Johns Hopkins University, Dept. of Biostatistics Working Papers

12-2-2004

The Proportional Odds Model for Assessing Rater Agreement with Multiple Modalities

Elizabeth Garrett-Mayer
Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, esg@jhu.edu

Steven N. Goodman
Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University

Ralph H. Hruban
Department of Pathology, Johns Hopkins University

The Proportional Odds Model for Assessing Rater Agreement with Multiple Modalities

Elizabeth Garrett-Mayer
Sidney Kimmel Comprehensive Cancer Center
Johns Hopkins University
Baltimore, MD, 21205
esg@jhu.edu

Steven N. Goodman
Sidney Kimmel Comprehensive Cancer Center
Johns Hopkins University
Baltimore, MD, 21205

Ralph H. Hruban
Department of Pathology
Johns Hopkins University
Baltimore, MD, 21205

Elizabeth Garrett-Mayer is Assistant Professor of Oncology and Biostatistics at the Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Suite 1103, 550 North Broadway, Baltimore, MD 21205 (esg@jhu.edu); Steven N. Goodman is Associate Professor of Oncology, Pediatrics, Epidemiology, and Biostatistics at the Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD 21205; Ralph H. Hruban is Professor of Pathology in the Johns Hopkins School of Medicine, Baltimore, MD 21205.

Abstract

In this paper, we develop a model for evaluating an ordinal rating systems where we assume that the true underlying disease state is continuous in nature. Our approach in motivated by a dataset with 35 microscopic slides with 35 representative duct lesions of the pancreas. Each of the slides was evaluated by eight raters using two novel rating systems (PanIN illustrations and PanIN nomenclature), where each rater used each systems to rate the slide with slide identity masked between evaluations. We find that the two methods perform equally well but that differentiation of higher grade lesions is more consistent across raters than differentiation across raters for lower grade lesions.

A proportional odds model is assumed, which allows us to estimate rater-specific thresholds for comparing agreement. In this situation where we have two methods of rating, we can determine whether the two methods have the same thresholds and whether or not raters perform equivalently across methods. Unlike some other model-based approaches for measuring agreement, we focus on the interpretation of the model parameters and their scientific relevance. We compare posterior estimates of rater-specific parameters across raters to see if they are implementing the intended rating system in the same manner. Estimated standard deviation distributions are used to make inferences as to whether raters are consistent and whether there are differences in rating behaviors in the two rating systems under comparison.

Keywords: latent trait, ordinal rating, Bayesian model, reliability.

1 INTRODUCTION

Histologic classification systems for precursor lesions help in the understanding of cancer progression. However, in some types of cancer lesions, attempts at standard classification systems have only recently occurred. As an example, proliferative epithelial lesions in the smaller caliber pancreatic ducts and ductiles did not have a standard nomenclature until a recent attempt by Hruban et al. (2001). Without a system for classifying lesions, further research including clinical, pathologic, and molecular studies proves difficult. In cases where there is not a standard classification system, one needs to be established and assessed for reliability before correlates of progression can be studied.

Hruban et al. (2001) propose a classification system in proliferative epithelial lesions in the smaller caliber pancreatic ducts. An initial study was performed where eight pathologists reviewed 35 microscopic slides with 35 representative duct lesions in the pancreas. They found that over 70 different diagnostic terms were used by the eight pathologists. This led the authors to develop a standard nomenclature and diagnostic system which they call "pancreatic intraepithelial neoplasia," or PanIN. The PanIN classification has the following levels: 0 (normal), 1-A, 1-B, 2, 3, and 4, where 4 can be thought of as an "other" category for the purposes of this paper. PanIN is presented in terms of "nomenclature" and as "illustrations" (Figure 5), allowing a rater to use either of these systems to assign a rating. (For the illustrations, there is another figure showing PanIN 2, 3, and 4, but we do not include this figure for brevity.) After developing the PanIN classification system, the reliability of the system was assessed by allowing the original eight raters to reevaluate of the slides using each of the two PanIN systems . Standard reliability methods (e.g., kappa statistics) found acceptable agreement and the new system was deemed as promising.

We reanalyze the PanIN rating data presented in Hruban et al. (2001), considering more subtle aspects of the study design. The eight raters evaluate each of 35 slides using two rating methods (nomenclature and illustrations) for total of $8 \times 35 \times 2 = 560$ ratings. Slide identity is masked between the illustration and nomenclature evaluations so that raters cannot use their first evaluation to influence their rating of the second evaluation. The 35 slides chosen were intended to reflect levels of disease between 1-A

and 4.

From a statistical point of view, this is an interesting problem in several ways: (1) there is no gold standard to which to compare the ratings, (2) it is assumed that there is variability due to both rater and method of rating (i.e., illustration versus nomenclature), (3) the scale is ordinal (as opposed to nominal or continuous), except for the fourth grade which refers to an "other" category. If there were a gold standard, it would be straightforward to compare the observed ratings to the true grade. Kappa statistics could be used or tabular methods summarizing sensitivity and specificity of diagnosis could be assessed for each of the rating methods. Standard kappa statistics will not address the issue of variability from two sources, requiring more elaborate methods for accounting for the rater and method effects.

The last point, regarding the ordinal form of the staging variable, is very common in medical research. We often treat this "observed" grade variables as representative of the true nature of the disease (i.e., we assume the underlying condition is categorical). However, it is natural to think that cancers are not truly "ordinal" in nature, but they are classified according to "cutoffs." As a result, true lesion grade can be thought of as a continuous variable that is observed as an ordinal variable where certain thresholds or cutoffs determine to which ordinal category the observed measure belongs. These assumptions coincide with the idea that true lesion grade is a continuous latent variable. In deciding to which category to assign a lesion, a rater applies a system of cutoffs—thresholds which dictate lesion assignment. Consider figure 5 where the rating cutoffs are shown for the intended rating system ("truth") and for two theoretical raters where true underlying disease scores range from 0 to 4.5 and samples are assigned to categories of 0, 1, 2, 3, and 4. The intended system has thresholds as follows: continuous scores between 0 and 0.5 are assigned to the 0 category, scores between 0.5 and 1.5 are assigned to the 1 category, and so on. Notice that the two raters impose slightly different cutoffs for assigning samples than the intended system: rater 1 has a wider 0 category and then each of the other categories is shifted up, and rater 2 has narrower 0 and 3 categories, and wider 1 and 4 categories. Although these raters will tend to agree on a large number of cases and be consistent with the intended system, there will be disagreement. To illustrate this, four underlying disease scores have been plotted using dotted vertical lines. Correct classification is denoted by "o" and incorrect by

"x." For disease score of 0.7, rater 1 misclassified the sample as a 0, and for disease score of 1.8 rater 2 misclassified the sample as a 1. The other two disease scores, 2.1 and 2.9, were correctly classified by both raters. Note that these thresholds are latent variables and cannot be directly observed.

Our goal is to estimate the thresholds for each rater. In this paper, we develop a model for assessing rater agreement for the Pan-IN system. We assume that the true underlying disease state is continuous in nature. That is, disease grade is a latent trait. A proportional odds model is assumed, which allows us to estimate rater-specific thresholds for comparing agreement. In this situation where we have two methods of rating, we can determine whether the two methods have the same thresholds and whether or not raters perform equivalently across methods. Unlike other model-based approaches for measuring agreement, we focus on the interpretation of the model parameters and their scientific relevance. We compare posterior estimates of rater-specific parameters across raters to see if they are implementing the intended rating system in the same manner. Estimated standard deviation distributions are used to make inferences as to whether raters are consistent and whether there are differences in rating behaviors in the two rating systems under comparison.

2 BACKGROUND

2.1 Measuring Agreement in Categorical Responses

Methods for quantifying agreement for continuous responses have been well studied. Regression techniques, intraclass correlations, and other measures provide means of evaluating agreement depending on the nature of the data. However, in many clinical applications, continuous measures cannot be observed. For evaluating agreement between categorical outcomes, kappa statistics are most prevalent. The most basic form of Cohen's kappa statistic (Cohen, 1960) compares two binary ratings: the difference between observed and expected agreement is divided by 1 minus the expected agreement, where expected agreement is based on the marginal prevalence of the ratings. The kappa is also generalizable to more than two ordinal or nominal categories A weighted kappa has been developed for use in situations where there are more than two

nominal categories, and also when there are more than two ordinal categories where "close" disagreements are penalized less than those further away (Cohen, 1968). Kappa statistics can be used when there are multiple raters grading the same items, or in the case where the same rater uses two methods for evaluating items. In our situation, we have both kinds of variability: multiple raters and multiple rating methods. As a result, there is no version of the kappa statistic that will allow us to evaluate agreement of readers while simultaneously considering differences between methods. Only if we analyzed the two datasets (i.e., those based on nomenclature and illustrations) separately will we get results based on standard kappa-based approaches, but these will not allow a rigorous comparison across methods.

Other model-based methods have also been devised for measuring agreement between raters. Tanner and Young (1985) used a log-linear model to describe association between raters, describing agreement of J raters on N subjects, 1 rater with J readings each on N subjects, or J raters compared to a gold-standard with ratings on N subjects. The log-linear model partitioned the observed data into two parts: the random or "chance" component, and the agreement component. Becker and Agresti (1992) also use a log-linear model in their approach, but they focus on modeling the overall structure of agreement, using second order dependence of the joint distribution of the ratings. Pairwise comparisons of raters are made to assess whether (1) all pairs of raters have the same structural pattern of agreement and (2) raters have same aggregate level of agreement with other raters. This model was then extended (Perkins and Becker, 2002) incorporating models for univariate marginal distributions and log non-linear models for modeling pairwise agreements. These modeling approaches have all been developed for the purposes of analyzing data from J raters on N subjects, or for J evaluations from a single rater.

Agresti (1988) extended the model proposed by Tanner and Young (1985) from nominal to ordinal data, which is a more appropriate approach for many kinds of rating data. However, the approach maintains the log-linear form, which still treats the underlying mechanism (i.e. grade) as categorical. A latent class model was developed that assumes that there is an underlying "true positive" or "true negative" rating for each rated object Agresti and Lang (1993). While this is an attractive approach because it incorporates a latent variable, the categorical nature of the latent variable

is not ideal in our setting.

Uebersax and Grove (1996) introduced a model that assumes that the underlying variable of interest is a latent trait, and more specifically, that there are a mixture of latent trait distribution. Similar to Agresti and Lang (1993) whose model is geared for the situation where the goal is to divide objects into categories of, for example, "diseased" and "not diseased," the Uebersax and Grove model assumes two components in the mixture and that binary classification is the primary interest. This model is rich in that it provides for bias, measurement error, and differing sensitivity of rater cutoffs. However, one drawback of this model is that it imposes a strong normality assumption on the mixture components which may not be appropriate due to sampling or small sample size.

A Bayesian analysis assuming that the underlying variable of interest is continuous is developed by Johnson (1996). A Gibbs Sampler approach is used for model estimation and the hierarchical structure allows ease of model estimation without problems of identifiability. Several aspects of Johnson's model include the assumption that the latent trait is normally distributed in the population and that the parameter of interest is reliability, as measured by the rater-specific rating variances. In other words, reliability is measured as the how well each rater adheres to the estimated thresholds. While Johnson introduces the concept of rater bias into the model, the approach that he develops assumes all rater biases to be 0 with estimation of rater variance the key parameter of interest. Another critical part of Johnson's model is that the estimated rater cutoffs are mapped from the rating metric to quantiles of the normal distribution. The normality assumption, while perhaps appropriate in some settings, will not be generally applicable. In our Pan-IN dataset, the ratings were sampled so that there are approximately equal numbers in each of the categories, and is more similar to a uniform distribution than a normal distribution.

The model that we develop is unlike the kappa-based and model-based approaches that focus on test-statistic inference for determination of agreement. It is more similar in nature to the models by Uebersax and Grove (1996) and Johnson (1996). We do not assume that true disease status is categorical, but instead is continuous (i.e. a latent trait). Unlike Johnson (1996), we focus on rater bias (i.e., validity) versus rater variance, our model does not rely on the normality assumption, and instead of using

a probit link, a logit link is used via the proportional odds model (i.e., ordinal logistic regression) (McCullagh, 1980). In this approach, we do not map the estimated cutoffs to quantiles of the normal distribution, but instead, they are mapped back to the original metric of the data. As a result, rater disagreement and bias can be better assessed because, especially to clinicians, the rating metric is clinically meaningful whereas the the interpretation of quantiles of the standard normal distribution is generally not.

2.2 Proportional Odds Model

The proportional odds model (POM) with logit link was introduced by McCullagh (1980), motivated by measures of bioassays where "the latent variable usually corresponds to a 'tolerance,' assumed to have a continuous distribution in the population. Tolerances themselves are not directly observable, but increasing tolerance is manifest in an increase in the probability of survival. The categories are envisaged as contiguous intervals on the continuous scale." McCullagh's model was developed in the context of regression, where the goal was to estimate the association between risk factors and an ordinal outcome of interest. It is a clever model, utilizing the simplicity of the logistic regression model with two categories, but allowing for more than two ordinal categories of outcome.

Define an ordinal outcome variable y with K ordinal categories where y takes values $1, 2, .., K$ and of interest is the association between a covariate x and outcome y. The proportional odds model in this case has the following form:

$$\log\left(\frac{P(Y > k)}{P(Y \leq k)}\right) = \alpha_k + \beta x; k = 1, ..., K-1 \quad (1)$$

There are $K-1$ intercepts and only one slope, β. The intercepts, $\alpha_1, ..., \alpha_{K-1}$ are the thresholds or cutpoints and have tended to be of little interest on their own. The β coefficient is usually of primary interest: it represents the log odds ratio of scoring $> k$ versus $\leq k$ for a one unit change in x. An important point to note is that the log odds ratio is independent of the value of k: a key feature of the POM is that β is constant for increasing values of k.

3 METHODS

3.1 Parameterizing the POM for comparison of raters and modalities

Unlike McCullagh's original applications of the proportional odds model, where the log odds ratio β was the primary quantity of interest, we focus on the predicted probabilities of ordinal assignment while incorporating a continuous latent variable into the model. Additional effects (or biases) of interest include rater effects, method effects, and the interaction between rater and method. The interactions allow for raters to have different thresholds for the different methods. In our example, we have data with three ordinal categories (levels 1, 2, and 3 of the PanIN syste), and an "other" category, which we treat as missing due to the inability to include it as an ordinal categories. With three rating categories, there are two cutpoints in the POM: between $k \leq 1$ and $k > 1$ (the lower threshold), and between $k \leq 2$ and $k > 2$ (the upper threshold). We can imagine that each rater has his own set of cutoffs that he applies when determining a rating for a slide. This arises from the idea that there is a true underlying continuous cancer grade which the rater observes, but s/he is forced to quantify it in the ordinal scale. This could be thought of as "rounding" to the closest integer between 1 and 3 in our case. For example, when deciding whether a slide is a 1 or a 2, one rater may be using a cutoff of 1.5 whereas another might use a cutoff of 1.7.

Returning to the parameters of interest, we would like to estimate the lower and upper cutoffs for each rater to determine if some tend to give higher or lower ratings than other raters, and we would like to do this for each of the modalities. Because we are assuming that each slide has an associated true continous grade, we must estimate the latent grades, denoted by η_i, in the model. We have one additional parameter to estimate, which is the β coefficient from the POM shown in equation (1). Although this is the parameter of interest in McCullagh's original implementation of this model, it is not of primary interest to us. In this application, we have assumed that β is constant across raters and will focus on the inferences associated with the parameters that tell us about differences in thresholds between raters and between methods. True grade for lesion i is denoted by η_i.

A POM which is consistent with the model described above can be written as equation (2), where $k = 1, 2, 3$ indexes the ratings, $m = 0, 1$ refers to nomenclature and illustrations modalities, and $i = 1, ..., N$ indexes the slides. Parameter interpretations are summarized in table 1.

$$\log\left(\frac{P(Y_{ijm} > k)}{P(Y_{ijm} \leq k}\right) = \beta\eta_i + \alpha_{jk} + m\delta_j + m(k-1)\gamma_j. \tag{2}$$

Note that k only takes values 1 and 2 in equation (2). To better understand the parameters in the model, we condition on values of m and k to show via equations the interpretation of the linear combinations of the parameters:

$$\begin{aligned}
logit(P(Y_{ijm} > k|k = 1, m = 0) &= \beta\eta_i + \alpha_{j1} && (3)\\
logit(P(Y_{ijm} > k|k = 1, m = 1) &= \beta\eta_i + \alpha_{j1} + \delta_j && (4)\\
logit(P(Y_{ijm} > k|k = 2, m = 0) &= \beta\eta_i + \alpha_{j2} && (5)\\
logit(P(Y_{ijm} > k|k = 2, m = 1) &= \beta\eta_i + \alpha_{j2} + \delta_j + \gamma_j && (6)
\end{aligned}$$

Equations (3) and (4) relate to the lower thresholds while equations (5) and (6) relate to the upper thresholds. The parameter α_{j1} represents the lower threshold for rater j for nomenclature and δ_j is the method effect for rater j for the lower threshold. In other words, δ_j measures whether or not rater j's estimated lower threshold differs across methods. Similarly, we can interpret equations (5) and (6): α_{j2} represents the upper threshold for rater j for nomenclature, and $\delta_j + \gamma_j$ is the method effect for the upper threshold for rater j.

Notice that due to the parameterization, there are contrasts which can easily be used to decide whether or not method effects exist and whether or not the method effects differ across the two thresholds. If it appears that δ_j is approximately 0, then we would conclude that there are not method effects for the lower threshold. If γ_j is approximately 0, then we could conclude that the method effects are the same for the upper and lower thresholds.

3.2 Model Assumptions and Estimation

The latent variable, true grade denoted by η, requires some modeling assumptions for estimation. For example, we could assume that grade is normally distributed with mean 0 and variance 1. However, given the nature of the study (i.e., samples were chosen to adequately reflect varying grades of disease with no interest in reproducing the distribution of samples in the population of individuals with these lesions), we decided to assume that true grade has a uniform distribution in the range 0.5 to 3.5. This is consistent with the idea that samples with true grade of 0.5 to 1.5 should be classified as 1, those with true grade of 1.5 to 2.5 as 2, and 2.5 to 3.5 as 3. A nice feature of this approach is that we have preserved the original scale of the ratings. However, other distributional assumptions can be chosen for estimation of η as will be discussed in section 4.2.

We have assumed a random effects POM in this application. Given the number of parameters and the size of the dataset, it is reasonable to assume that the rater effects come from a common distribution. We assume hierarchical distributions on α_{jk}, δ_j, and γ_j.

A Markov chain Monte Carle estimation procedure was used to estimate the POM in equation (2). Priors and hierarchical distributions on the parameters described are as follows:

$$\begin{aligned}
\alpha_{j1} &\sim N(\alpha_1, \sigma^2_{\alpha 1})I(-\infty, \alpha_{j2}) && (7)\\
\alpha_{j2} &\sim N(\alpha_2, \sigma^2_{\alpha 2})I(\alpha_{j1}, \infty) && (8)\\
\delta_j &\sim N(0, \sigma^2_{\delta})\\
\gamma_j &\sim N(0, \sigma^2_{\gamma})\\
\eta_i &\sim U[0.5, 3.5]\\
\beta &\sim N(0, 4)\\
\alpha_1 &\sim N(0, 100)\\
\alpha_2 &\sim N(0, 100)\\
1/\sigma^2_{\alpha 1} &\sim Gamma(0.01, 0.01)\\
1/\sigma^2_{\alpha 2} &\sim Gamma(0.01, 0.01)
\end{aligned}$$

$$
\begin{aligned}
1/\sigma_{\delta}^{2} &\sim Gamma(0.01, 0.01) \\
1/\sigma_{\gamma 2}^{2} &\sim Gamma(0.01, 0.01)
\end{aligned}
$$

(Equations (7) and (8) are truncated normals.)

The model was estimated using WinBugs 1.3 (Imperial College of Science, Technology and Medicine, 2000). A standard approach for MCMC estimation was implemented as follows. We performed a burn-in of 1000 iterations and determined the model had converged by examination of traceplots. An additional 5000 iterations were run and every 10th iteration was saved for analysis. These chains of length 500 were then entered into the R statistical package (The R Development Core Team, 2003). For each parameter in the model in equation (2), the posterior distribution was estimated based on the MCMC chain results. Point estimates and standard errors of parameters were calculated to be the means and standard deviations of the posterior distributions. For functions of the parameters, we first calculated the function of the parameters for each iteration of the chain. We then estimated the posterior distribution, point estimate, and standard error as above.

3.3 Estimating Thresholds on the Rating Scale

Although the rater-specific threholds are the quantities of interest, the direct interpretation of the parameters in the model in equation (2) do not coincide with the ordinal 1,2,3 scale of the observed data due to the logit link. As a result, we estimate more intuitive quantities for comparing the raters by tranforming the model parameters to the metric of the rating scale. Of interest are the values of true grade that serve as the cutoffs for each rater. These can be considered the values for the true grade where the probability that the rating is k is the same as for $k+1$ $(k = 1, 2)$. Via the POM in (2), we find the value of η that satisfies the following equalities for each combination of j and m:

$$
\begin{aligned}
2P(y_{ijm} > 2) &= P(y_{ijm} > 1) \\
2P(y_{ijm} > 1) &= 1 + P(y_{ijm} > 2)
\end{aligned}
\tag{9}
$$

4 RESULTS

4.1 Model-based Inferences

We applied the above methods to the dataset described. Figure 5 shows the point estimates and 95% posterior intervals (i.e., the middle 95% of the posterior distribution) of the thresholds for the eight raters for the nomenclature and the illustration methods with true grade η on the x-axis. The posterior means for the cutoff for each rater are denoted on the x-axis for raters $1, ..., 8$. The posterior distributions have varying standard deviations, suggesting greater imprecision in the estimates of some of the raters' cutoffs as compared to others. Rater 1's threshold of 0.52 is the lowest estimated lower threshold, whereas the largest is 1.93 by rater 4. The average lower threshold for nomenclature is 1.42. The range of estimates for the upper thresholds is 1.85 to 2.85, with a mean of 2.34. For the illustration method, the range of lower thresholds is 0.78 to 2.15, with a mean of 1.58. For the upper thresholds, the range is 2.04 to 3.06 with mean of 2.51.

The average thresholds are approximately 1.5 and 2.5 as we had hypothesized, but this is really due to the parameterization of the model. The key issue is the spread of the thresholds. Notice that the estimated cutoffs (i.e., the posterior means in figure 5) are relatively wide considering the range of η. For instance, in the nomenclature lower threshold, the difference between the largest and smallest thresholds is 1.41 which is considerably wide. As it turns out, all of the ranges are quite wide: the width of the other three ranges are 1.37, 1.00, and 1.02. As a result, we can conclude that there is significant spread among the thresholds.

Additional information is provided about how large or small the true variation may be among these raters in the posterior distributions of the standard deviation of the cutoffs. This is seen in figure 5A by noticing the relative spread of standard deviations and the ranges. The nomenclature lower cutoffs appear to have the largest variation in rater cutoffs with an estimated standard deviation of approximately 0.51, while the nomenclature upper cutoffs have the smallest variability, estimated at 0.32. Unlike the nomenclature threshold standard deviations, the variation in the cutoffs in the illustration methods are almost the same for the lower and upper cutoffs, with estimates of 0.39 and 0.42.

In looking more closely at the figure 5, we can see that raters 3, 4, 5, and 7 generally tend to have higher thresholds than raters 2, 6, and 8. We also see that rater 1 tends to have very low thresholds for the nomenclature definition, but relatively high thresholds for the illustrations. It was found upon further discussion with authors from Hruban et al. (2001), that rater 1 had misunderstood the rating system for nomenclature: instead of assigning both PanIN-1A and PanIN-1B a rating of 1, PanIN-2 a rating of 2, etc., he assigned PanIN-1B slides a rating of 2, PanIN-2 slides a rating of 3, etc. Therefore, his ratings of the slides using nomenclature tend to be biased upwards causing his thresholds to be quite low as compared to the other raters. The analyses were repeated excluding rater 1 from the dataset with results in figures 5B and 5. The estimates of thresholds for raters 2 through 7 remain almost the same, as can be seen in figure 5. However, in comparing figures 5A and 5B, we can see that the estimated standard deviations have changed: the standard deviation of the lower nomenclature cutoffs is much smaller (posterior mean of 0.52 versus 0.40), the standard deviation of the illustrations upper cutoff is smaller (0.41 versus 0.33), and the standard deviation of ilustrations lower and nomenclature upper cutoffs are approximately the same (0.44 versus 0.42, and 0.34 versus 0.30, respectively).

Based on the findings shown in figure 5B, we can conclude that both the nomenclature and the illustrations methods behave approximately the same. We see that there is greater variability in lower thresholds than in the upper thresholds, and that there is little difference in the nomenclature and illustrations standard deviations for a given cutoff. But, had we seen, for example, that the variation in the illustrations cutoffs was less than in the nomenclature cutoffs, we would have concluded that defining disease grade based on illustrations (versus nomenclature) is a better way of assigning grades. However, based on our model where rater 1 was removed, it appears that there is comparable variability in the two methods.

We also conclude that the variation in rater cutoffs tends to be high. For example, an estimated standard deviation of 0.40 and a mean lower threshold of 1.5 imply that in the population of raters (assuming normally distributed rater cutoffs), 95% of the lower cutoffs will be in the range (0.7, 2.3). Only 68% will be between 1.1 and 1.9. Therefore, many of the thresholds tend to be quite far from the the intended cutoff of 1.5.

4.2 Model Fit and Sensitivity Analysis

To demonstrate that the metric for the cutoffs is consistent with the original data scale and to provide a sense of model fit, we have plotted the predicted values of true disease score versus the average rating for each lesion, excluding observations from rater 1. This is shown in figure 5 in the left panel. The solid line is the $x = y$ line, which indicates perfect agreement between the average rating and our fitted value of η. Notice that there appear to be deviations from linearity at the extremes: when the average rating is very close to 1 or close to 3. Although this may appear as a deviation from model fit, this is actually consistent with the hypothesized model and the floor and ceiling effects imposed by the rating scale. Recall for example that the intended system assigns lesions with true disease states of 0.5 to 1.5 to the 1 category. In the case where all raters agree that the lesion is 1, which is the lowest possible average rating, it is likely (according to the model), that this lesion has a true disease grade of less than 1. The same occurs for lesions consistently graded as 3 by the raters: this lesion is likely to be on the high end of the 3 category, with a true disease score closer to 3.5.

Also included in the figure are the 95% posterior intervals for the fitted latent grade variable (shown as vertical lines), and the weighted least squares regression line, with weights chosen to reduce the floor and ceiling effects mentioned in the previous paragraph:

$$w_i = (0.5 - \mid q_i - 0.5 \mid)^2 \tag{10}$$

where w_i is the weight for slide i and q_i is the quantile for slide i, based on the average ranks. The correlation noted in the top left corner is based on the weighted least squares analysis. It is clear that the metrics match up quite well: in the left panel of figure 5 the regression line and the $x = y$ line are practically indistinguishable, and the correlation is > 0.99.. In this case, we can conclude that the fitted η values are on the same scale as the original data, as we had earlier hypothesized.

We performed several sensivitity analyses to see the effect of changing the prior distribution on η. Recall that $\eta \sim U[0.5, 3.5]$ in our model. But, as noted, this is not a necessary assumption and might be unrealistic in some cases depending on the sampling of subjects. To see the effect of this assumption, we tried two different priors:

(1) $\eta \sim N(2, 0.56)$, and (2) $\eta \sim N(2, 100)$. The first prior tested, with mean 2 and variance 0.56, is consistent with the distribution where most samples are truly in the 2 range, and 95% of samples are within 0.5 and 3.5. This could be another feasible distribution in our example. The second prior tested, with mean 2 and variance 100, is quite unrealistic, but still an interesting choice to see how the choice of prior affects the fitted model and inferences. The results based on using these two priors are shown in the center and right panels of figure 5.

The middle panel of 5 looks very similar to the left panel. We see that there is simply a slight shift up of the ratings compared to the $x = y$ line, however, the slope is approximately equal to 1 and the relationship is linear ($r = 0.99$). In the right panel, we see that allowing the variance of η to be large causes the relationship between the average ratings and the ηs to deviate from the $x = y$ line. However, it is critical to note that the relationship is still linear ($r = 0.98$). This implies that we can transform our results to the appropriate scale with a simple linear transformation of the ηs. We have done this, by pivoting the regression line so that it lies on top of the $x = y$ line. The linear transformed values are shown by the X's in the right panel. The transformation that achieves $x = y$ is based on $e_i + \bar{y_i}$ where e_i is the residual from the weighted least squares model and $\bar{y_i}$ is the average rating for the i^{th} slide. So, even with an inappropriate prior, the model can be salvaged to provide meaningful results. Note, however, that (a) the remainder of the model would need to be transformed to make inferences about the cutoffs shown in the Model Based Inferences section, and (b) it is much more sensible to choose a prior for η that is consistent with the scale under scrutiny. Due to our use of an MCMC approach for model fit, transformation of all parameters of interest to the appropriate scale is technically very simple, yet it adds another level of complexity to the analysis which is not particularly appealing.

5 DISCUSSION

We have presented a model that allows us to assess rater agreement in the case of multiple readers and multiple reading modalities. Our conclusion is that both the nomenclature method and the illustrations methods do about equally well for assigning ratings. However, we do find that there is quite a lot of variability in raters in both

methods. Especially when defining lesions of grade 1 from grade 2, raters appear to have quite different thresholds. This does not appear to be due to random chance, as rater biases tended to be consistent across methods and thresholds. For example, rater 2 tended to give high ratings for both methods and for both thresholds, while rater 4 tended to give low ratings. These results lead us to think that we might be able to improve the nomenclature method, for example, by more clearly distinguishing between grade 1 and grade 2 in the definition.

There are a number of appealing characteristics of the method that we have proposed for evaluating rater agreement with ordinal data arising from a latent variable. The POM discussed takes account of the ordinal (as opposed to continuous or nominal) nature of the data while assuming that disease state is a continuous latent variable. This is a natural assumption in many applications in medical research, with cancer staging being a prime example.

We see that our model is robust to outliers. Despite the inclusion of rater 1 in the model, the estimated thresholds for the other raters were not affected. The only area where inferences changed were in the case of the variation in thresholds, which was expected: assuming rater 1's ratings were not due to confusion of the rating system, the variation in the thresholds should be large.

The methods of interest (e.g., nomenclature and illustrations) can be directly compared to one another. In this case, we see by looking at the posterior distributions of the standard errors of the thresholds that there is little difference in the accuracy of these methods. Additionally, we evaluated the estimated method effects (δ_j and γ_j), finding that there were no method effects in the second analysis (i.e. removing rater 1). This suggests that both methods have the same average lower thresholds and the same average upper thresholds. Had we seen that estimates of δ_j or γ_j appeared to be significant, we could identify if one of the methods tended to give consistently lower or higher ratings than the other, and we could quantify the strength of the evidence for this finding.

Considering that the method effects did not appear to be significant in the analysis with rater 1 removed, we could simplify the model at this point, removing the method effects from the model. We have seen that they are not important, and as a result we woould gain parsimony and precision by removing them. Additionally, the model

presented can be changed in a number of ways to fit other related problems. We developed this model needing to account for both rater and method effects. However, this model would be perfectly appropriate in the case of just one of these effects. Similarly, additional effects can be added to the model, presuming there is enough information in the data to estimate all of the parameters of interest, and the number of thresholds can be increased. In summary, despite our specific application, the model presented is general and will fit many cases of ordinal ratings where an underlying continuous latent variable is assumed.

A very appealing part of our model is the interpretation of the parameters of interest. The parameters that we have focused on are the thresholds and their standard deviations which have intuitive meanings to clinicians and to statisticians. When comparing raters or comparing methods, the observed differences in thresholds are easily understood simply by looking at plots like figure 5 and figure 5. In other models where interpretation of the parameters of interest is complicated, it can be hard to put the results into a clinical perspective by keeping the results in the metric of the original scale. And, even with an inappropriate prior, we can reclaim the original scale through transformation of the estimated paramaters. This is a nice result of our finding that the model is insensitive to the choice of prior on η: we examined a uniform and two normal distributions, and all three showed very similar model fit. However, it is much simpler to choose an appropriate prior rather than transform model parameters back to the original scale post hoc.

From our model, we can see where improvements in the rating system can be made. In particular, as noted above, due to the relatively large variation in the lower thresholds, making the definitions and/or illustrations more precise may improve agreement. Additionally, we can identify particular raters who tend to be inconsistent with other raters, or who tend to have consistent biases in one direction. Perhaps in some settings, those raters could be retrained to adhere more closely to the PanIN standards.

References

Agresti, A. (1988), "A Model for Agreement Between Ratings on an Ordinal Scale," *Biometrics*, 44, 539–548.

Agresti, A. and Lang, J. B. (1993), "Quasi-Symmetric Latent Class Models, with Application to Rater Agreement," *Biometrics*, 49(1), 131–139.

Becker, M. P. and Agresti, A. (1992), "Log-Linear Modelling of Pairwise Interobserver Agreement on a Categorical Scale," *Statistics in Medicine*, 11, 101–114.

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Tables," *Educational and Psychological Measurement*, 20, 37–46.

— (1968), "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit," *Psychological Bulletin*, 70, 213–220.

Hruban, R., Adsay, N., Albores-Saavedra, J., Compton, C., Garrett, E., Goodman, S., Kern, S., Klimstra, D., Kloppel, G., Longnecker, D., Luttges, J., and Offerhaus, J. (2001), "Pancreatic Intraepithelial Neoplasia: A New Nomenclature and Classification System for Pancreatic Duct Lesions," *The American Journal of Surgical Pathology*, 25, 579–586.

Imperial College of Science, Technology and Medicine (2000), *WinBugs version 1.3*, Cambridge, UK.

Johnson, V. (1996), "On Bayesian Analysis of Multi-Rater Ordinal Data: An Application to Automated Essay Grading," *The Journal of the American Statistical Association*, 91, 42–51.

McCullagh, P. (1980), "Regression models for ordinal data (with discussion)," *Journal of the Royal Statistical Society, Series B*, 42, 109–142.

Perkins, S. M. and Becker, M. P. (2002), "Assessing Rater Agreement using Marginal Association Models," *Statistics in Medicine*, 21, 1743–1760.

Tanner, M. A. and Young, M. A. (1985), "Modeling Agreement Among Raters," *The Journal of the American Statistical Association*, 80, 175–180.

The R Development Core Team (2003), *R version 1.6.2*, StatLib: http://lib.stat.cmu.edu/R/CRAN.

Uebersax, J. S. and Grove, W. M. (1996), "A Latent Trait Finite Mixture Model for the Analysis of Rating Agreement," *Biometrics*, 49(3), 823–835.

Table 1: Interpretation of parameters in proportional odds model for rater agreement.

α_{j1}	lower nomenclature threshold for rater j.
α_{j2}	upper nomenclature threshold for rater j.
δ_j	difference between lower thresholds for nomenclature and illustrations for rater j.
$\delta_j + \gamma_j$	difference between upper thresholds for nomenclature and illustrations for rater j.

Figure 1: Example of differences in rater thresholds from the intended rating system.

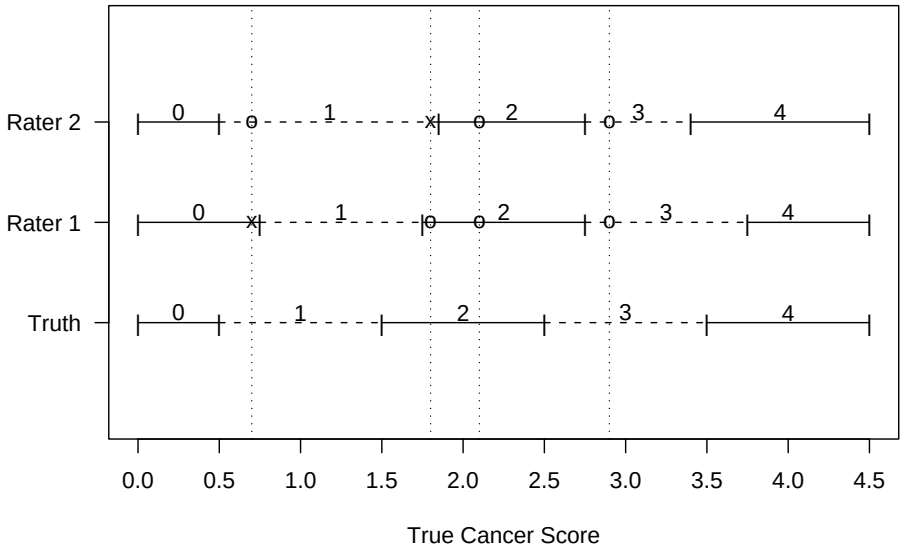
Figure 2: Nomenclature form of PanIN classification (left) and Illustration form of PanIn classification (right).

Figure 3: Point estimates and 95% posterior intervals of (A) nomenclature lower thresholds (black) and upper thresholds (gray), and (B) illustrations lower thresholds (black), and upper thresholds (gray).

Figure 4: Posterior distributions of standard deviations of rater thresholds (A) including all eight raters, and (B) with rater 1 removed.

Figure 5: Point estimates and 95% posterior intervals of (A) nomenclature lower thresholds (black) and upper thresholds (gray), and (B) illustrations lower thresholds (black), and upper thresholds (gray) with rater 1 removed from the analysis.

Figure 6: Model fit and sensitivity analysis: Model-based (posterior) estimate of true disease state (η) versus the average of ratings for each slide for different choices of priors for η. Left panel: $\eta \sim U[0.5, 3.5]$; Middle panel: $\eta \sim N(2, 0.56)$; right panel: $\eta \sim N(2, 100)$. The solid $x = y$ lines and dotted weighted least squares regression lines are plotted to show degree of association and departure from scale of original ratings. Vertical bars represent 95% credible intervals (i.e. the middle 95% of the posterior interval). In right panel, X's indicate the linear transformation of η to the original data scale.



Normal: The normal ductal and ductular epithelium is a cuboidal to low-columnar epithelium with amphophilic cytoplasm. Mucinous cytoplasm, nuclear crowding, and atypia are not seen.

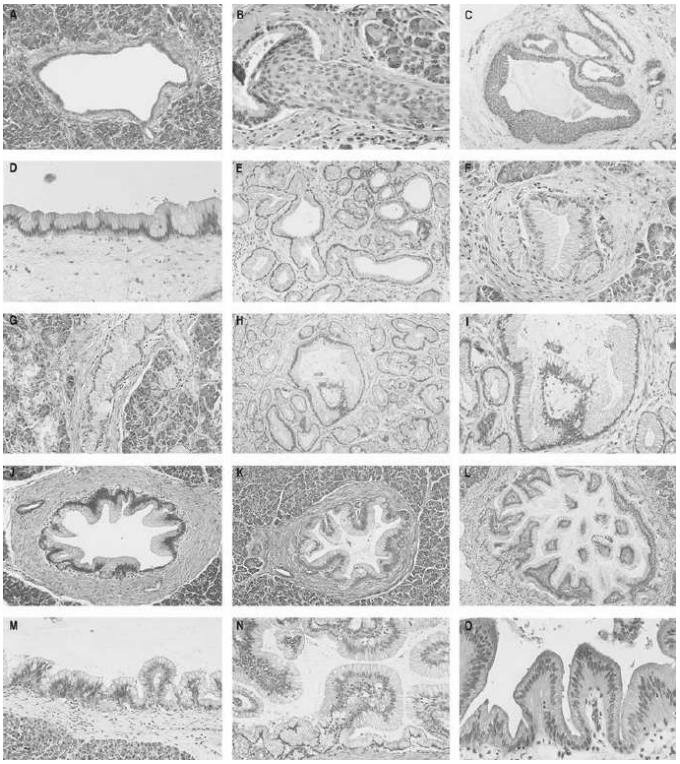
Squamous (transitional) metaplasia: A process in which the normal cuboidal ductal epithelium is replaced by mature stratified squamous or pseudostratified transitional epithelium without atypia.

PanIN-1A (pancreatic intraepithelial neoplasia 1-A): These are flat epithelial lesions composed of tall columnar cells with basally located nuclei and abundant supranuclear mucin. The nuclei are small and round to oval in shape. When oval, the nuclei are oriented perpendicular to the basement membrane. It is recognized that there may be considerable histologic overlap between non-neoplastic flat hyperplastic lesions and flat neoplastic lesions without atypia. Therefore, some may choose to designate these entities with the modifier term "lesion" ("PanIN/L-1A") to acknowledge that the neoplastic nature of many cases of PanIN-1A has not been unambiguously established.

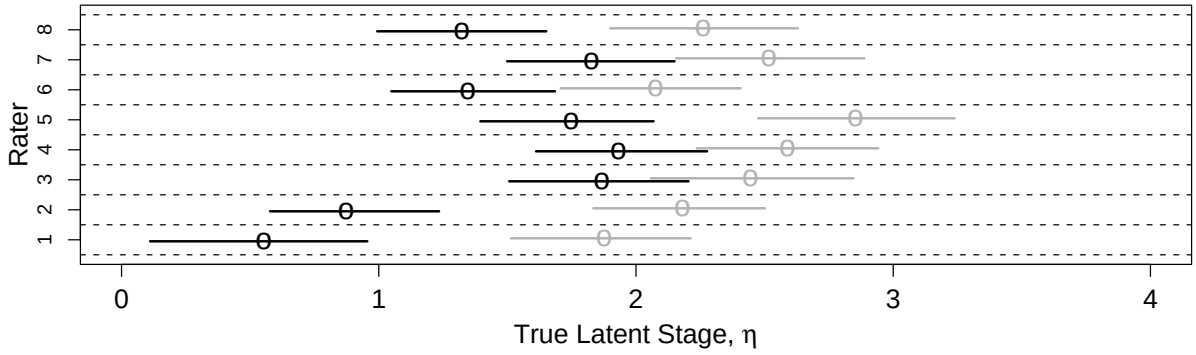
PanIN-1B (pancreatic intraepithelial neoplasia 1-B): These epithelial lesions have a papillary, micropapillary, or basally pseudostratified architecture but are otherwise identical to PanIN-1A.

PanIN-2 (pancreatic intraepithelial neoplasia 2): Architecturally these mucinous epithelial lesions may be flat but are mostly papillary. Cytologically, by definition, these lesions must have some nuclear abnormalities. These abnormalities may include some loss of polarity, nuclear crowding, enlarged nuclei, pseudo-stratification, and hyperchromatism. These nuclear abnormalities fall short of those seen in PanIN-3. Mitoses are rare, but when present are nonluminal (not apical) and are not atypical. True cribriform structures with luminal necrosis and marked cytologic abnormalities are generally not seen and, when present, should suggest the diagnosis of PanIN-3.

PanIN-3 (pancreatic intraepithelial neoplasia 3): Architecturally, these lesions are usually papillary or micropapillary; however, they may rarely be flat. True cribriforming, the appearance of "budding off" of small clusters of epithelial cells into the lumen, and luminal necrosis should all suggest the diagnosis of PanIN-3. Cytologically, these lesions are characterized by a loss of nuclear polarity, dystrophic goblet cells (goblet cells with nuclei oriented toward the lumen and mucinous cytoplasm oriented toward the basement membrane), mitoses that may occasionally be abnormal, nuclear irregularities, and prominent (macro) nucleoli. The lesions resemble carcinoma at the cytonuclear level, but invasion through the basement membrane is absent.



(A)



(B)