Study Design Considerations for Estimation of Agreement with Contributions of Variability to Establish Reliability in a Medical Device Pivotal Study

Mat D. Davis, MS
Jeffrey Joseph, MS
Theorem Clinical Research

September 23, 2013

Abstract

For a diagnostic imaging agent, the major considerations for efficacy analysis in a Phase III clinical trial are validity (sensitivity, specificity, agreement) and reproducibility of the result as used in clinical practice for diagnosis of disease or patient management. Masked image evaluation by multiple independent readers is performed to provide information about the reproducibility. In order to have a successful clinical development program for regulatory approval, both results need to be prospectively demonstrated. Sources of variation are misclassification (the agent result is inconsistent with the truth) and reader interpretation (each interprets the results with a certain sensitivity and specificity). Given the sensitivity and specificity of two readers, the intraclass correlation coefficient (ICC) and kappa will be determined. The effect of adding additional readers on the expected ICC will be demonstrated as well as the resulting difference if some readers interpret the results with poor sensitivity and specificity. The resulting bias and standard error of the ICC estimator from increasing the number of subjects and readers will be presented.

Keywords: Sensitivity, Specificity, Agreement, Intraclass Correlation, Kappa Statistic, Reliability

1 Introduction

The Guidance for Industry: Developing Medical Imaging Drug and Biological Products (2004) [1] is intended to assist developers of medical imaging agents in planning and coordinating their clinical investigations and preparing and submitting regulatory documents (INDs, NDAs, BLAs). The clinical development plan needs to be designed specifically to reflect use in diagnosis or monitoring of a disease as opposed to treating a disease. The major considerations for efficacy analysis in the pivotal trial are validity and reproducibility of the diagnostic agent result. Validity refers to the degree which the diagnostic agent accurately represents the underlying pathology of the phenomenon in question. Reproducibility ensures that repeating the same test multiple times will result in the same conclusion. For a successful clinical program, both results need to be prospectively demonstrated.

The validity of the diagnostic agent can be measured by sensitivity, specificity, accuracy, negative and positive predictive value comparing the diagnostic finding to a standard of truth.

The reproducibility of the diagnostic agent is assessed through the use of blinded image evaluation by multiple independent readers. This is performed to provide information on the

reproducibility of the diagnostic and accuracy parameters. These blinded read evaluations are designed to specifically answer the questions of inter-reader and intra-reader reliability. Inter-reader reliability refers to the measure by which multiple readers agree when interpreting the same scan, while intra-reader reliability measures the level of agreement of multiple diagnoses provided by the same reader on the same image. This discussion will be limited to the qualitative (binary outcome) where it is important just to know whether the disease is present or not.

When setting up a blinded image read evaluation, the balance between the number of readers per subject and the number of subjects necessary is always a study design issue. How many readers are reasonable for a blinded read image evaluation that would be representative of the medical community at large? How does the number of readers affect the number of subjects needed to ensure appropriate statistical power? These questions will be addressed and discussed.

Based on a given sensitivity and specificity, the agreement for qualitative outcomes can be measured by kappa statistics and intracluster correlation coefficients can be explicitly determined. In regards to validity measures of percent agreement, most confidence intervals will be constructed using methods appropriate for binomial proportions. These parameters will be helpful to provide a recommendation for the number of independent readers and number of subjects necessary for a clinical trial.

2 Motivation

A blinded image read evaluation needs to be performed to determine the acceptable inter-reader reliability for a new diagnostic imaging agent. Each image will be read by all independent readers and provide a binary response (positive or negative for disease). Each reader has an inherent sensitivity and specificity for determination if disease is present or not. In the past, a blinded image read evaluation would have 3 independent readers. Recently, the question has been raised as to whether inter-reader reliability be established with only 3 readers. If it cannot, how many independent readers are needed to establish inter-reader reliability?

Currently, methods exist to determine expected kappa [2, 3] and ICC [4] given an inherent sensitivity, specificity and prevalence of disease. While these methods are shown to accurately estimate expected agreement given those parameters, these methods are hindered by two limiting factors. First, the methods do not take into account that the two readers could potentially read at different levels of sensitivity and specificity, a common case in blinded image evaluation. Second, they do not consider the case where more than two raters evaluate the same image. In order to accurately plan for the appropriate number of readers needed for blinded image evaluation, these two factors need to be addressed to allow for accurate estimation of agreement among raters.

3 Methods

3.1 Clustered Variance of Repeated Measurements

Let y_{ij} be a binary response variable (1 for positive, 0 for negative) for subject i ($i : 1 \dots n$) from rater j ($j : 1 \dots m_i$), and let $\rho_{j,k}$ be the interclass correlation coefficient (ICC) between raters j and k when assessing the same target. Also, let y_i be the total number of positive

ratings per subject defined as

$$y_i = \sum_{j=1}^{m_i} y_{ij}$$

Then, given m_i ratings per subject, the resulting correlation matrix will be a symmetric m_i x m_i matrix in the form of

$$\Sigma_i = \begin{bmatrix} 1 & \rho_{1,2} & \dots & \dots & \rho_{1,m_i} \\ \rho_{1,2} & \ddots & \dots & \dots & \vdots \\ \vdots & \dots & 1 & \dots & \vdots \\ \vdots & \dots & \dots & \ddots & \rho_{m_{i-1},m_i} \\ \rho_{1,m_i} & \dots & \dots & \rho_{m_{i-1},m_i} & 1 \end{bmatrix}$$

Assuming that each observation is drawn from the Bernoulli distribution with probability π, the variance of y_i can be written as

$$\text{Var}(y_i) = m_i \pi (1-\pi) [1 + (m_i - 1)\overline{\rho}], \text{ where } \overline{\rho} = \frac{\sum_{j=1}^{m_i} \sum_{\substack{k=1 \\ j<k}}^{m_i} \rho_{jk}}{\binom{m_i}{2}} \tag{1}$$

Under the assumption that $m_i = m \ \forall \ i$, which is not an unreasonable assumption in the controlled clinical trial setting, and that all subjects are independent, the variance of $y = \sum_{i=1}^{n} y_i$ can be written as

$$\text{Var}(y) = n m_i \pi (1-\pi) [1 + (m_i - 1)\overline{\rho}] \tag{2}$$

Therefore the cluster-level ICC is in fact an average of the ICC between each unique pair of readers. Therefore, specifying the ICC between each pair of readers should be sufficient to determine the overall ICC.

3.2 Determining ICC Using Sensitivity and Specificity

The ICC and the kappa statistic are closely linked in determining inter-rater agreement. Research has already been conducted examining the effect of sensitivity and specificity on determining the kappa statistic. Given the sensitivity (S) and specificity (C) of two readers (assumed to be equivalent between the two raters) in addition to the prevalence of disease (π), Küchenhoff et. al. (2012) [3] determined the following expression for the kappa statistic (κ)

$$\kappa = \frac{\pi (1-\pi) (S + C - 1)^2}{(C - \pi (S + C - 1)) (1 - C + \pi (S + C - 1))} \tag{3}$$

From the following definition of the ICC

Similar results are achieved for the intraclass correlation coefficient. First, recall the definition of the ICC:

$$\rho_{j,k} = \frac{P(y_{ij} = 1, y_{ik} = 1) - P(y_{ij} = 1) P(y_{ik} = 1)}{\sqrt{P(y_{ij} = 1)(1 - P(y_{ij} = 1)) P(y_{ik} = 1)(1 - P(y_{ik} = 1))}} \tag{4}$$

As is standard practice, sensitivity will be defined as the probability of detecting disease given the subject is truly diseased whereas specificity will be defined as the probability of declaring a subject disease-free given the subject is truly free of disease. Let D_i be the disease status for subject i. Then D_i is defined as

$$D_i = \begin{cases} 1: & \text{Subject is diseased} \\ 0: & \text{Subject is disease-free} \end{cases}$$

In order to proceed, the assumption of independence on two different ratings on the same subject given the true disease status is made, that is

$$P\left(y_{ij}=x_{ij},y_{ik}=z_{ik}|D_i=d_i\right)=P\left(y_{ij}=x_{ij}|D_i=d_i\right)P\left(y_{ik}=z_{ik}|D_i=d_i\right) \tag{5}$$

This assumption is reasonable in most circumstances. Once a rater is given the true disease status of a subject, other ratings on that subject should not influence their decision. On the other hand, if the true disease status is not available, other ratings on the same subject could make an impact on the specific rater's decision.

Therefore, using the assumption of conditional independence in addition to the law of total probability, the following properties can be derived:

$$\begin{aligned}
P\left(y_{ij}=x_{ij}\right) &= P\left(y_{ij}=x_{ij}|D_i=1\right)P\left(D_i=1\right)\\
&\quad +P\left(y_{ij}=x_{ij}|D_i=0\right)P\left(D_i=0\right)\\
P\left(y_{ij}=x_{ij},y_{ik}=z_{ik}\right) &= P\left(y_{ij}=x_{ij},y_{ik}=z_{ik}|D_i=1\right)P\left(D_i=1\right)\\
&\quad +P\left(y_{ij}=x_{ij},y_{ik}=z_{ik}|D_i=0\right)P\left(D_i=0\right)\\
&= P\left(y_{ij}=x_{ij}|D_i=1\right)P\left(y_{ik}=z_{ik}|D_i=1\right)P\left(D_i=1\right)\\
&\quad +P\left(y_{ij}=x_{ij}|D_i=0\right)P\left(y_{ik}=z_{ik}|D_i=0\right)P\left(D_i=0\right)
\end{aligned}$$

Using these properties, the ICC can be rewritten in terms of the sensitivity and specificity for each rater as well as the overall prevalence of disease. Let $S_j=P\left(y_{ij}=1|D_i=1\right)$ be the sensitivity of rater j, $C_j=P\left(y_{ij}=0|D_i=0\right)$ be the specificity of rater j and $\pi=P\left(D_i=1\right)$ be the prevalence of disease. Then, starting from the definition in equation (4), the ICC between two raters can be written as follows:

$$\begin{aligned}
p^*_{iz} &= P\left(y_{iz}=1|D_i=1\right)P\left(D_i=1\right)+P\left(y_{iz}=1|D_i=0\right)P\left(D_i=0\right)\\
&= S_z\pi+\left(1-C_z\right)\left(1-\pi\right)\\
p^*_{ix,iz} &= P\left(y_{ix}=1|D_i=1\right)P\left(y_{iz}=1|D_i=1\right)P\left(D_i=1\right)\\
&\quad +P\left(y_{ix}=1|D_i=0\right)P\left(y_{iz}=1|D_i=0\right)P\left(D_i=0\right)\\
&= S_xS_z\pi+\left(1-C_x\right)\left(1-C_z\right)\left(1-\pi\right)\\
\rho_{j,k} &= \frac{p^*_{ij,ik}-p^*_{ij}p^*_{ik}}{\sqrt{p^*_{ij}p^*_{ik}\left(1-p^*_{ij}\right)\left(1-p^*_{ik}\right)}}
\end{aligned}$$

Assuming that the sensitivity and specificity for each subject are equal, the ICC between the two raters would be equal to κ as displayed in equation 3.

$$\begin{aligned}
\rho_{j,k} &= \frac{S^2\pi+\left(1-C\right)^2\pi^2-\left(S\pi+\left(1-C\right)\pi\right)^2}{\left(S\pi+\left(1-C\right)\pi\right)^2\left(1-S\pi-\left(1-C\right)\pi\right)^2}\\
&= \frac{\pi\left(1-\pi\right)\left(S+C-1\right)^2}{\left(C-\pi\left(S+C-1\right)\right)\left(1-C+\pi\left(S+C-1\right)\right)}\\
&= \kappa \text{ [from equation (3)]}
\end{aligned}$$

Therefore, using sensitivity and specificity of raters to determine the appropriate ICC is more flexible than that of determining κ since the ICC allows for different sensitivity and specificity for each rater, yet is equivalent to κ when the sensitivity and specificity for each rater are the same.

Given these results, the overall cluster-level ICC is expressed as

$$\overline{\rho}=\frac{\sum_{j=1}^{m_i}\sum_{\substack{k=1\\j<k}}^{m_i}\frac{p^*_{ij,ik}-p^*_{ij}p^*_{ik}}{\sqrt{p^*_{ij}p^*_{ik}\left(1-p^*_{ij}\right)\left(1-p^*_{ik}\right)}}}{\binom{m_i}{2}} \tag{6}$$

3.3 Sensitivity, Specificity and Probability of Agreement

While agreement among readers is important to show the reliability of the test in question, circumstances arise where the probability of perfect or near-perfect agreement is of interest. For m raters and $z+1$ possible ordinal outcomes, the probability of perfect agreement can be written as

$$P_{perf} = \sum_{x=0}^{z} P\left(y_{i1} = x, y_{i2} = x, \ldots, y_{im} = x\right)$$

With binary outcomes, the rating x can only take the value 1 or 0. Under the framework of conditional independence specified in the previous section, this probability can also be expressed in terms of sensitivity, specificity and prevalence of disease.

$$\begin{aligned}
P_{perf} =& \sum_{x=0}^{1} P\left(y_{i1} = x, y_{i2} = x, \ldots, y_{im} = x | D_i = 1\right) P\left(D_i = 1\right) \\
&+ \sum_{x=0}^{1} P\left(y_{i1} = x, y_{i2} = x, \ldots, y_{im} = x | D_i = 0\right) P\left(D_i = 0\right) \\
&= \sum_{x=0}^{1} \left(\prod_{j=1}^{m} P\left(y_{ij} = x | D_i = 1\right) \right) P\left(D_i = 1\right) \\
&+ \sum_{x=0}^{1} \left(\prod_{j=1}^{m} P\left(y_{ij} = x | D_i = 0\right) \right) P\left(D_i = 0\right) \\
&= \left(\prod_{j=1}^{m} S_j \right) \pi + \left(\prod_{j=1}^{m} (1 - C_j) \right) (1 - \pi) \\
&+ \left(\prod_{j=1}^{m} (1 - S_j) \right) \pi + \left(\prod_{j=1}^{m} C_j \right) (1 - \pi) \\
&= \left(\prod_{j=1}^{m} S_j + \prod_{j=1}^{m} (1 - S_j) \right) \pi + \left(\prod_{j=1}^{m} (1 - C_j) + \prod_{j=1}^{m} C_j \right) (1 - \pi)
\end{aligned}$$

Similar arguments could be used to find the probability of near-perfect agreement (the probability that $m-1$ of m raters agree) or any specified level of agreement among raters. Upon estimating the expected probability of perfect agreement, the degree of error aournd the estimated proportion could be attained using one of the myriad of methods to obtain confidence intervals around a binomial proportion.

4 Simulations

4.1 Simulation Methods

In order to simulate data that mimics the situations previously referenced, the following input parameters are needed:

1. Number of subjects (n)
2. Number of raters per subject (m)
3. Prevalence of disease (p for disease status d_i)

4. Sensitivity and specificity of each rater (S_i,C_i)

The simulations were carried out for this analysis by first generating each subject's disease status (d_i for n subjects based on p) using a Bernoulli distribution. Second, a rating for each subject based on the rater's individual sensitivity and specificity was generated from a Bernoulli distribution with S_i and C_i as success probabilities given the subject's disease status d_i.

Simulations were carried out using R programming software [5]. The package Rlab [6] was used to sample values from the Bernoulli distribution. At each combination of parameters, 5000 simulations were carried out and the estimate, standard error, bias and coverage of each were captured and summarized. In addition, percent perfect and near-perfect agreement (defined as exactly one rater differing from all other raters) were recorded.

Given these data, agreement among raters was calculated in three distinct fashions:

- The beta-binomial ICC was calculated among all raters within a subject assuming the beta-binomial distribution. This estimate was obtained by finding the value of $\overline{\rho}$ that sets the gradient of the maximum likelihood [7] equation to zero with assumed prevalence

$$\hat{p} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij}}{\sum_{i=1}^{n} m_i} \tag{7}$$

  The standard error was calculated to be the inverse square root of the Fisher information matrix assuming the same likelihood. The 95% confidence interval was calculated as

$$\left[\hat{\overline{\rho}} - Z_{1-\alpha} \cdot SE\left(\hat{\overline{\rho}}\right), \hat{\overline{\rho}} + Z_{1-\alpha} \cdot SE\left(\hat{\overline{\rho}}\right)\right] \tag{8}$$

- The overall ICC was then calculated by determining the pair-wise ICC for each set of readers. This ICC was calculated by directly applying the formula given in (4) to the each pair of readers, then averaging all results to obtain $\hat{\overline{\rho}}$. Because of the efficiency properties of the variance estimate derived from the beta-binomial distribution [8], the same formula was used to calculate the standard error and confidence interval around $\hat{\overline{\rho}}$.

- Finally, Cohen's kappa statistic was computed each pair-wise set of readers using the standard formula

$$\kappa = \frac{p_o - p_c}{1 - p_c} \tag{9}$$

  where p_o is the observed percent agreement and p_c is the chance percent agreement [9]. As shown earlier, κ and ρ should be theoretically equivalent when assessing the same data and is therefore reasonable to assume that equivalent variance and confidence interval assumptions should work on both. Therefore, the method of obtaining the variance for the previous two ICC's will also be used to determine the variance of κ.

Figure 1: Simulation Results for Bias

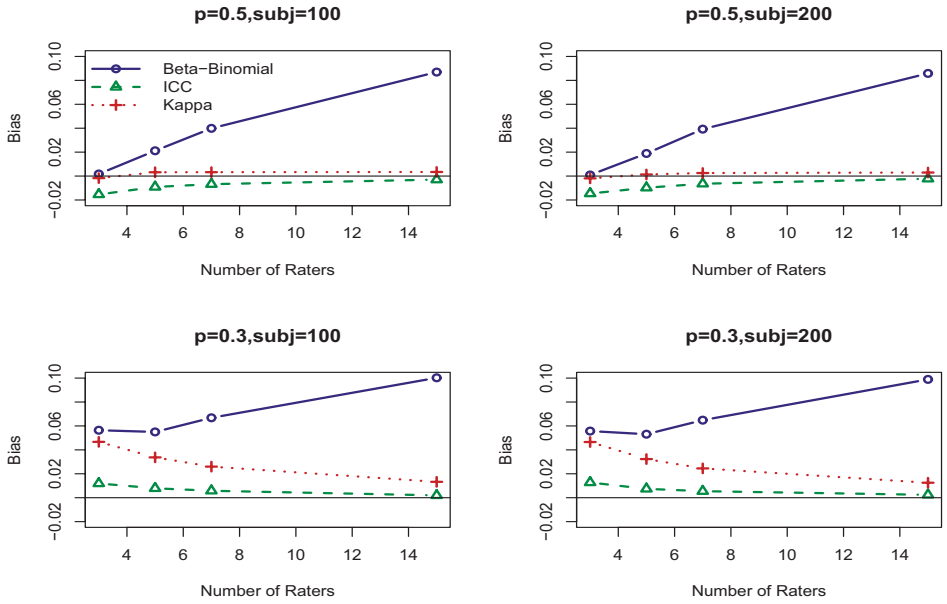
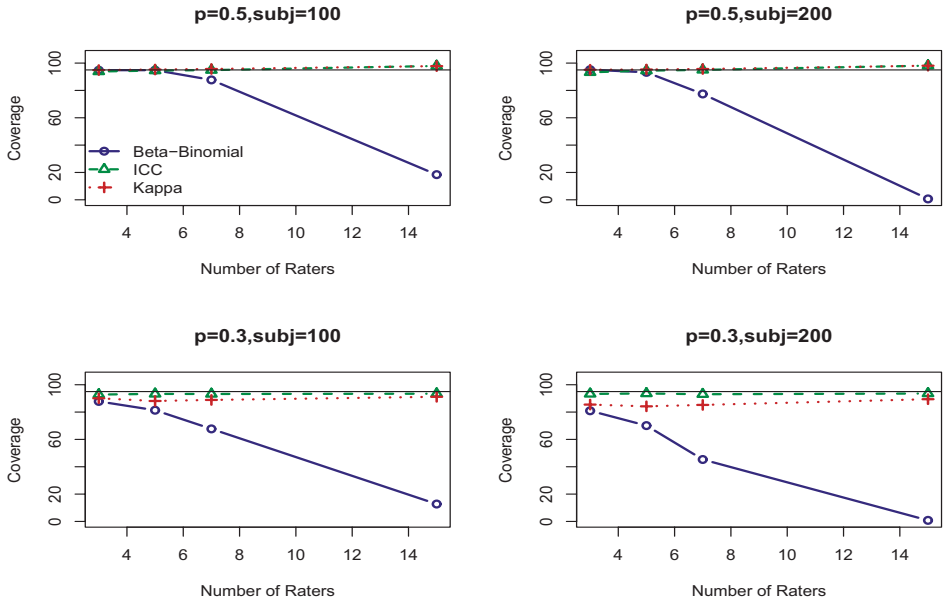


Figure 2: Simulation Results for Coverage



4.2 Simulation Results

4.2.1 Point Estimate

When the prevalence is set at 0.5, all three estimates have neglible bias. However, as the number of raters increases, the bias for the beta-binomial estimate increases, while the bias tends toward zero for both the pairwise ICC and kappa estimates. However, when the

prevalence decreases to 0.3, both the beta-binomial estimate and the pairwise kappa estimate underestimate the true ICC. Presumably, this difference is due to the misspecification of the true prevalence. As the sensitivities and specificities are not set to 1, the observed prevalence is different that the true prevalence. Neither method is able to account for this difference. However, the pairwise ICC estimate outlined earlier is able to account for this misspecification and more accurately estimates the true ICC.

4.2.2 Standard Error and Coverage

As expected, as either the number of subjects or the number of raters increase, the standard error decreases. For the pairwise ICC estimate, the coverage of the 95 % confidence interval is close to the nominal 95% level. The pairwise kappa performs less optimally as the bias increases. As mentioned before, the bias for this estimate when the prevalence is close to 0.5 is close to zero, and the coverage performs as expected. For the beta-binomial estimate, as mentioned before, the bias increases as the number of raters increases. Therefore, the coverage tends toward zero as the number of raters increases.

5 Conclusion

In most cases, it is a reasonable expectation that all readers within a trial will not read images with the same sensitivity and specificity. If that is the case, the standard ICC methods will probably be inadequate to estimate the true level of agreement among raters in the study. If this is expected, then the pairwise ICC method of determining overall agreement should be used instead of standard methods. In addition, the use of the efficient variance obtained from the beta-binomial distribution to obtain a 95% confidence interval around the pairwise ICC estimate achieves the expected 95% level and is an acceptable confidence level for the measure.

Regarding the number of raters needed for the study, the only benefit to increasing the number of raters needed to determine agreement is to shorten the confidence interval width of the estimator. However, this effect can also be achieved by increasing the number of subjects in the study. Therefore the balance between increasing the number of subjects and the number of raters should be based solely on desired confidence interval width and other clinical factors.

Table 1: Simulation Results

Simulation Results with $p = 0.5$									
			N = 100			N = 200			
Raters	True ICC	Result	BB ICC	PW ICC	PW Kappa	BB ICC	PW ICC	PW Kappa
3	.449	E/S	.447 (.066)	.464 (.063)	.451 (.066)	.448 (.047)	.464 (.045)	.451 (.047)
		B/C	.002 (94.8)	-.015 (93.9)	-.002 (94.9)	.001 (95.2)	-.015 (93.3)	-.002 (95.1)
5	.465	E/S	.444 (.050)	.475 (.047)	.462 (.048)	.447 (.035)	.475 (.033)	.464 (.034)
		B/C	.021 (94.8)	-.009 (94.5)	.003 (95.3)	.019 (93.2)	-.001 (94.5)	.001 (95.3)
7	.472	E/S	.433 (.043)	.479 (.040)	.469 (.041)	.433 (.031)	.479 (.028)	.470 (.029)
		B/C	.040 (87.7)	-.007 (94.8)	.003 (95.7)	.039 (77.4)	-.006 (95.0)	.003 (95.7)
15	.482	E/S	.395 (.034)	.485 (.031)	.478 (.032)	.396 (.024)	.484 (.022)	.479 (.022)
		B/C	.087 (18.4)	-.003 (97.9)	.003 (97.8)	.086 (.006)	-.002 (98.1)	.003 (98.1)
Simulation Results with $p = 0.3$								
			N = 100			N = 200		
3	.422	E/S	.365 (.073)	.410 (.066)	.375 (.071)	.366 (.052)	.409 (.047)	.375 (.051)
		B/C	.056 (87.8)	.012 (92.8)	.047 (90.1)	.056 (80.9)	.013 (93.3)	.047 (85.5)
5	.432	E/S	.377 (.052)	.424 (.048)	.398 (.050)	.379 (.037)	.424 (.034)	.399 (.036)
		B/C	.055 (81.3)	.008 (93.3)	.034 (88.1)	.053 (70.1)	.007 (93.6)	.032 (84.2)
7	.436	E/S	.369 (.044)	.430 (.041)	.410 (.042)	.371 (.031)	.430 (.029)	.411 (.030)
		B/C	.067 (67.7)	.006 (93.2)	.026 (88.9)	.065 (45.3)	.005 (93.1)	.025 (.85.2)
15	.442	E/S	.341 (.033)	.440 (.031)	.428 (.032)	.343 (.023)	.439 (.022)	.429 (.022)
		B/C	.100 (12.7)	.002 (93.4)	.013 (91.2)	.099 (0.8)	.002 (93.6)	.013 (89.4)

All raters had sensitivity/specificity of 85% except for one with sens/spec 70%/90%
N: Number of subjects, BB: Beta-Binomial, PW: Pairwise
E/S: Estimate (SE), B/C: Bias (95% CI Coverage)

References

[1] U. D. of Health, F. Human Services, C. f. D. E. Drug Administration, C. f. B. E. Research (CDER), and R. (CBER), "Guidance for industry: Developing medical imaging drug and biological products," 2004.

[2] I. Gardner, H. Stryhn, P. Lind, and M. T. Collins, "Conditional dependence between tests affects the diagnosis and surveillance of animal diseases," *Preventive Veterinary Medicine*, vol. 45, pp. 107–122, 2000.

[3] H. Küchenhoff, T. Augustin, and A. Kunz, "Partially identified prevalence estimation under misclassification using the kappa coefficient," *International Journal of Approximate Reasoning*, 2012.

[4] A. Branscum, I. Gardner, B. Wagner, P. McIntruff, and M. Salman, "Effect of diagnostic testing error on intracluster correlation coefficient estimation," *Preventive Veterinary Medicine*, vol. 69, pp. 63–75, 2005.

[5] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[6] D. D. Boos, A. D. Brooks, and D. Nychka, *Rlab: Functions and Datasets Required for ST370 class*, 2009. R package version 2.9.0.

[7] D. Smith, "Algorithm AS 189: Maximum likelihood estimation of the parameters of the beta binomial distribution," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 32, no. 2, pp. 196–204, 1983.

[8] M. S. Ridout, C. G. Demetrio, and D. Firth, "Estimating intraclass correlation for binary data," *Biometrics*, vol. 55, no. 1, pp. 137–148, 1999.

[9] J. Cohen *et al.*, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.