeling (Gentleman, et al., 1994; Kalbfleisch and Lawless, 1985; Kay, 1986) is perhaps the most appropriate statistical approach for such research (Deuffic-Burban, Poynard and Valleron, 2002; Terrault et al., 2008). These methods can estimate misclassification probabilities in addition to the parameters governing transition rates between states (Jackson and Sharples, 2002; Jackson et al., 2003; Satten and Longini, 1996), but the data available for multi-state modeling typically provide only indirect information about misclassification rates.

Studies focused specifically on misclassification may provide better estimates of misclassification rates. In the case of HCV, misclassification of fibrosis stage arises from *reading error*, where the stage of the specimen is misclassified by a pathologist, and from *sampling error*, which refers to sampling of the liver and means that the biopsy specimen does not accurately represent the true stage of the liver as a whole. Hepatologists think of the true stage of the liver as being the stage of the most diseased part, as reflected by description of disagreements between the stages of two specimens as "understaging" or "underdiagnosis" (Regev, et al., 2002; Skripenova, et al., 2007). There are studies that focused specifically on reading error via multiple readings of the same liver biopsy specimens and studies that focused specifically on sampling error via examination of two specimens from the same liver. Existing studies, however, have not combined the two approaches, have not been analyzed in detail, and do not attempt to estimate the overall misclassification rates that would be most relevant and interpretable for both clinical use and research—they instead provide only simple tabulations and kappa statistics, which are not directly useful. We therefore present here an analysis of data from four studies, along with overall estimated misclassification rates.

2 Data Sources

We obtained data from four studies that focused on patients with HCV, used methods that rate fibrosis from 0 (no fibrosis) to 4 (cirrhosis), and provided usable data. We denote two studies of reading error as R1 (Rousselet et al., 2005) and R2 (Netto et al., 2006) and two studies of sampling error as S1 (Skripenova et al., 2007) and S2 (Regev et al., 2002).

From study R1 we utilize the data from their substudy 1B, which had 157 liver biopsy specimens staged using the Metavir system (Bedossa et al., 1994) by both a junior and a senior expert pathologist, with consensus reached in a second, common reading. We treat the consensus reading as the true fibrosis stage; this is optimistic whenever both pathologists were wrong, and so may tend to understate reading error. The specimens were from chronic hepatitis C patients who had not been treated with antiviral or antifibrotic drugs. Eleven specimens had true stage F0, 55 had stage F1, 48 had stage F2, 16 had stage F3, and 27 had stage F4. Appendix Table A1 provides the raw data.

Study R2 reports 17 readings on each of 6 specimens, staged according to the Batts and Ludwig schema (Batts and Ludwig, 1995). Five of the specimens were from post-transplant recurrences of HCV and one from chronic HCV infection. One reader was the central pathologist for a multisite clinical trial, and the other 16 were local pathologists from 13 of the centers participating in the trial. We used the majority of the readings to define the "true" stage for each specimen, which minimizes estimated misclassification rates and may therefore be optimistic. The numbers of correct readings for the six specimens by this definition were 8, 9, 10, 10, 11, and 14. The one with only 8 correct had 8 stage F0, 8 stage F1, and 1 stage F2, so we assumed that the true stage was F1. In one case, the stage from 10 of the 16 local pathogists did not agree with the central pathologist's stage. Raw data are in the original publication (Netto et al., 2006).

Study S1 examined left and right lobe pairs of liver biopsy specimens from 60 patients with chronic hepatitis C. These were staged by one pathologist, blinded to the pairings, using the Batts and Ludwig schema. Re-readings were not used to reduce or eliminate potential reading error in the paired scores that were analyzed, but an intraobserver agreement rate of 106/120 (88%) resulted from re-readings two weeks after the original readings. The left and right stages were equal for 42 (70%) of the pairs and differed by one point for the other 18. Appendix Table A2 provides the raw data. Notably, there were no readings of stage F0 and only 5 of F1, with only one liver read as stage F1 on both sides.

Study S2 staged left-right pairs of biopsies from 124 patients with chronic hepatitis C. Using the Batts and Ludwig schema, one experienced pathologist scored 50 pairs and another scored the other 74 pairs. Re-readings were not used to reduce or eliminate potential reading error in the paired scores that were analyzed, but intraobserver agreement rates of 48/50 (96%) and 47/50 (94%) resulted from re-readings of some specimens 3-4 months after the original readings. Raw data are not available for this study, but a variety of summaries given in the original publication permit the analysis described in Section 3.3. The left and right stages were equal for 83 (67%) of the pairs, differed by 1 point for 38 (31%), and differed by 2 points for 3 (2%).

DOI: 10.2202/1557-4679.1139

3 Estimation Methods and Results

We focus here on methods that explicitly assess and estimate variation between readers and between specimens, using either fixed or random effects. This is for two main reasons. First, left-right disagreements in the sampling studies can result not only from sampling error but also from reading error on the samples. Our estimates of reading error will therefore serve as important inputs into estimation of sampling error rates. Because left and right samples from the same patient were always read by the same reader, individual-specific reading error rates will be needed, and marginal rates would be inappropriate. Second, our models will allow direct estimation and concrete illustration of reader-to-reader variation, which is potentially important. Where marginal estimates are needed, we generate these from random effects models.

3.1 Reading Error Methods

We wish to model the quantity

$r_{jkuv} = \Pr\{\text{specimen } k \text{ read as stage } v \mid \text{true stage } u, \text{ reader } j\}$.

In addition to depending on u and v, this quantity may also depend on reader effects and specimen effects. We suppose that each reader may have a bias toward tending to read specimens too high or too low, reflected by a reader effect β_j. We also suppose that the skill of readers may vary so that some tend to be more accurate or less accurate than others; this is reflected by a random effect γ_j. Finally, we allow for a specimen effect, σ_k, that allows some specimens to be "borderline" in the sense of being much easier to read too high or too low than others. This could reflect the discretization of an underlying continuous disease process. Some specimens are near the bottom of the range of continuous values for their stage, so they are more likely to be read as a lower stage, while other specimens may be near the top, in which case they are more likely to be read as a higher stage. In principle, there could also be specimen effects to allow for some specimens to simply be harder to read than others, but we have not included this here because the reading errors in study R2 are clearly directional (rather than some specimens showing greater symmetric spread than others), while study R1 provides little information on specimen effects, as discussed in the next section. Our multinomial model for reading probabilities is

$$r_{jkuv} = \exp(\eta_{jkuv}) / \sum_w \exp(\eta_{jkuw}) \ , \qquad (1)$$

where

$$\eta_{jkuv} = \begin{cases} 0 & u = v \\ \alpha_{uv} + \text{sgn}(v-u)\beta_j + \gamma_j + \text{sgn}(v-u)\sigma_k & u \neq v \end{cases}. \quad (2)$$

The intercept parameters α_{uv} are unconstrained, and sgn(v-u) is the signum function equal to 0 if u=v, 1 if v>u, and -1 if v<u. This typical multinomial framework has larger values of η_{jkuv} for $u \neq v$ corresponding to greater chances of incorrect readings and the numerator in (1) always equal to unity for correct readings. Although more constrained models have been used for modeling misclassification with sparse ordinal data (Albert, Hunsberger and Biro, 1997; Mwalili, Lesaffre and Declerck, 2008), this is not necessary with our data. With a large number of readers, the β_j and γ_j can be treated as random effects generated from a joint distribution function $G(\beta, \gamma)$ that is a bivariate normal with mean (0,0) and covariance matrix $\mathbf{V} = \begin{bmatrix} V_\beta & V_{\beta\gamma} \\ V_{\beta\gamma} & V_\gamma \end{bmatrix}$. We choose this form because it can be handled by the NLMIXED procedure in the SAS statistical package (SAS Institute, Cary, NC, USA). We treat specimen effects as fixed. Alternatively, with few readers and many specimens, the σ_k can be modeled as random effects with reader effects fixed. We were not able to model both reader and specimen effects as random because no software was readily available for fitting multinomial models with crossed random effects.

For the model with random reader effects, we estimate the parameters, $\{\alpha_{uv}\}$, $\mathbf{V}$, and $\{\sigma_k\}$, by maximum likelihood, using the general likelihood feature of the SAS NLMIXED procedure (SAS Institute, Cary, NC). The likelihood is

$$L_R = \prod_j \int \left(\prod_k r_{jku_k v_{jk}} \right) dG(\beta, \gamma),$$

where the inner product is over all k read by reader j, the outer product is over all readers in the study, u_k is the true stage for specimen k, and v_{jk} is the stage assigned to specimen k by reader j. This assumes that all readings are independent given the reader and specimen effects. We use similar methods for the case with fixed reader effects and random specimen effects. Appendix 2 provides example code for fitting a reading error model.

3.2 Reading Error Estimates

Study R1 had 96 specimens where the two readers agreed (and therefore agreed with the "true" consensus stage by definition), 35 specimens where the junior reader misclassified the stage, and 27 specimens where the senior reader misclassified the stage. We fit a model with fixed β and γ, specified so the senior

DOI: 10.2202/1557-4679.1139

reader has $+\beta$ and the junior $-\beta$, and similarly for γ. We used a random specimen effect assumed to follow a normal distribution with mean 0 and variance V_σ. The estimates (95% confidence intervals) were $\hat{\beta}$ = 0.67 (0.33 to 1.00), $\hat{\gamma}$ = -0.30 (-0.66 to 0.05), and $\hat{V}_\sigma$ = 0. The apparent lack of any specimen effects, however, may be due to the procedures used in study R1. The consensus process may have only been applied in cases of initial disagreement, and it never produced a consensus (which we take as the true state) that was below or above both initial readings. With one exception, the consensus value was always equal to one of the readers' original choices, and the exception was original readings of 2 and 4 that produced a consensus of 3. Thus, it may not have been possible to detect the type of cases that would be most indicative of specimen effects, where both readers were too low or too high.

Because the data do not appear to permit estimates of specimen effects, and because the estimated $\hat{\gamma}$ does not achieve 5% statistical significance, we estimated a simpler model without those terms. This gave $\hat{\beta}$ = 0.61 (0.31, 0.91), and Appendix Table A3 shows the estimated $\hat{\alpha}_{uv}$. Combinations of u and v that never occurred in the data are estimated to have probability zero ($\hat{\alpha}_{uv}$ = -∞).

Study R2 is limited by having only true stages 1, 2, and 3 represented (each by 2 specimens), so we have mainly used it in conjunction with study R1. For study R2 alone, we fit a model with random β and γ and fixed σ_k for one specimen in each of the 3 true stages, and then we fit various simpler models. This estimated $\hat{V}_\beta$ = 1.71 and $\hat{V}_\gamma$ = 0.58. A likelihood ratio test for V_β = 0 (Stram and Lee, 1994) produced p=0.0012, for V_γ=0 produced p=0.12, and for no specimen effects versus 3 effects produced p<0.0001. The results contrast with R1 in suggesting the possibility of important specimen effects, and the design of R2 is better able to show such effects, if they exist, because of the large number of readers for each specimen.

To model both R1 and R2 together, we included a random β along with fixed specimen effects σ_k for each of the specimens in R2 (6 parameters). (Also including random γ did not reach statistical significance by likelihood ratio test, p=0.26, so we focus on this more parsimonious model for simplicity.) The estimated $\hat{V}_\beta$ is 0.98, and a likelihood ratio test for V_β=0 produces p<0.0001. If one outlying reader from study R2 is excluded from the analysis, then the estimated $\hat{V}_\beta$ drops to 0.52 (p=0.0001 for V_β=0). The estimated specimen effects in the model with all readers were -2.6, -2.4, -1.2, -1.0, 0.4, and 2.8. A likelihood ratio test for all 6 specimen effects being zero produced p<0.0001.

Our assumed distribution of specimen effects for subsequent use in modeling sampling errors required some additional consideration and analysis. If specimen effects arise from discretization of an underlying continuous disease process, then the distribution of specimen effects might be expected to be symmetric. In addition, computational issues in modeling sampling error necessitated use of a very simple form for the distribution of specimen effects. We modeled the R1 and R2 data together using fixed reader effects and a random specimen effect that only pertained to specimens in R2, obtaining an estimated normal distribution of specimen effects with mean zero and standard deviation 1.76. We divided this normal distribution into thirds, and represented each third by the conditional expectation within that third. This produces a distribution of specimen effects equally likely to be -1.92, 0, or +1.92. We use this in all subsequent analyses.

Table 1. Fitted and tabulated classification rates (percentages) reflecting only reading error.

		Estimated percentage in each read stage, given each true stage*														
True stage u:		0		1			2			3				4		
Read stage v:		0	1	0	1	2	1	2	3	1	2	3	4	3	4	
Specimen effects[†]	β															
No	-1	**93**	7	15	**79**	6	20	**77**	2	6	40	**52**	1	16	**84**	
	0	**83**	17	5	**78**	16	8	**85**	7	3	20	**72**	5	6	**94**	
	+1	**65**	35	2	**63**	36	3	**79**	18	1	8	**76**	15	2	**98**	
	m[‡]	**81**	19	7	**73**	19	10	**80**	9	3	23	**67**	7	8	**92**	
Yes	-1	**86**	14	24	**63**	13	29	**64**	7	6	41	**48**	5	25	**75**	
	0	**74**	26	12	**62**	26	16	**69**	15	4	28	**56**	12	13	**87**	
	+1	**60**	40	5	**54**	41	7	**65**	27	2	16	**58**	24	6	**94**	
	m[‡]	**74**	26	14	**60**	27	17	**66**	16	4	28	**54**	14	15	**85**	
Raw tabulations across all readers and specimens:																
Study R1		**82**	18	6	**75**	18	9	**82**	8	3	22	**69**	6	7	**93**	
Study R2**				47	**50**	3	12	**74**	9	6	15	**59**	21			

* Values shown are $100\times \hat{r}_{j\cdot uv}$ as defined at the beginning of Section 3.1, where j corresponds to a reader with the β shown for that row, and the $\cdot$ subscript indicates averaging over specimen effects if present. Bold entries are the percentage correctly classified for each true stage.

[†] If present, specimen effects are assumed to be -1.92, 0, and +1.92 each with probability 1/3.

[‡] Marginal rates, averaged over readers with β= -1, β=0, and β= +1.

** Two observations (6%) with true stage 2 and read stage 0 not shown due to space constraints and no occurrence of this combination in Study R1.

DOI: 10.2202/1557-4679.1139

Table 1 shows estimated misclassification rates based on the $\hat{\alpha}_{uv}$ shown in Table A3, with either no specimen effects or the simplified specimen effect distribution described in the previous paragraph, and three different types of readers: low (β= -1), medium (β= 0), and high (β= +1). The marginal rates are the unweighted average over the three types of readers; averaging instead over a normal distribution of β's with mean 0 and variance $\hat{V}_{\beta}$=0.98 produced similar rates.

Appendix Table A4 shows confidence intervals for the estimates and describes the method for obtaining them. For the fourth row of Table 1, upper 95% confidence bounds on the percentage read correctly for stages 0 to 4 are 92%, 81%, 87%, 79%, and 97%. For the row with specimen effects that is marginal over β , the upper confidence bounds on correct classification are 86%, 69%, 76%, 68%, and 93%.

3.3 Sampling Error Methods

Because studies S1 and S2 did not attempt to determine the true stage of each specimen, left-right disagreements can arise from reading error even if the true stages of the specimens agree. To estimate sampling error, we must therefore calculate the likelihood of the pattern of left and right observed stages in terms of both reading error and sampling error parameters. Because we do not know the true stage of each person's liver, there are also nuisance parameters for the prevalence of true states in the study. For a given patient and a given study, define

o_{vw} = Pr{observe left stage v, right stage w}
p_t = Pr{true state of liver is t}
s_{tu} = Pr{obtain specimen with true stage u | true stage of liver is t}.

The s_{tu} are the sampling error probabilities that we wish to estimate, and we assume that these are the same for left and right specimens and for all patients in a study. We also assume that s_{tu}=0 for $u>t$ and $u<t$-1 (generalizing to allow $s_{tu}>0$ for $u=t$-2 often made estimation more difficult even though the estimates ended up being infinitesimal). The assumed downward direction of all sampling errors reflects the idea that sampling error only arises due to missing the most diseased part of the liver. In order to deal with the paired data, let 1 index the left specimen and 2 the right specimen. We can then calculate

$$o_{vw} = \sum_t p_t \left(\sum_{u_1} \sum_{u_2} \sum_{\sigma_1} \sum_{\sigma_2} s_{tu_1} r_{j1u_1v} s_{tu_2} r_{j2u_2w} f(\sigma_1, \sigma_2 \mid u_1, u_2) \right), \quad (3)$$

with the reading probabilities r_{j1u_1v} and r_{j2u_2w} defined by equations (1) and (2). Here, $f(\sigma_1,\sigma_2 \mid u_1,u_2)$ is the probability of having specimen effects σ_1 and σ_2 given true states u_1 and u_2. Equation (3) simply adds the probabilities of all the possible combinations of true liver stages, stages of specimens from the liver, and readings of the specimens that produce observed stages v on the left and w on the right. We note that it assumes that the chance of sampling error is independent on the left and right of the same patient. This may reduce estimated sampling error rates compared to allowing for dependence, because it reduces the possibility that concordant pairs arise from both specimens' stages being lower than the liver's stage. We also assume that reading probabilities on the left and right are conditionally independent given the reader and specimen effects in (2).

We use multinomial models analogous to (1) for modeling p_t and s_{tu}:

$$p_t = \exp(\theta_t)/\sum_{\tau}\exp(\theta_\tau) \text{ , and}$$

$$s_{tu} = \exp(\lambda_{tu})/\sum_{w}\exp(\lambda_{tw}) \text{ .} \qquad (4)$$

The θ_t and λ_{tu} are not themselves modeled analogously to (2) but are instead the parameters of the models, with reference categories defined by setting $\theta_2 = 0$ and $\lambda_{tt} = 0$ for all t.

We evaluate three possible assumptions for $f(\sigma_1,\sigma_2 \mid u_1,u_2)$. We assume that the marginal distributions of σ_1 and σ_2 follow the discrete distribution described in the previous section, but they may be correlated. We incorporate estimation of this dependence into the sampling error estimation, which we denote as the *estimated dependence* case. We also evaluate an assumption of *complete independence*. Finally, we have allowed the distribution to depend on u_1 and u_2 in order to evaluate a biologically plausible exceptional case: that σ_1= +1.92 and σ_2= -1.92 whenever $u_1<u_2$, and vice versa. This assumes that a specimen with a true stage less than that of the liver as a whole will be near the top of the underlying continuous range for the specimen's stage, and that a liver capable of providing under-staged specimens will provide correctly-staged specimens that are near the bottom of the underlying continuous range for the liver's stage. We couple this assumption with the additional assumptions that $\sigma_1=\sigma_2$ with the marginal distribution from the previous section whenever $u_1=u_2=t$ and that $\sigma_1=\sigma_2$= +1.92 whenever $u_1=u_2<t$, in order to define the case we denote as *full dependence*. Note that the estimated dependence case does not include our full dependence case, due to the specialized assumption when $u_1\neq u_2$.

Equation (3) requires specification of reading error rates. The readings in study S1 were all done by a single reader, and those in S2 were done by only two readers, 50 pairs by one reader and 74 by the other (but there is no way of telling

DOI: 10.2202/1557-4679.1139

which were done by which). Use of marginal misclassification rates would therefore not be appropriate. Although allowing for a random reader bias β in the model would be possible in principle, this would be difficult due to the presence of random specimen effects as described in the previous paragraph. We therefore separately evaluate use of the low, medium, and high reader misclassification rates from Table 1. Both studies S1 and S2 provided information on intra-reader disagreement rates from re-readings, and these were lower than would be expected from the models of Table 1. We therefore also evaluated models that had an added reader accuracy effect γ in equation (2) chosen to produce an expected intra-reader disagreement rate that exactly matches S1's or S2's reported rate. These assumed the marginal distribution of true stages was equal to each study's reported distribution of left and right read stages combined.

Given a particular assumed reading error model and a particular assumption about dependence between the specimen effects, we estimate the parameters of the sampling error model (4) by maximum likelihood. Letting c_{vw} denote the number of patients in a given study who have stage v observed on the left and w on the right, and $\mathbf{c}$ denote the vector of all those counts, we have a multinomial likelihood

$$L_S(\mathbf{c}) = N(\mathbf{c}) \prod_{v,w} (o_{vw})^{c_{vw}} , \quad (5)$$

where $N(\mathbf{c})$ is the combinatorial term denoting the number of possible ways of dividing the total number of patients in the study into the cell counts c_{vw}.

We know $\mathbf{c}$ for study S1, but for study S2, the authors were not able to locate the original data and we only have partial information about $\mathbf{c}$, including: summaries of how many left-right pairs had $|v\text{-}w|$ equal to 0, 1, or 2; mean readings for left and right; the kappa statistic for left-right agreement; and various summaries of specific types of discordance such as stage 3 on one side and stage 4 on the other. For study S2, we therefore estimate the sampling error model by maximizing the likelihood of the reported information. Let C denote the set of all possible vectors $\mathbf{c}$ consistent with the provided information. The likelihood of the available information is the sum of the likelihoods of all the possible specific ways in which it could have arisen. We then have the likelihood for study S2

$$L_{S2} = \sum_{\mathbf{c} \in C} L_S(\mathbf{c}) . \quad (6)$$

Note that the combinatorial term in (5), while not needed for study S1, is important here because it reflects how many different ways each vector $\mathbf{c}$ could have arisen.

3.4 Sampling Error Results

The reading error rates from Table 1 predicted intra-reader disagreement rates that ranged from 2.6 to 4.5 times higher than the observed rates reported for studies S1 and S2. Estimation of sampling error rates using the unmodified reading distributions from Table 1 often produced implausible estimates, with very high or even certain estimated undersampling probabilities $\hat{s}_{tu}$ for some t, non-zero $\hat{p}_t$ for only 2 true liver states t, and/or most s_{tu} estimated to be zero, with little consistency in which were nonzero across different scenarios. Table 2 shows the estimated undersampling probabilities $\hat{s}_{tu}$ when the estimation uses modified distributions that are tuned (via addition of negative accuracy effects γ as described in the previous section) to match the observed intra-reader disagreement rates. Confidence intervals shown are from the Wald intervals around the $\hat{\lambda}$, except that a profile likelihood confidence bound is shown if $\hat{s}_{tu}=0$. For study S2, the estimates are based on maximizing likelihood (6) using 2342 vectors in the set C that are consistent with the reported information.

The estimates $\hat{s}_{32}$ of undersampling risk from livers with true stage 3 are high for both studies and many different possible assumptions. These estimates and others, however, have very wide confidence intervals. Our biologically-motivated full dependence assumption for the specimen effects does not fit as well as the other assumptions. We note that the zero estimates for $\hat{s}_{10}$ using S1 reflect the fact that no reading had stage 0 in that study. The profile likelihood confidence bound extends to 100% in all those cases because models with $\hat{s}_{10}=1$ worsen $-2l$ by less than the 3.84 worsening needed for a profile likelihood confidence bound. The zero estimates for $\hat{s}_{21}$ using study S2 do not have an obvious explanation, and one scenario instead has $\hat{s}_{10}=0$. There were only 90 vectors (4%) in C that had $c_{12}+c_{21}=0$. Similarly to the zero estimates in S1, upper confidence bounds are often 100%, because setting the probability to 1 worsened $-2l$ by less than 3.84.

DOI: 10.2202/1557-4679.1139

Table 2. Estimated sampling error rates for 18 combinations of data used, assumed reading effects β, and assumed dependence between specimen effects.

Study	β	Assumed dependence, σ_1 with σ_2	Sampling error estimate, % (95% confidence interval) $\hat{s}_{10}$	$\hat{s}_{21}$	$\hat{s}_{32}$	$\hat{s}_{43}$	$-2l$*
S1	-1	None	0 (0, 100)	3 (0, 31)	45 (21, 72)	14 (2, 54)	28.01
		Estimated	0 (0, 100)	3 (0, 28)	46 (22, 72)	14 (2, 53)	27.82
		Full	0 (0, 100)	2 (0, 41)	43 (19, 69)	13 (2, 54)	29.99
	0	None	0 (0, 100)	3 (0, 30)	42 (17, 72)	12 (1, 58)	28.40
		Estimated	0 (0, 100)	3 (0, 30)	43 (18, 72)	12 (1, 58)	28.12
		Full	0 (0, 100)	2 (0, 38)	37 (14, 67)	10 (1, 59)	30.48
	+1	None	0 (0, 100)	3 (0, 26)	17 (1, 87)	6 (0, 95)	28.50
		Estimated	0 (0, 100)	3 (0, 24)	26 (3, 80)	7 (0, 88)	28.11
		Full	0 (0, 100)	1 (0, 43)	6 (0, 99)	5 (0, 97)	30.00
S2	-1	None	18 (7, 41)	0 (0, 100)	46 (14, 81)	25 (15, 40)	52.49
		Estimated	18 (7, 41)	0 (0, 100)	46 (14, 81)	25 (15, 40)	52.49
		Full	17 (7, 39)	0 (0, 27)	42 (16, 73)	25 (14, 39)	53.83
	0	None	16 (5, 43)	0 (0, 100)	45 (13, 82)	26 (15, 40)	51.13
		Estimated	16 (5, 43)	0 (0, 100)	45 (13, 82)	26 (15, 40)	51.13
		Full	15 (5, 39)	0 (0, 100)	39 (13, 72)	25 (15, 39)	52.93
	+1	None	15 (3, 47)	0 (0, 100)	31 (5, 80)	28 (16, 44)	51.44
		Estimated	16 (4, 49)	0 (0, 100)	34 (4, 87)	28 (16, 43)	51.41
		Full	0 (0, 29)	9 (3, 24)	15 (1, 80)	26 (15, 41)	51.96

* -2 times the maximized log likelihood.

3.5 Composite Population-Averaged Misclassification Rates

We suppose that misclassification rates are required for a study involving many specimens, each read by one of many different readers. We therefore want population-averaged probabilities

$$e_{tv} = \Pr\{\text{reading of } v \mid \text{true stage of liver is } t\}$$
$$= \sum_{u=0}^{t} s_{tu} r_{uv}\text{, where} \quad (7)$$
$$r_{uv} = \iint r_{jkuv}\, dF(\sigma_k)\, dG(\beta_j, \gamma_j). \quad (8)$$

For our purposes, the integrals in equation (8) are really just sums of 3 terms, because we have assumed γ=0 and used simple, discrete forms for $G(\beta_j)$ and $F(\sigma_k)$. To obtain confidence intervals for the $\hat{e}_{tv}$, we use a simple importance sampling algorithm similar to that given in the Appendix, but we randomly generate both $\{\tilde{\alpha}_{uv}\}$ and $\{\tilde{\lambda}_{tu}\}$ from separate, independent multivariate normal distributions with means and covariance matrices as estimated for particular entries in Tables 2 and 3. This ignores some potential dependence between the estimated reading and sampling errors. The need for re-calibration of reading errors, as described at the beginning of the previous section, seems likely to us to minimize the impact of ignoring such dependence.

Table 3 shows estimated composite misclassification rates based on reading errors from the marginal estimate with specimen effects in Table 1 and sampling errors from the best-fitting β=0 entry for study S2 (dependence=None) in Table 2. Because this entry is quite uninformative about s_{21}, we substitute the estimate and variance of λ_{21} from study S1 with the same assumptions (β=0 and dependence=None), and set its covariance with other λ_{tu} to be zero, reflecting the fact that it came from a different study.

The estimated probabilities of misclassification are quite high, particularly when the true stage is 3. We note that blank cells are those that cannot occur due to the assumption of only downward sampling errors of one stage and the assumption that some types of reading errors cannot occur because they were never present in the raw data. We also note that upper confidence bounds, particularly for cells toward the lower left, are smaller than they would be if some uncertainty about the latter assumption had been included in the confidence interval estimation. Estimated composite rates that use sampling error estimates from the best-fitting β=0 entry for study S1 (dependence=Estimated) in Table 2 (but with the λ_{10} estimate from study S2 replacing the largely uninformative one from study S1) are identical for true stages 0, 1, and 2, as they are based on all the

DOI: 10.2202/1557-4679.1139

same parameter estimates. Rates are very similar for true stage 3, with slightly narrower confidence intervals, and correct classification is better for true stage 4, 77%, but with a wider confidence interval, 43% to 89%.

Table 3. Fitted misclassification rates (percentages) reflecting both reading error and sampling error.

		$100 \times \hat{e}_{tv}$ (95% Confidence Interval)				
Read stage v		0	1	2	3	4
True stage t	0	**74 (58, 86)**	26 (14, 42)			
	1	24 (14, 40)	**54 (42, 63)**	22 (14, 29)		
	2	0 (0, 4)	19 (12, 32)	**65 (51, 74)**	16 (9, 24)	
	3		10 (4, 20)	45 (29, 61)	**37 (21, 54)**	8 (2, 20)
	4		1 (0, 6)	7 (3, 13)	25 (16, 35)	**67 (54, 78)**

4 Discussion

We originally envisioned that this analysis would be relatively straightforward and would produce reasonably accurate estimates of fairly small misclassification rates. Instead, our results suggest that liver biopsy may be rather unreliable for assessing the actual state of HCV-related liver disease, and we found a number of limitations in the available data and difficulties in performing analyses that would properly accommodate important features in the data.

4.1 Substantive Implications

Analyses of biopsy-measured fibrosis progression have generally ignored misclassification. Although one reason for this may be technical difficulties in accounting for misclassification within some of the simple statistical approaches that have been used, a lack of any estimates of actual misclassification rates has been another barrier—the abstract concordance measures typically provided in studies of biopsy reliability are of no use in modeling progression. We have focused here on trying to fill this gap, providing estimated misclassification rates that reflect both of the recognized sources or error, reading and sampling.

Despite some recognition of inaccuracies in biopsy-measured fibrosis, it is still used as a gold standard (Cross, Antoniades and Harrison, 2008; Parkes et al., 2006). A recent review states explicitly in its conclusion, "Liver biopsy remains the gold standard for assessment of liver fibrosis" (Manning and Afdhal, 2008). The misclassification estimates obtained here indicate that biopsy is too inaccurate to play such a role. Even the estimate of reading error alone in the fourth row of Table 1 shows error rates that seem too high for use as a gold standard, and those estimates are likely to be very optimistic because they 1) assume no specimen effects, 2) assume no liver sampling error, and 3) are based on optimistic definitions of the true stages of the specimens. The possibly more realistic estimates in Table 3 show dismal performance overall, most notably when the true stage of the liver is F3.

Although one report did characterize agreement between 1 expert and 10 nonacademic pathologists as "very poor" (Rousselet et al., 2005), previous analyses, sometimes using the same raw data that we analyzed here, have generally reached more optimistic conclusions. Several factors may have contributed to this. First, the sampling studies S1 and S2 did show high rates of intra-observer agreement. Second, previous studies focused on reading or sampling error in isolation and did not assess possible reader and specimen effects. Third, previous work relied heavily on abstract concordance measures, rather than estimating actual misclassification rates. Unfortunately, concordance measures that appear quite high are consistent with the poor substantive performance found here, which can produce severe misunderstandings. Study R2, for example, shows substantial raw error rates (see Section 2, above) with strong evidence of both reader and specimen effects, but its authors note an "almost perfect" Kendall Coefficient of Correlation (0.85) and kappa (0.76, if fibrosis stage is grouped into two categories), concluding that, "Acceptable interobserver agreement … should help ensure consistency in patient management" (Netto et al., 2006). Study S2 assumes that all left-right disagreements must be due to sampling error, because they obtained "almost perfect" kappas for intraobserver agreement (Regev et al., 2002).

Because of the higher risk and expense of liver biopsy, there is considerable interest in non-invasive measures of fibrosis (Cross et al., 2008; Manning and Afdhal, 2008; Parkes et al., 2006). Unfortunately, such methods have typically been assessed by receiver operating characteristic (ROC) curve analyses that use biopsy as a gold standard. The area under the ROC curve (AUC) suffers from several drawbacks: 1) it requires dichotomizing the supposed true stage; 2) it has no concrete, practical interpretation; and 3) it does not account for the consequences of correct and mistaken classifications (Vickers and Elkin, 2006). Moreover, errors in biopsy-measured stage will cause poorer performance by AUC (or other measures) even for superior non-invasive measures. Indeed,

DOI: 10.2202/1557-4679.1139

non-invasive measures are specifically thought to have poor ability to distinguish intermediate levels of fibrosis (Bissell, 2004), but this is precisely where biopsy itself appears to be most unreliable. Thus, fair evaluation of non-invasive fibrosis measures would seem to require assessment of long-term clinical and scientific utility, not just direct comparison to biopsy results.

The time of HCV infection is typically unknown (Bacchetti et al., 2007). The focus in studies of biopsy-measured fibrosis is usually on progression over the entire course of infection, making unknown infection time an important limitation. Non-invasive measures that can be performed more frequently could permit focusing instead on trajectories during the measured period, which could mitigate this limitation. In addition, frequent measurement could help mitigate the effects of measurement error, particularly if error is largely independent from one occasion to the next.

4.2 Limitations and Possible Enhancements

The studies analyzed here fall short of ideal in several respects. First, the true stage of specimens is not known with certainty and is particularly suspect for study R1 with regard to estimating specimen effects (see Section 3.2). Because study R2 only had 6 specimens, this limits our assessment of specimen effects, and we do not attempt to estimate non-directional specimen-based accuracy effects as we do for reader effects. Second, the studies of sampling error did not take any steps to eliminate or reduce reading error. Estimation of sampling error from paired biopsies is already challenging due to the true state of the liver being unknown, and the possibility of discrepancies arising from reading error adds further complication. Third, studies R2 and S1 did not represent all stages of fibrosis. Fourth, complete data were not available for study S2. Finally, the studies were heterogeneous in several respects. Study R1 used a different scoring system than the others. Study R2 included mostly post-transplant specimens, although the one from a chronic HCV patient was not read more accurately than the others (10 of 17 correct and all errors downward). Specimen length impacts accuracy (Manning and Afdhal, 2008), and study S1 used smaller specimens overall (median length 14mm) than study S2 (all $\geq$15mm), while study R1 included many (31%) that were <10mm and study R2 did not report on specimen length. In addition, the different study populations may differ in ways that we cannot discern.

The data limitations leave uncertainty about two crucial assumptions for our estimates: the existence of specimen effects outside of the post-transplant setting and the existence and magnitude of any sampling error. A key concern in estimating sampling error is the reason why intraobserver agreement was much higher in studies S1 and S2 than would be predicted by our models of reading

error based on studies R1 and R2. Under our models, intraobserver agreement upon independent re-reading of a specimen would be influenced only by the specimen's true stage, the α parameters for that stage, one reader parameter, and one specimen parameter. In reality, the read stage may also be influenced by multifaceted aspects of the specimen, the reader, and interactions of those aspects. This could produce high intraobserver agreement without indicating high accuracy—reading the same specimen the same way twice may not be the same as reading it correctly twice. Despite this and the fact that analysis of studies R1 and R2 provided some evidence against the existence of large reader accuracy effects, we estimated sampling error as if the source of those studies' high intraobserver agreement is improved accuracy. Without this assumption, estimates of sampling error appeared to be unstable and implausible.

We encountered limitations and technical challenges that necessitated simplifications. Due to lack of strong evidence and the minimal amount of useful data on specimen effects, we assumed no non-directional accuracy effects for both readers and specimens. We assumed that undersampling risk was independent and equal on either side of the liver. In obtaining composite estimates and confidence intervals, we neglected any dependence between reading error estimation and sampling error estimation. Because we found no software that would easily include both reader and specimen random effects (crossed random effects) simultaneously in a multinomial model, we performed sampling error analyses separately for three different types of readers. Including a full normal distribution of specimen effects in the sampling error estimation, particularly for study S2 using likelihood (6), appeared to be technically infeasible, so we represented the specimen effect with a discrete 3-point distribution.

Sampling error estimation shares some features with the challenging situation of comparing diagnostic tests when there is no gold standard (Albert and Dodd, 2004; Hui and Walter, 1980; Pepe and Janes, 2007), notably that the true state of the liver is not known. Our situation is more favorable than comparison of different diagnostic tests without a gold standard in that we can reasonably assume identical left and right sampling error probabilities, halving the number of parameters of interest. Nevertheless, we still had to estimate latent parameters (prevalences of true liver states) and, as noted in Section 3.3, assume conditional independence both of left and right sampling errors given the true liver state and of left and right reading errors given the reader and specimen effects. (We were able to perform some investigation of dependence between left and right specimen effects.) Even with the simplifications noted above and in the previous paragraph, estimation remained difficult for many models, requiring extensive computing resources and evaluation of multiple, randomly-perturbed starting values to ensure identification of global rather than local maxima in the likelihoods. Despite all these difficulties, we believe that our results suggest that

DOI: 10.2202/1557-4679.1139

sampling error may substantially increase misclassification rates. Avoiding the difficulties by simply ignoring sampling error would therefore seem likely to be a dangerous strategy.

More comprehensive and elegant statistical methods are possible in principle, though probably not feasible for these data sets and not worth the extensive effort that would be needed, given the limited amount and quality of the available data. Rather than assigning a true stage for studies R1 and R2, one could perhaps generalize the latent class methods discussed in the previous paragraph to the multi-state case. This might require careful parameterization to preserve identifiability, particularly because study R1 has only two readers, and such an approach probably could not make any use of the consensus readings. A more customized approach, possibly using Markov-chain Monte Carlo methods, might be able to estimate models that include both random reader and random specimen effects, perhaps even with both directional and non-directional effects for each, such as the β and γ in equation (2). Joint estimation of reading and sampling parameters could utilize all four studies at once. Any future studies of sampling error, however, would be much more informative if they eliminated dependence on estimation of reading error by ensuring correct readings of all specimens. (The invasiveness and risk of taking two biopsies would seem to require optimization of the information obtained from the specimens, justifying any extra costs from use of multiple readers.)

For any future studies of reading error, the potential importance of both reader and specimen effects argues for inclusion of large numbers of both readers and specimens (in contrast to the severe asymmetries in studies R1 and R2). Such studies need not have each specimen read by each reader, but optimizing allocation of numbers of specimens per reader, readers per specimen, and patterns of overlap could pay off with improved accuracy and cost efficiency. Because liver biopsy is already unpopular with clinicians and patients (Cross et al., 2008), such careful study may never occur, but similar considerations may also apply to other multi-state situations.

4.3 Conclusions and Recommendations

There appears to be a considerable possibility that biopsy is far too inaccurate to be considered a gold standard for measuring fibrosis in patients with HCV, and biopsy reading appears to differ systematically between readers. We acknowledge, however, that the accuracy of biopsy is difficult to estimate with the data we were able to obtain. Many uncertainties about basic modeling assumptions, noted in Section 4.2, are difficult to quantify, and even the stochastic uncertainty alone, as shown by confidence intervals in Table 3, is considerable. Ideally, accurate external estimates of misclassification probabilities could

improve the performance of models of fibrosis progression. Because of the uncertainty encountered here, however, we would recommend performing such modeling with both optimistic estimates, such as the fourth line of Table 1, and less optimistic estimates such as in Table 3. Because these are so uncertain, use of indirect estimation as part of a multi-state modeling process may be as good as or better than using external estimates. In addition, study of fibrosis progression in patients with HCV may be as informative with non-invasive fibrosis measures as it would be with biopsy assessment. Many of the non-invasive measures are continuous and would therefore avoid the need for multi-state modeling methods altogether.

Appendix 1 – Raw data analyzed

Table A1. Raw data on read and consensus stages for Study R1.

Fibrosis stage by			
Consensus	Reader 1	Reader 2	Count
0	0	0	7
0	0	1	1
0	1	0	3
1	0	1	1
1	1	0	6
1	1	1	28
1	1	2	6
1	2	1	14
2	1	2	1
2	2	1	8
2	2	2	31
2	2	3	4
2	3	2	4
3	3	1	1
3	3	2	6
3	3	3	7
3	4	2	1
3	4	3	1
4	3	4	2
4	4	3	2
4	4	4	23

Raw data for study R2 have already been published (Netto, et al., 2006).

DOI: 10.2202/1557-4679.1139

Table A2. Raw counts for each combination of read stages for left liver lobe and right liver lobe specimens from Study S1.

		Read stage on right				
		0	1	2	3	4
Read stage on left	0	0	0	0	0	0
	1	0	1	2	0	0
	2	0	2	30	8	0
	3	0	0	4	6	1
	4	0	0	0	1	5

Raw data for study S2 are not available. We have instead analyzed the published information (Regev, et al., 2002), as described in Section 3.3.

Appendix 2 – Example SAS code fitting a reading model

The code below illustrates use of the SAS NLmixed procedure to fit a model like the one described in Section 3.2 leading to the estimates shown in Table A3. For illustrative purposes, we include here estimation of a random specimen effect, even though that was not included in the model, and it is estimated to have zero variance, implying no specimen effects.

```
data R1; input true reading1 reading2 count;
do i=1 to count;
  specimenID+1;
  reading=reading1; reader=1; output;
  reading=reading2; reader=2; output;
end;
cards;
<< Data from Table A1 >>
run;

proc nlmixed data=R1 tech=nrridg absgconv=1e-9[5] gconv=1e-10[5];
      title Study R1 with fixed reader, random specimen effects;
              /* Starting values of parameters */
      parms alpha01=-2 alpha10=-2 alpha12=-2 alpha21=-2
            alpha23=-2 alpha31=-3 alpha32=-2 alpha34=-2
            alpha43=-2 beta=0 specimenSD=0.1;
              /* Identifiability constraint for Reader effects */
      if reader=1 then shift=beta; else shift=-beta;

      select (true); * stage by consensus, "true" stage ;


/* block for observations where true stage = 0 */
```

```
      when (0) do; * true stage = 0 ;
              /* Numerator of likelihood if read stage=1 */
        lognumer01 = alpha01 + shift + specimen;
              /* Denominator of likelihood */
        logdenom = log(1 + exp(lognumer01)) ;
              /* ll is the log-likelihood */
              /* Numerator is 1 if read correctly */
        if reading = 0 then ll = -logdenom;
        else if reading = 1 then ll = lognumer01 -logdenom;
        else ll=.;
      end;

/* block for observations where true stage = 1 */
      when (1) do;
        lognumer10 = alpha10 - shift - specimen;
        lognumer12 = alpha12 + shift + specimen;
        logdenom = log(1 + exp(lognumer10) + exp(lognumer12)) ;
        if reading = 0 then ll = lognumer10 -logdenom;
        else if reading = 1 then ll = -logdenom;
        else if reading = 2 then ll = lognumer12 -logdenom;
        else ll=.;
      end;
/* block for observations where true stage = 2 */
      when (2) do;
        lognumer21 = alpha21 - shift - specimen;
        lognumer23 = alpha23 + shift + specimen;
        logdenom = log(1 + exp(lognumer21) + exp(lognumer23)) ;
        if reading = 1 then ll = lognumer21 -logdenom;
        else if reading = 2 then ll = -logdenom;
        else if reading = 3 then ll = lognumer23 -logdenom;
        else ll=.;
      end;
/* block for observations where true stage = 3 */
      when (3) do;
        lognumer31 = alpha31 - shift - specimen;
        lognumer32 = alpha32 - shift - specimen;
        lognumer34 = alpha34 + shift + specimen;
        logdenom = log(1 + exp(lognumer31) +
                   exp(lognumer32) + exp(lognumer34)) ;
        if reading = 1 then ll = lognumer31 -logdenom;
        else if reading = 2 then ll = lognumer32 -logdenom;
        else if reading = 3 then ll = -logdenom;
        else if reading = 4 then ll = lognumer34 -logdenom;
        else ll=.;
      end;
/* block for observations where true stage = 4 */
      when (4) do;
        lognumer43 = alpha43 - shift - specimen;
        logdenom = log(1 + exp(lognumer43)) ;
        if reading = 3 then ll = lognumer43 -logdenom;
```


DOI: 10.2202/1557-4679.1139

```
        else if reading = 4 then ll = -logdenom;
        else ll=.;
      end;
/* Wrap up likelihood calculations */
      otherwise ll=. ;
      end;  * Close select statement from above ;

/* Specify random specimen effect */
      random specimen ~ normal(0, specimenSD*specimenSD)
                        subject=specimenID;
/* Use general optimization capability */
       model true ~ general(ll);
run;
```

Appendix 3 – Fitted intercept parameters

Table A3. Fitted intercept parameters $\hat{\alpha}_{uv}$ as defined at equation (2), with (95% Confidence intervals), for study R1. Blank cells had no specimen with that combination of u and v and therefore all have estimates of $\hat{\alpha}_{uv} = -\infty$; diagonal cells have $\hat{\alpha}_{uv} = 0$ by definition.

		Read stage v				
	$\hat{\alpha}_{uv}$	0	1	2	3	4
True stage u	0	0	-1.6 (-2.7, -0.5)			
	1	-2.7 (-3.5, -1.9)	0	-1.6 (-2.1, -1.0)		
	2		-2.3 (-3.1, -1.6)	0	-2.5 (-3.2, -1.7)	
	3		-3.2 (-5.2, -1.2)	-1.3 (-2.1, -0.4)	0	-2.6 (-4.1, -1.2)
	4				-2.7 (-3.7, -1.6)	0

Appendix 4 – Confidence intervals for Table 1

Table A4. Confidence intervals for estimates in Table 1.

		$100 \times \hat{r}_{j \cdot uv}$													
		Lower 95% confidence bound Upper 95% confidence bound													
True stage u:		0		1			2			3				4	
Read stage v:		0	1	0	1	2	1	2	3	1	2	3	4	3	4
Specimen effects†	β														
No	-1	**93**	7	15	**79**	6	20	**77**	2	6	40	**52**	1	16	**84**
		82	2	7	**66**	4	11	**64**	1	1	21	**30**	0	6	**65**
		98	18	28	**87**	10	34	**86**	5	31	60	**70**	6	35	**94**
	0	**83**	17	5	**78**	16	8	**85**	7	3	20	**72**	5	6	**94**
		63	6	2	**69**	11	4	**75**	4	0	9	**50**	1	2	**84**
		94	37	11	**85**	25	15	**90**	14	18	36	**83**	19	16	**98**
	+1	**65**	35	2	**63**	36	3	**79**	18	1	8	**76**	15	2	**98**
		38	15	1	**50**	25	1	**65**	10	0	3	**50**	4	1	**93**
		85	62	3	**73**	48	5	**88**	32	8	16	**88**	43	7	**99**
	m‡	**81**	19	7	**73**	19	10	**80**	9	3	23	**67**	7	8	**92**
		61	8	3	**63**	13	5	**70**	5	0	11	**46**	2	3	**81**
		92	39	14	**81**	28	18	**87**	17	19	37	**79**	22	19	**97**
Yes	-1	**86**	14	24	**63**	13	29	**64**	7	6	41	**48**	5	25	**75**
		73	6	14	**50**	9	20	**53**	3	1	23	**32**	1	13	**60**
		94	27	35	**73**	19	39	**74**	12	29	55	**60**	15	40	**87**
	0	**74**	26	12	**62**	26	16	**69**	15	4	28	**56**	12	13	**87**
		58	13	6	**51**	19	9	**57**	8	1	15	**36**	4	6	**75**
		87	42	21	**71**	33	25	**79**	23	21	40	**70**	29	25	**94**
	+1	**60**	40	5	**54**	41	7	**65**	27	2	16	**58**	24	6	**94**
		43	24	2	**45**	33	4	**53**	18	0	8	**36**	9	2	**86**
		76	57	10	**63**	49	13	**76**	38	13	25	**75**	45	14	**98**
	m‡	**74**	26	14	**60**	27	17	**66**	16	4	28	**54**	14	15	**85**
		58	14	8	**49**	20	11	**55**	10	1	16	**35**	5	7	**74**
		86	42	22	**69**	34	25	**76**	24	21	40	**68**	30	26	**93**

† If present, specimen effects are assumed to be -1.92, 0, and +1.92 each with probability 1/3.
‡ Marginal rates, averaged over readers with β= -1, β=0, and β= +1.

To obtain the above confidence intervals for the $\hat{r}_{j \cdot uv}$ in Table 1, we use a very simple importance sampling algorithm (Evans and Swartz, 1995):

DOI: 10.2202/1557-4679.1139

Algorithm for obtaining confidence intervals

1. Randomly generate $\{\tilde{\alpha}_{uv}\}$ from a multivariate normal distribution with means and covariance matrix as estimated for Table A3. (For cases where $\hat{\alpha}_{uv} = -\infty$, we set $\tilde{\alpha}_{uv} = -\infty$ with probability 1.)
2. Calculate $\{\tilde{r}_{j\cdot uv}\}$ using equations (1) and (2), along with the same assumptions about reader and specimen effects as used for the entries in Table 1.
3. Repeat steps 1-2 a total of 10,000 times.
4. For each $\hat{r}_{j\cdot uv}$, estimate its confidence bounds as the 2.5 and 97.5 percentiles of its 10,000 calculated values $\tilde{r}_{j\cdot uv}$.

References

Albert, P. S., and Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* **60**, 427-435.

Albert, P. S., Hunsberger, S. A., and Biro, F. M. (1997). Modeling repeated measures with monotonic ordinal responses and misclassification, with applications to studying maturation. *Journal of the American Statistical Association* **92**, 1304-1311.

Bacchetti, P., Tien, P. C., Seaberg, E. C.*, et al.* (2007). Estimating past hepatitis C infection risk from reported risk factor histories: implications for imputing age of infection and modeling fibrosis progression. *BMC Infectious Diseases* **7**, 145.

Batts, K. P., and Ludwig, J. (1995). Chronic hepatitis - an update on terminology and reporting. *American Journal of Surgical Pathology* **19**, 1409-1417.

Bedossa, P., Bioulacsage, P., Callard, P.*, et al.* (1994). Intraobserver and interobserver variations in liver-biopsy interpretation in patients with chronic hepatitis-C. *Hepatology* **20**, 15-20.

Bissell, D. M. (2004). Assessing fibrosis without a liver biopsy: Are we there yet? *Gastroenterology* **127**, 1847-1849.

Cross, T., Antoniades, C., and Harrison, P. (2008). Non-invasive markers for the prediction of fibrosis in chronic hepatitis C infection. *Hepatology Research* **38**, 762-769.

Deuffic-Burban, S., Poynard, T., and Valleron, A. J. (2002). Quantification of fibrosis progression in patients with chronic hepatitis C using a Markov model. *Journal of Viral Hepatitis* **9**, 114-122.

Evans, M., and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science* **10**, 254-272.

Gentleman, R. C., Lawless, J. F., Lindsey, J. C., and Yan, P. (1994). Multistate Markov-models for analyzing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine* **13**, 805-821.

Hui, S. L., and Walter, S. D. (1980). Estimating the error rates of diagnostic-tests. *Biometrics* **36**, 167-171.

Jackson, C. H., and Sharples, L. D. (2002). Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine* **21**, 113-128.

Jackson, C. H., Sharples, L. D., Thompson, S. G., Duffy, S. W., and Couto, E. (2003). Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society Series D-the Statistician* **52**, 193-209.

Kalbfleisch, J. D., and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* **80**, 863-871.

Kay, R. (1986). A Markov model for analyzing cancer markers and disease states in survival studies. *Biometrics* **42**, 855-865.

Manning, D. S., and Afdhal, N. H. (2008). Diagnosis and quantitation of fibrosis. *Gastroenterology* **134**, 1670-1681.

Mwalili, S. M., Lesaffre, E., and Declerck, D. (2008). The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Statistical Methods in Medical Research* **17**, 123-139.

Netto, G. J., Watkins, D. L., Williams, J. W*., et al.* (2006). Interobserver agreement in hepatitis C grading and staging and in the Banff grading schema for acute cellular rejection - The "Hepatitis C 3" multi-institutional trial experience. *Archives of Pathology & Laboratory Medicine* **130**, 1157-1162.

Parkes, J., Guha, I. N., Roderick, P., and Rosenberg, W. (2006). Performance of serum marker panels for liver fibrosis in chronic hepatitis C. *Journal of Hepatology* **44**, 462-474.

Pepe, M. S., and Janes, H. (2007). Insights into latent class analysis of diagnostic test performance. *Biostatistics* **8**, 474-484.

Regev, A., Berho, M., Jeffers, L. J*., et al.* (2002). Sampling error and intraobserver variation in liver biopsy in patients with chronic HCV infection. *American Journal of Gastroenterology* **97**, 2614-2618.

Rousselet, M. C., Michalak, S., Dupre, F*., et al.* (2005). Sources of variability in histological scoring of chronic viral hepatitis. *Hepatology* **41**, 257-264.

DOI: 10.2202/1557-4679.1139

Satten, G. A., and Longini, I. M. (1996). Markov chains with measurement error: Estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease. *Applied Statistics-Journal of the Royal Statistical Society Series C* **45**, 275-295.

Skripenova, S., Trainer, T. D., Krawitt, E. L., and Blaszyk, H. (2007). Variability of grade and stage in simultaneous paired liver biopsies in patients with hepatitis C. *Journal of Clinical Pathology* **60**, 321-324.

Stram, D. O., and Lee, J. W. (1994). Variance-components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171-1177.

Terrault, N., Im, K., Boylan, R.*, et al.* (2008). Fibrosis progression in African Americans and Caucasian Americans with chronic hepatitis C. *Clinical Gastroenterology and Hepatology* **6,** 1403-1411.

Vickers, A. J., and Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making* **26**, 565-574.