Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2007 *Paper* 22

Coronary Evaluation Using Multi-detector Spiral Computed Tomography Angiography: Statistical Design and Analysis

William F. McCarthy* Douglas R. Thompson†

Bruce A. Barton‡

*Maryland Medical Research Institute, dr.w.f.mccarthy@gmail.com

†Maryland Medical Research Institute, dthompson@mmri.org

‡Maryland Medical Research Institute, bbarton@mmri.org

Coronary Evaluation Using Multi-detector Spiral Computed Tomography Angiography: Statistical Design and Analysis

William F. McCarthy, Douglas R. Thompson, and Bruce A. Barton

Abstract

Contrast-enhanced multi-detector row spiral computed tomography (MDCT) has been introduced as a method for non-invasive visualization of coronary artery stenosis. To determine the diagnostic accuracy of MDCT coronary angiography, as compared to the "gold standard" invasive coronary angiography, sensitivity and specificity are estimated (95% Confidence Intervals). Three separate levels of estimation are computed: at the patient level, at the coronary artery level, and at the coronary artery segment level. We review the methodology for the estimation of sensitivity and specificity of non-clustered binary data (patient level analysis) and present a methodology for the estimation of sensitivity and specificity that considers the patient as a cluster and the coronary arteries (or coronary artery segments) as the diagnostic units of the study (DUOS) within each cluster. We also present how to estimate the weighted kappa for the comparison of ordinal measures of stenosis when non-clustered and clustered data are considered and the mean difference for the comparison of continuous measures of stenosis when non-clustered and clustered data are considered. Finally, we present a methods for determining the statistical precision of estimates sensitivity and specificity, weighted kappa and mean difference when clustered data are considered.

I. Introduction

Contrast-enhanced multi-detector row spiral computed tomography (MDCT) has been introduced as a method for non-invasive visualization of coronary artery stenosis. To determine the diagnostic accuracy of MDCT coronary angiography, as compared to the "gold standard" invasive coronary angiography, sensitivity and specificity are estimated (95% Confidence Intervals). Three separate levels of estimation are computed: at the patient level, at the coronary artery level, and at the coronary artery segment level. We review the methodology for the estimation of sensitivity and specificity of non-clustered binary data (patient level analysis) and present a methodology for the estimation of sensitivity and specificity that considers the patient as a cluster and the coronary arteries (or coronary artery segments) as the diagnostic units of the study (DUOS) within each cluster. We also present how to estimate the weighted kappa for the comparison of ordinal measures of stenosis when non-clustered and clustered data are considered and the mean difference for the comparison of continuous measures of stenosis when non-clustered and clustered data are considered. Finally, we present a methods for determining the statistical precision of estimates sensitivity and specificity, weighted kappa and mean difference when clustered data are considered.

The conventional binomial variance estimate [Equations 1.10, 1.11], which assumes that all measurements are independent, will not be valid for calculating the variance of sensitivity and specificity of clustered binary data. As an example, consider the assessment of sensitivity. When considering clustered binary data, the point estimate of sensitivity [Equation 1.15] gives the same estimate that would result if we had ignored the clustering, assumed independence, and used the conventional binomial point estimate [Equation 1.8]. However, if the variance of the sensitivity is estimated using the conventional binomial variance estimate [Equation 1.10], the variance will be underestimated if the correlation between the diagnostic units is positive (i.e., the estimated variance will be smaller than the true variance) or will be overestimated if the correlation between the diagnostic units is negative (i.e., the estimated variance will be larger than the true variance). The most likely scenario for our study is that the correlation between the diagnostic units is positive. A positive correlation indicates that if one diagnostic unit tests as a true positive (true negative) in a particular cluster, the probability of testing other diagnostic units in the same cluster as true positives (true negatives) will increase. To correct for this underestimated/overestimated variance problem (bias), a ratio estimator for the variance of clustered binary data has been derived [Equation 1.16] (Cochran, 1977; Rao and Scott, 1992).

In a talk presented by J M Bland to the RSS Medical Section and the RSS Liverpool Local Group, 12 NOV 2003, Bland noted that the magnitude of the effect of clustering is measured by the design effect, *Deff*, given by the following: $Deff = 1 + (n - 1)(ICC)$ where n is the number of observations in a cluster and ICC is the intra-cluster correlation coefficient. The ICC is the correlation between pairs of subjects chosen at random from the same cluster. If n=1, cluster size one, in other words, no clustering, then $Deff=1$, otherwise *Deff* will exceed 1 (this assumes a positive correlation between pairs of subjects). In analysis, if we analyse the clustered data as if there were no clusters, the variances of the estimates must be multiplied by *Deff*, hence the standard error must be multiplied by the square root of Deff.

II. Definitions and Terms

The **sensitivity** of a diagnostic test is its ability to detect the condition of interest when it is **present** in patients.

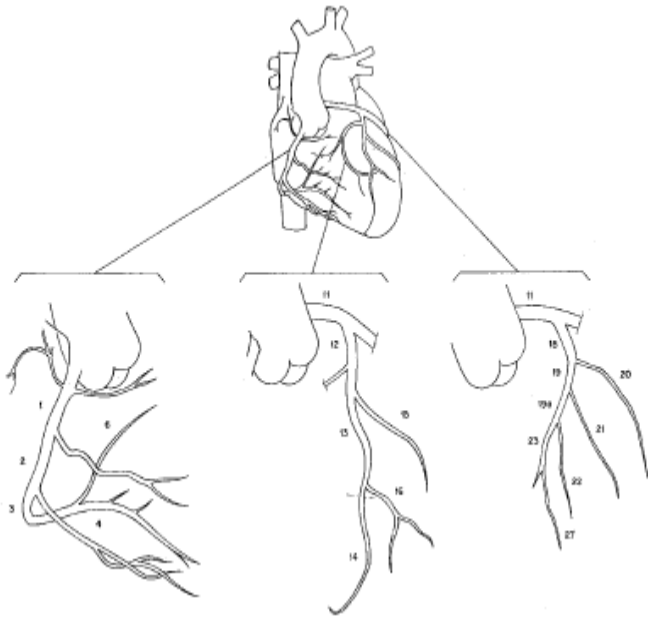
The **specificity** of a diagnostic test is its ability to exclude the condition of interest in patients **without** the condition.

The **diagnostic unit of study** (DUOS) is the smallest "unit" that is tested (Beam, 1998).

Gold standard: we determine the true condition status by means of a gold standard – "a source of information completely different from the test or tests under evaluation and which tells us the true condition status of the patient" (Zhou et al, 2002). For this paper, we shall say there is concordance between the test being evaluated and the "gold standard" if the same (identical) coronary artery or same (identical) coronary artery segment has the event of interest by both the test being evaluated and the "gold standard". For example, if coronary artery segment 4 is determined to be positive by the test being evaluated as well as by the "gold standard", we have concordance.

Stenosis: Narrowing or constriction of a coronary artery (coronary artery segment).

Coronary Tree with coronary arteries and coronary artery segments (Wang et al, 2004).



III. Conventional Binomial Estimator Method: Estimation in a Single Sample, Binary-Scale Data

This method is used for the patient level analysis. It is a non-clustered method. As an example, we will use the following binary-scale event definition: Event= Patient has a $\geq 50\%$ stenosis and Non-event= Else. The same (identical) coronary artery or same (identical) coronary artery segment has the event. Only one coronary artery with an event or one coronary artery segment with an event is required. Table 1 shows how the non-clustered data for binary diagnostic test results can be displayed.

The true condition status of the patient is determined by the use of the "gold standard".

The sensitivity and specificity estimates are:

$$\hat{Se} = \frac{s_1}{n_1} \qquad \text{sensitivity} \qquad [1.8]$$

and

$$\hat{Sp} = \frac{r_0}{n_0} \qquad \text{specificity} \qquad [1.9]$$

The variance of the sensitivity and specificity is the variance of a proportion:

$$\hat{Var}\left(\hat{Se}\right) = \frac{\hat{Se}(1-\hat{Se})}{n_1} = \frac{s_1 s_0}{n_1^3} \qquad [1.10]$$

and

$$\hat{Var}\left(\hat{Sp}\right) = \frac{\hat{Sp}(1-\hat{Sp})}{n_0} = \frac{r_1 r_0}{n_0^3} \qquad [1.11]$$

Table 1. Data Layout for Non-Clustered Binary-Scale Data

		Test Result		
True Condition Status	Positive (T=1)		Negative (T=0)	Total
Present (D=1)	s_1		s_0	n_1
Absent (D=0)	r_1		r_0	n_0
Total	m_1		m_0	N

The usual approach to constructing a confidence interval for a measure of diagnostic accuracy assumes a large sample size (i.e., follows a normal distribution asymptotically). This confidence interval is referred to as an asymptotic interval. It has the following form:

$$\hat{\Theta} - z_{1-\alpha/2}\sqrt{\hat{Var}\left(\hat{\Theta}\right)}\ ,\ \hat{\Theta} + z_{1-\alpha/2}\sqrt{\hat{Var}\left(\hat{\Theta}\right)}$$

[1.12]

where $\hat{\Theta}$ is the estimate of the accuracy measure (either sensitivity or specificity), Θ; $z_{1-\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution, and $100(1-\alpha)\%$ is the confidence level.

Agresti and Coull (1998) have noted that the "Wald [confidence] interval performs poorly unless n is quite large (e.g., Ghosh, 1979; Blyth and Still, 1983)". Construction of a confidence interval based on Equation 1.12 and using Equations 1.8 and 1.10 and Equations 1.9 and 1.11, is based on the Wald confidence interval.
If the sample size is small, then the confidence limits for the sensitivity are estimated with the following equation (Agresti and Coull, 1998):

$$\frac{\hat{Se} + z^2_{1-\alpha/2}/(2n_1) \pm z_{1-\alpha/2}\sqrt{[\hat{Se}(1-\hat{Se}) + z^2_{1-\alpha/2}/(4n_1)]/n_1}}{1 + z^2_{1-\alpha/2}/n_1}$$

[1.13]

If the sample size is small, then the confidence limits for the specificity are estimated with the following equation (Agresti and Coull, 1998):

$$\frac{\hat{Sp} + z^2_{1-\alpha/2}/(2n_0) \pm z_{1-\alpha/2}\sqrt{[\hat{Sp}(1-\hat{Sp}) + z^2_{1-\alpha/2}/(4n_0)]/n_0}}{1 + z^2_{1-\alpha/2}/n_0}$$

[1.14]

IV. Estimation in a Single Sample: Clustered Binary-Scale Data

As an example, we will use the following binary-scale event definitions: Event= same (identical) coronary artery segment vessel with $\geq 50\%$ stenosis and Non-Event= else. All coronary artery segments with an event are considered.

Sensitivity and Specificity

If one assesses sensitivity and specificity of a diagnostic test when clustered binary-scaled data is used, the data layout presented in Table 1 is modified as follows in Tables 2a and 2b:

Table 2a: Data Layout for the Assessment of Sensitivity

Patient (cluster)	No. of TP_i	No. of Segments (N_i) with Event	$\hat{Se}_i$	$N_i / \bar{N}$	$(N_i / \bar{N})^2 (\hat{Se}_i - \hat{Se})^2$

Each patient is considered a cluster (i=1,2,..,I clusters); within the cluster are a number of segments.

No. of TP_i refers to the number of segments that are test positives in cluster i, as determined by the test being evaluated (MDCT angiography).
No. of Segments (N_i) with Event refers to the number of segments that are true positives in cluster i, as determined by the "gold standard" (invasive coronary angiography).

$\hat{Se}_i$ = No. of TP_i / No. of segments with event (N_i).

$N_i / \bar{N}$ = the cluster size for patient$_i$ / mean cluster size.

$$\hat{Se} = \frac{\sum_{i=1}^{I} N_i \hat{Se}_i}{\sum_{i=1}^{I} N_i} \qquad [1.15]$$

$$\hat{Var}(\hat{Se}) = \frac{1}{I(I-1)} \sum_{i=1}^{I} \left[\left(N_i / \bar{N} \right)^2 (\hat{Se}_i - \hat{Se})^2 \right]$$

[1.16]

where $\bar{N} = \Sigma \frac{N_i}{I}$ is the mean cluster size.

Table 2b: Data Layout for the Assessment of Specificity

Patient (cluster)	No. of TN_i	No. of Segments (N_i) without Event	$\hat{Sp}_i$	$N_i / \bar{N}$	$(N_i / \bar{N})^2 (\hat{Sp}_i - \hat{Sp})^2$

Each patient is considered a cluster (i=1,2,..,I clusters); within the cluster are a number of segments.

No. of TN_i refers to the number of segments that are test negatives in cluster i, as determined by the test being evaluated (MDCT angiography).

No. of Segments (N_i) without Event refers to the number of segments that are true negatives in cluster i, as determined by the "gold standard" (invasive coronary angiography).

$\hat{Sp}_i$ = No. of TN_i / No. of segments without event (N_i).

$N_i / \bar{N}$ = the cluster size for patient$_i$ / mean cluster size.

$$\hat{Sp} = \frac{\sum_{i=1}^{I} N_i \hat{Sp}_i}{\sum_{i=1}^{I} N_i} \qquad [1.17]$$

$$\hat{Var}(\hat{Sp}) = \frac{1}{I(I-1)} \sum_{i=1}^{I} \left[\left(N_i / \bar{N} \right)^2 (\hat{Sp}_i - \hat{Sp})^2 \right] \qquad [1.18]$$

where $\bar{N} = \Sigma \frac{N_i}{I}$ is the mean cluster size.

The usual approach to constructing a confidence interval for a measure of diagnostic accuracy assumes a large sample size (i.e., follows a normal distribution asymptotically). This confidence interval is referred to as an asymptotic interval. It has the following form [Equation 1.12]:

$$\hat{\Theta} - z_{1-\alpha/2} \sqrt{\hat{Var}\left(\hat{\Theta}\right)}, \ \hat{\Theta} + z_{1-\alpha/2} \sqrt{\hat{Var}\left(\hat{\Theta}\right)}$$

where $\hat{\Theta}$ is the estimate of the accuracy measure, Θ; $z_{1-\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution, and $100(1-\alpha)\%$ is the confidence level.

V. Sample Size Requirements for Estimation in a Single Sample: Binary-Scale Data

A general formula for sample size estimation for constructing a 2-sided Confidence Interval (CI) for a single diagnostic test measure of accuracy is:

$$m = \frac{z_{1-\alpha/2}^2 \left[V\left(\hat{\nu}\right) \right]}{L^2} \qquad [1.19]$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution, α is the confidence level, $V\left(\hat{\nu}\right)$ is the variance function of $\hat{\nu}$, ν is the conjectured measure of accuracy (Se or Sp) and L is the desired width of one-half of the CI. Often, one wishes to construct a 95% CI, in which case α = 0.05 and $z_{1-\alpha/2}$ =1.96.

With respect to $\hat{Se}$, m will be the number of patients with the condition of interest.

With respect to $\hat{Sp}$, m will be the number of patients without the condition of interest.

One can use

$$\hat{V}\left(\hat{\nu}\right) = \nu\left(1-\nu\right), \qquad [1.20]$$

to calculate $V\left(\hat{\nu}\right)$.

VI. Sample Size Requirements for Estimation in a Single Sample: Clustered Binary-Scale Data

In a talk presented by J M Bland to the RSS Medical Section and the RSS Liverpool Local Group, 12 NOV 2003, Bland noted that "the magnitude of the effect of clustering is measured by the design effect, *Deff*, given by the following: *Deff* = 1 + (n - 1)(ICC) where n is the number of observations in a cluster and ICC is the intra-cluster correlation coefficient. The ICC is the correlation between pairs of subjects chosen at random from the same cluster. It is usually quite small, 0.04 is a typical figure. This was the median ICC reported in the review by Eldridge *et al*. (2004). If n=1, cluster size one, in other words, no clustering, then *Deff*=1, otherwise *Deff* will exceed 1 (this assumes a positive correlation between pairs of subjects).

We can use this in two ways. In design, if we estimate the required sample size ignoring clustering, we must multiply it by the design effect to get the sample size required for the clustered sample. Alternatively, we can say that if the sample size is estimated ignoring the clustering, the clustered sample has the same power as for a simple sample of size equal to what we get if we divide our sample size by the design effect.

In analysis, if we analyse the data as if there were no clusters, the variances of the estimates must be multiplied by *Deff*, hence the standard error must be multiplied by the square root of Deff.

From this formula, we can see that clustering may have a large effect if the ICC is large or if the cluster size is large. Only one of these conditions need be met. For example, if the ICC is 0.001, a very small correlation, and the cluster size is 500, the design effect will be 1 + (500-1)x0.001 = 1.5 and we would need to increase the sample size by 50% to achieve the same power as an unclustered trial.

In addition, we need to estimate variances both within and between clusters. If the number of clusters is small, the between clusters variance will have few degrees of freedom and we will be using the t distribution in inference rather than the Normal. This too will cost in terms of power."

ICC (ρ) is the correlation between pairs of patients chosen at random from the same cluster.

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

where σ_B^2 is the between cluster variability

where σ_W^2 is the within cluster variability

the size of the ICC is generally larger for smaller clusters

small cluster $\sim$ 0 to 0.3 (large ICC)

medium cluster $\sim$ 0 to 0.05 (medium ICC)

large cluster $\sim$ 0 to 0.001 (small ICC)

Illustration of the use of Deff to Determine the Statistical Precision of Estimates of Sensitivity and Specificity

Patient-Level Binary-Scale: Event= Same coronary artery with $\geq$50% stenosis with an event rate=35% (Only one coronary artery with an event is required for the analysis) and Non-event= Else.

Table 3. Event Rate for Patient-Level Analysis (35% of 350)

Events	Non-Events
123	227

Coronary Artery-Level Binary-Scale: Event= Same coronary artery with $\geq 50\%$ stenosis with an event rate=18% (All coronary arteries with an event are considered in the analysis) and Non-event= Else.

The 1050 sample size for coronary arteries is based on 350 patients x 3 coronary arteries=1050.

Table 4. Event Rate for Coronary Artery-Level Analysis (18% of 1050)

Events	Non-Events
189	861

Table 5. Precision Expected for Sensitivity 95%Confidence Interval when Expected Patient-Level (Single Sample: Binary-Scale Data) and Expected Coronary Artery-Level (Single Sample: Clustered Binary-Scale Data) Sample Sizes are Considered.

Sensitivity				
	Patient-Level 123 Patients with an Event (35% of 350)		C.A.-Level 189 C.A.s with an Event (18% of 1050)	
$\hat{Se}$	+/- Precision	+/-Precision Large ICC	+/-Precision Medium ICC	+/- Precision Small ICC
0.85	0.063	0.065	0.054	0.051
0.87	0.059	0.061	0.050	0.048
0.90	0.053	0.054	0.045	0.043
0.91	0.051	0.052	0.043	0.041
0.92	0.048	0.049	0.041	0.039
0.93	0.045	0.046	0.038	0.036
0.94	0.042	0.043	0.036	0.034
0.95	0.039	0.039	0.033	0.031
0.96	0.035	0.035	0.029	0.028

In all likelihood, the ICC associated with nesting of coronary arteries in a patient will be large. Therefore, it may be best to determine expected +/- precision for the coronary artery-level analysis based on the results of the +/- Precision Large ICC column.

The standard error of a non-clustered coronary artery-level design was multiplied by the square root of Deff to get the proper expected +/- precision for the coronary artery-level clustered design. For the coronary artery-level columns above, the +/- precision was computed by multiplying the standard error by the square root of Deff as specified by Bland JM (2003).

Table 6. Precision Expected for Specificity 95% Confidence Interval when Expected Patient-Level (Single Sample: Binary-Scale Data) and Expected Coronary Artery-Level (Single Sample: Clustered Binary-Scale Data) Sample Sizes are Considered.

Specificity				
	Patient-Level 227 Patients without an Event (65% of 350)		C.A.-Level 861 C.A.s without an Event (82% of 1050)	
$\hat{Sp}$	+/- Precision	+/-Precision Large ICC	+/- Precision Medium ICC	+/-Precision Small ICC
0.85	0.046	0.030	0.025	0.024
0.87	0.044	0.029	0.024	0.023
0.90	0.039	0.025	0.021	0.020
0.91	0.037	0.024	0.020	0.019
0.92	0.035	0.023	0.019	0.018
0.93	0.033	0.022	0.018	0.017
0.94	0.031	0.020	0.017	0.016
0.95	0.028	0.019	0.016	0.015
0.96	0.026	0.016	0.014	0.013

Illustration of Analyzing Clustered Binary-Scaled Data

Using the data layout outlined in Table 2a and the equations [1.15], [1.16], we illustrate the effect of clustering on binary-scaled data (Refer to Table 7 and 8). We will estimate the 95% CI of sensitivity at the coronary artery segment level.

Table 7. Data Layout of Example
(Based on an adaptation of data and example presented by Zhou et al, 2002)

Patient (cluster)	No. of TP (MDCT)	No. of segments (N_i) with Event (invas. c. angio.)	$\hat{Se}_i$	$N_i / \bar{N}$	$(N_i / \bar{N})^2 (\hat{Se}_i - \hat{Se})^2$
1	0	3	0.000	1.235	0.9391
2	2	3	0.667	1.235	0.0211
3	3	3	1.000	1.235	0.0710
4	1	1	1.000	0.412	0.0079
5	2	3	0.667	1.235	0.0211
6	4	4	1.000	1.647	0.1263
7	3	3	1.000	1.235	0.0710
8	2	2	1.000	0.824	0.0316
9	2	2	1.000	0.824	0.0316
10	1	1	1.000	0.412	0.0079
11	2	3	0.667	1.235	0.0211
12	2	2	1.000	0.824	0.0316
13	3	3	1.000	1.235	0.0710
14	2	2	1.000	0.824	0.0316
15	0	2	0.000	0.824	0.4174
16	2	3	0.667	1.235	0.0211
17	2	3	0.667	1.235	0.0211
18	2	3	0.667	1.235	0.0211
19	2	2	1.000	0.824	0.0316
20	1	1	1.000	0.412	0.0079
21	2	2	1.000	0.824	0.0316
Sum	40	51			2.0357

Each patient is considered a cluster (i=1,2,..,I clusters); within the cluster are a number of segments. We have 21 patients (clusters) with an event based on invasive coronary angiography; we have 51 coronary artery segments (DUOS) with an event based on invasive coronary angiography.

No. of TP_i refers to the number of segments that are test positives in cluster i, as determined by the test being evaluated (MDCT angiography).

No. of Segments (N_i) with Event refers to the number of segments that are true positives in cluster i, as determined by the "gold standard" (invasive coronary angiography).

We shall say there is concordance between the test being evaluated and the "gold standard" if the same (identical) coronary artery segment has the event of interest. That is, if coronary artery segment 4 is determined to be positive by the test being

evaluated as well as by the “gold standard”, we have concordance. There must be a one-to-one correspondence between the coronary artery segment determined to be positive by the test being evaluated and the “gold standard” in order for the diagnostic accuracy assessment to be meaningful!

$$\hat{Se} = \frac{\sum_{i=1}^{I} N_i \hat{Se}_i}{\sum_{i=1}^{I} N_i}$$

= No. of TP / No. of segments with event (N_i)

=40/51=0.7843.

$N_i / \bar{N}$ = the cluster size for patient$_i$ / 2.43.

$$\hat{Var}(\hat{Se}) = \frac{1}{I(I-1)} \sum_{i=1}^{I} \left[\left(N_i / \bar{N} \right)^2 (\hat{Se}_i - \hat{Se})^2 \right]$$

=2.0357/[(21)(20)]

=0.0048

where $\bar{N} = \Sigma \frac{N_i}{I}$ is the mean cluster size.

If the clustering effect was ignored, $\hat{Var}(\hat{Se})$

=0.7843(0.2157)/51

=0.0033 (based on Equation 1.10).

Table 8. 95% Confidence Limits (using Equation 1.12)

Clustering Effect not Considered	Clustering Effect Considered
0.7843 ± 0.1126 (0.6717, 0.8969)	0.7843 ± 0.1358 (0.6485, 0.9201)

One should note that if the clustering effect is not taken into account in this example, the calculated variance (thus the calculated confidence limits) is inappropriately small because the calculation ignores the positive correlation among the coronary artery segments nested within the patient. Even if the amount of correlation among the coronary artery segments nested within the patient is small, the variance (confidence limits) will be biased if the clustering effect is not taken into account.

VIII. Additional Methods to Estimate Diagnostic Accuracy

Comparison of Ordinal MDCT Angiography Data to Ordinal “Gold Standard” Invasive Coronary Angiography Data

Diagnostic accuracy for ordinal MDCT angiography and invasive coronary angiography data (type I, less than 25% stenosis; type II, 25% - 50% stenosis; type III, 51% - 74% stenosis; type IV, 75% - 89% stenosis; type V, 90% - 99% stenosis; and type VI, 100% stenosis) can be assessed by the use of a weighted kappa analysis (Liebetrau, 1983). The standard error used to compute the 95% confidence interval for the weighted kappa when clustered data are considered will be adjusted for clustering by the use of the Deff as specified by Bland JM (2003).

Landis and Koch (1977) have indicated a kappa greater than 0.75 may be taken to represent excellent agreement beyond chance. A kappa below 0.40 may be taken to represent poor agreement beyond chance and a kappa between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance.

Comparison of Continuous MDCT Angiography Data to Continuous "Gold Standard" Invasive Coronary Angiography Data

Diagnostic accuracy for continuous MDCT angiography and invasive coronary angiography data (any continuous stenosis value between 0% and 100%, inclusive) can be assessed by the using the methods outlined by Bland (1995). Using a graphic approach, we will generate a scatter diagram in which a "line of equality" is superimposed. If there is perfect agreement between both the invasive coronary angiography and MDCT angiography, the plotted point will sit directly on the "line of equality". The degree that the plotted point sits above or below the "line of equality" indicates the limit of agreement. Also, A 95% confidence interval of the mean difference between MDCT angiography and invasive coronary angiography percent stenosis will be computed (Bland and Altman, 1986). The standard error used to compute the 95% confidence interval for the mean difference when clustered data are considered will be adjusted for clustering by the use of the Deff as specified by Bland JM (2003).

IX. Multivariable Modeling

Sensitivity and specificity of MDCT angiography may vary by gender, race/ethnicity, age and so forth. Multivariable modeling will be used to identify factors associated with differences in the sensitivity and specificity of MDCT angiography.

To identify factors associated with sensitivity, the analysis will be limited to the subset of cases that are classified as events by invasive coronary angiography ($Y_{AN} = 1$). Using this subset of cases, logistic regression can be used to model the probability that the case is also classified as an event by CT ($pr[Y_{CT} = 1]$, where $Y_{CT} = 1$ means that a case is an event by MDCT angiography). Modeling $pr(Y_{CT} = 1)$ among cases where $Y_{AN} = 1$ is a logistic regression approach to sensitivity analysis. Specifically, the logistic model is $\log(p/1-p) = X*Beta$, where $p = pr(Y_{CT} = 1 \mid X)$, X is the covariate design matrix, and Beta is a vector of parameters to be estimated. The covariates can be either continuous (e.g., age) or categorical (e.g., gender). A similar approach can be used to identify factors associated with specificity, except that the analysis is limited to the subset of cases not classified as events by invasive coronary angiography ($Y_{AN} = 0$) and logistic regression is used to model the probability that a case is classified as a non-event by MDCT angiography ($pr[Y_{CT} = 0]$). The discussion here focuses on sensitivity, but these methods can also be used in specificity analyses.

In a logistic regression model with an intercept (beta0) but no other predictors, the estimate [exp(beta0) / (1+exp(beta0))] is identical to the sensitivity estimate computed by the methods described above. This is a "population average" estimate of sensitivity.

Adding covariates, one can determine whether sensitivity varies among different subpopulations. The Wald chi-square will be used for hypothesis tests about subpopulation differences (likelihood ratio tests are also possible using mixed models). Subpopulation differences will be described using two methods. First, odds ratios will be computed, indicating the increase in sensitivity associated with a 1-unit increase in a given covariate. Second, conditional marginals produced by the logistic regression procedure in SUDAAN will be used to estimate sensitivity for specific subpopulation, e.g., different countries. The conditional marginal is an estimator of the expected response conditional on having a particular set of covariate values.

Analyses at the coronary artery and coronary artery segment levels must take clustering into account (coronary arteries are nested within patients, and coronary artery segments are nested within coronary arteries). Methods for logistic regression with clustered data have been developed for analysis of multistage survey data (e.g., the SUDAAN logistic regression procedure or SAS PROC SURVEYLOGISTIC). These methods produce the same regression coefficients (betas) as ordinary logistic regression, but the standard errors are adjusted to take into account variation among the clusters. If there is substantial variation among clusters, the inter-cluster variation will result in increased variance estimates. The regression coefficients are interpreted as population average parameters, i.e., averaging over all clusters. We will assume that all cases have a weight of one and all are drawn from the same population stratum. Patient and vessel will be represented as clusters in the analysis.

X. Imputation: Handling of Missing Data

All analyses will be conducted on non-imputed and imputed data sets. The results from both types of data sets will be reported. Imputation relates to a family of techniques for replacing missing values in a data set by 'synthetic values' obtained from some

kind of model. Such a model describes a relationship between the variable having missing values and other variables for which the values are available.

Imputation of binary-scaled outcome data (non-clustered and clustered) will use a worst score method of imputation similar to that proposed by McCarthy et al, 2004. This method will allow for an unbiased assessment of sensitivity and specificity when MDCT angiography binary-scaled data is missing. In essence, the imputed value for the missing MDCT angiography binary-scaled value will be the opposite of what was reported for the "gold standard" invasive coronary angiography. For example, if a patient, vessel or segment has a missing MDCT angiography binary-scaled value, and the corresponding gold standard binary-scaled value was positive, the imputed value for the MDCT angiography would be negative, and vice versa.

For continuous missing outcome data, Hot-deck imputation will be used *(*Kalton and Kasprzyk, 1986). This is a special implementation of random imputation within groups. For each group, a donor record is maintained. The records in the file are processed sequentially. If the field of the variable to be imputed contains a 'real' value, the value is copied to the donor record. If the value in the field is missing, the value from the donor record is copied to the field.

XI. Statistical Software and FDA Compliance

Statistical computing will be done using the following statistical software packages: SAS (Version 9.0) ; SUDAAN (Release 9); and StatXact (Version 6). Statistical computing programs for the analyses of diagnostic accuracy (non-clustered, clustered, and multivariable) will be reviewed and validated in accordance to the SAS Testing element of the FDA Regulation 21 CFR Part 11. The statistical design and all statistical analyses will comply with the International Conference on Harmonization (ICH) Guidance for Industry: E9 Statistical Principles for Clinical Trials.

XII. References:

Agresti A and Coull B (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *Am. Stat.* **52:** 119-126.

Beam CA (1998). Analysis of clustered data in receiver operating characteristic studies, *Stat. Methods Med. Res.* **7:** 324-336.

Bland JM (2003). *Cluster Randomized Trials in the Medical Literature*. Talk presented to the RSS Medical Section and the RSS Liverpool Local Group, 12 NOV 2003. http://www-users.york.ac.uk/~mb55/talks/clusml.htm

Bland JM (1995). An introduction to medical statistics. 2nd Edition. Oxford: Oxford Medical Publications, 269-273.

Bland JM and Altman DG (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, Feb 8; 1 (8476): 307-10.

Blyth CR and Still HA (1983). Binomial confidence intervals, *Journal of the American Statistical Association*, **78:** 108-116.

Bogaerts K. (2004). An introduction to the analysis of cluster randomized trials. Seminars in Epidemiology, Catholic University, Leuven, Belgium (25 MAR 2004) http://www.med.kuleuven.ac.be/biostat/courses/Bogaerts25MAart2004.pdf

Centor, R.M. and Schwartz, J.S., An evaluation of methods for estimating the area under the Receiver Operating Characteristic (ROC) curve, Medical Decision Making 5 (1985) 149-156.

Centor, R.M., Signal detectability: The use of ROC curves and their analyses, Medical Decision Making 11 (1991) 102-106.

Cochran W. (1977). *Sampling Techniques*, 3rd ed., John Wiley and Sons, New York.

DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L., Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach, Biometrics 44 (1988) 837-845.

Efron, B. and Tibshirani, R.J., An Introduction to the Bootstrap. New York: Chapman and Hall, 1993.

Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC: Lessons for cluster randomised trials in the 21st century: a systematic review of trials in primary care. *Clin Trials* 2004, **1:**80-90.

FDA Regulation 21 CFR Part 11.
http://www.fda.gov/cder/guidance/5667fnl.htm

Goddard M.J. and Hinberg, I., Receiver operating (ROC) curves and non-normal data: an empirical study, Statistics in Medicine 9 (1990) 325-337.

Ghosh BK (1979). A comparison of some approximate confidence intervals for the binomial parameter, *Journal of the American Statistical Association*, **74:** 894-900.

Hanley, J.A. and McNeil, B.J., The meaning and use of the area under a receiver-operating characteristic (ROC) curve. Radiology 143(1982): 29-36.

Hanley, J.A. and McNeil, B.J., A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148(1983): 839-843.

Hanley, J.A., The robustness of the "binormal" assumptions used in fitting ROC curves, Medical Decision Making 8 (1988) 197-203.

International Conference on Harmonization (ICH) Guidance for Industry: E9 Statistical Principles for Clinical Trials. http://www.fda.gov/cder/guidance/ICH_E9-fnl.PDF

Kalton, G. and D. Kasprzyk (1986): The Treatment of Missing Survey Data. Survey Methodology 12, pp.1-16.

Kerry SM and Bland JM (1998). The intraclass correlation coefficient in cluster randomization. BMJ, 316: 1455-1460 (9 May) http://bmj.bmjjournals.com/cgi/content/full/316/7142/1455

Lachin JM (1999). Worst-rank score analysis with informatively missing observations in clinical trials. Controlled Clinical Trials 20:408-422.

Liebetrau AM (1983). Measures of association. Sage Publications.

Landis JR and Koch GG (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Little, R.J.A. and D.B. Rubin (1987): Statistical Analysis with Missing Data. Wiley, New York.

McCarthy WF, SA Forman, BB Barton, SP Schulman, and ML Terrin (2004). A method for an unbiased assessment of treatment effect in the vascular interaction with age in myocardial infarction (VINTAGE MI) clinical trial. Clinical Trials, Volume 1; Supplement 1, 242, P 24.

Obuchowski NA and McClish DK (1997). Sample size determination for diagnostic accuracy studies involving binormal roc curve indices. *Statistics in Medicine*, Vol. 16, pp 1529-1542.

Obuchowski, N., Confidence Intervals for the Receiver Operating Characteristic Area in Studies with Small Samples. Academic Radiology 5 (1998) 561-571.

Rao NK and Scott AJ (1992). A simple method for the analysis of clustered binary data, *Biometrics* **48:** 577-585.

Rockette, H.E., Obuchowski, N.A. and Gur, D., Nonparametric estimation of degenerate ROC data sets used for comparison of imaging systems, Invest Radiology 25 (1990) 835-7.

Rubin, D.B. (1979): Illustrating the use of multiple imputations to handle non-response in sample surveys. Bulletin of the International Statistical Institute, Book 2, pp. 517-532.

SAS, Version 9. SAS Institute, Cary, NC.

StatXact, Version 6. Cytel Statistical Software, Cambridge, MA.

SUDAAN, Release 9. RTI International, RTP, NC.

Vida, S., A computer program for non-parametric receiver operating characteristic analysis. Computer Methods and Programs in Biomedicine 40 (1993) 95-101.

Wang JC, Normand S-L T, Mauri L and Kuntz RE (2004). Coronary artery spatial distribution of acute myocardial infarction occlusions, *Circulation* **110:** 278-284.

Zhou X-H, Obuchowski NA, McClish DK (2002). *Statistical Methods in Diagnostic Medicine*. Wiley, New York.