Year 2007 *Paper* 23

The Analysis of Pixel Intensity (Myocardial Signal Density) Data: The Quantification of Myocardial Perfusion by Imaging Methods.

William F. McCarthy[*] Douglas R. Thompson[†]

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

[†] Maryland Medical Research Institute, dthompson@mmri.org

The Analysis of Pixel Intensity (Myocardial Signal Density) Data: The Quantification of Myocardial Perfusion by Imaging Methods.

William F. McCarthy and Douglas R. Thompson

Abstract

This paper described a number of important issues in the analysis of pixel intensity data, as well as approaches for dealing with these. We particularly emphasized the issue of clustering, which may be ubiquitous in studies of pixel intensity data. Clustering can take many forms, e.g., measurements of different sections of a heart or repeated measurements of the same research participant. Clustering typically has the effect of increasing variance estimates. When one fails to account for clustering, variance estimates may be unrealistically small, resulting in spurious significance. We illustrated several possible approaches to account for clustering, including adjusting standard errors for design effects and modeling the covariance structure within clusters using mixed models. These methods offer great flexibility for dealing with a wide variety of research designs and include the capability for adjusting for covariates and different case weights. Similar methods can be used to account for clustering in both superiority and equivalence analyses. In situations where clustering affects the true cluster mean, μ, but not the difference between measures of the mean, it is possible that clustering will have a much greater impact on superiority analyses than on equivalence analyses.

A. Introduction

Superiority: the goal of a superiority analysis is to determine if one image region (new imaging technology) is significantly different (superior) to another image region (an established imaging technology or gold standard). For example, the Wilcoxon Signed Rank Test can be performed to determine if there were significant differences in signal density between remote and infarct regions at one time point. The Wilcoxon Signed Rank Test is a non-parametric test used to test the median difference in paired data (Wilcoxon, 1945 and Conover, 1980). One could also use the parametric paired t-test (Armitage and Berry, 1994; Altman, 1991). One could define statistical significance as a 2-sided p-value<0.05.

Another approach to superiority analysis is modeling. For example, if the goal was to determine whether blood flow was less in one region of the heart ("infarct") than in a comparison region ("remote"), one could use the model:

$$\text{Perfusion measure}_i = \gamma_0 + \gamma_1 * \text{Region}_i + \varepsilon_i$$

Where γ_0 is an estimated intercept and γ_1 is the estimated region difference (0 = remote, 1 = infarct). Superiority is supported if the null hypothesis (H_0: $\gamma_1 = 0$) is rejected. Advantages of modeling include the ability to adjust for covariates and the availability of many approaches to take clustered (non-independent) data into account.

Method Comparison (Equivalence): In contrast to a superiority analysis, a method comparison analysis is used to determine if a new imaging technology is statistically interchangeable (give clinically equivalent results) to an established imaging technology or gold standard. It is important to know whether the imaging technologies give measurements that are in some sense comparable. For example, one may wish to see whether a new imaging technology produces measurements that are clinically in close agreement with those from an alternative method. As noted by Bland and Altman (1999), "Some lack of agreement between different methods of measurement is inevitable. What matters is the amount by which methods disagree. We want to know by how much the new method is likely to differ from the old, so that if this is not enough to cause problems in clinical interpretation we can replace the old method by the new, or even use the two interchangeably."

Agreement between a new imaging technology, denoted as A, and an established imaging technology or gold standard, denoted as B, can be assessed by a Bland Altman plot as well as a 95% confidence interval of the mean difference between the new imaging technology and the established imaging technology or gold standard can be computed as well (Bland and Altman, 1986). One can also use the regression methodologies as well: Deming regression, Passing-Bablok regression, Mountain plots and a regression method outlined by Hawkins (2002) which gets results comparable to Bland and Altman (1986). Statistical analyses for method comparison were performed using MedCalc for Windows, version 8.2.1.0 (MedCalc Software, Mariakerke, Belgium); SAS 9.1 (PROC MIXED for the cluster-specific models); and SUDAAN 9.01.

Altman and Bland (1983;1986;1999) proposed an approach for analyzing measurement comparison data. In order to assess the agreement of the two technologies graphically, the difference between the two measurements of each of the two methods are plotted against the mean of the two values of the methods for each subject. Limits of agreement, defined as twice the standard deviation of the difference between the technologies, are calculated and plotted in the figure. If we suppose that these differences follow a normal distribution, 95% of the differences will lie between the limits of agreement. If the differences are Normally distributed (Gaussian), 95% of differences will lie between the limits of agreement (or, more precisely, between mean difference – 2*standard deviation [$\overline{d}$ -2s] and mean difference+2*standard deviation [$\overline{d}$ +2s]). Such differences are likely to follow a Normal distribution because we have removed a lot of the variation between subjects and are left with the measurement error. The measurements themselves do not have to follow a Normal distribution, and often they will not. We can check the distribution of the differences by drawing a histogram. If this is skewed or has very long tails the assumption of Normality may not be valid. We can also perform a Test of Normality. From this type of plot, it is easy to see if there is any tendency for the variation to change with the magnitude of the measurements. If the differences are symmetrical around zero, then there is no systematic bias. If the differences fall within the limits of agreement and the limits of agreement are considered to be clinically acceptable in terms of agreement, then one can say the two methods are in some sense comparable. If there is no relationship between the differences and the averages, the agreement between the two methods may be summarized using the means and standard deviations of the method measurements.

Deming regression (Combleet & Gochman, 1979) is a method of linear regression that finds a line of best fit for a set of related data. It differs from simple linear regression in that it accounts for error in both the *x* and the *y*-axis. The line of regression (or line of best fit) must begin where your x and y axis meet (zero). If both sets of data contained error Deming regression would be more appropriate than linear regression.

Passing & Bablok (1983) have described a linear regression procedure with no special assumptions regarding the distribution of the samples and the measurement errors. The result does not depend on the assignment of the methods (or instruments) to X and Y. The slope

B and intercept A are calculated with their 95% confidence interval. These confidence intervals are used to determine whether there is only a chance difference between B and 1 and between A and 0.

A M**ountain Plot** (folded empirical cumulative distribution plot) is created by computing a percentile for each ranked difference between a new method and a reference method. To get a folded plot, the following transformation is performed for all percentiles above 50: percentile = 100 - percentile. These percentiles are then plotted against the differences between the two methods (Krouwer & Monti, 1995). The mountain plot is a useful complementary plot to the Bland & Altman plot. In particular, the mountain plot offers the following advantages: It is easier to find the central 95% of the data, even when the data are not Normally distributed and Different distributions can be compared more easily.

Other regression methods can also be used for method comparison (Hawkins, 2002). Assume that two methods are used with a sample of n subjects ($i = 1,2,\ldots,n$). X_i is the measure using the first method in patient i and Y_i is the measure using the alternative method in patient i. Matched pairs are assumed such that each patient i has measures using both the first methodology (X) and an alternative methodology (Y) (missing data is possible). To conduct a method comparison analysis using regression, one computes the sum and difference of the two methods ($\text{Sum}_i = X_i + Y_i$, $\text{Difference}_i = Y_i - X_i$), then the difference is regressed on the sum, i.e., one estimates the model $\text{Diff}_i = \alpha + \text{Sum}_i{*}\beta + \varepsilon_i$. Covariates can also be included and cases can have different weights. Equivalence is supported if one *fails* to reject the null hypotheses, H_0: $\beta = 0$ and $\alpha = 0$.

In addition to the advantages for superiority analysis (i.e., adjustment for covariates and availability of a range of methods to account for clustered data), Hawkins argued that regression models are especially useful for method comparison because regression diagnostics can be used to identify and correct for violations of assumptions of method comparison analysis.

B. Critical Issue for Analysis

Any sampling scheme (or hierarchical structure) different from simple random sampling will require an adjustment to be made to the variance of the estimate or the variance used in the test statistic. If the adjustment is not made, the variance of an estimate or the p-value of the hypothesis test will be incorrect.

A multi-stage sampling scheme best describes how the data was generated. The first stage is the dog's heart. The second stage consists of two regions, infarct and remote. The third stage consists of five "rings" parallel to and encircling the long axis. The fourth and final stage is made up of the CT slices (images), generated when the heart is divided into five 1 cm slices perpendicular to the short axis, with the number of slices varying from dog to dog. This type of data collection mechanism is not simple random sampling and therefore an adjustment needs to be made during analysis when estimation or hypothesis testing is performed.

A variance estimate that assumes simple random sampling of non-correlated data will not be valid for calculating the variance of an estimate (or the variance used in a test statistic) generated by a multi-staged (clustered) sampling scheme. The variance will be underestimated if the correlation between the measurements is positive (i.e., the estimated variance will be smaller than the true variance) or will be overestimated if the correlation between the measurements is negative (i.e., the estimated variance will be larger than the true variance). A positive correlation indicates that as one measurement increases the other increases as well. A negative correlation indicates that as one measurement increases the other decreases. To correct for this underestimated (overestimated) variance problem (bias), an adjustment to the variance needs to be made.

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: $\text{Deff} = 1 + (n - 1)(\text{ICC})$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 (the square root of the Deff is sometimes called the "design factor," abbreviated DEFT). If this Deff adjustment is not made, the variance of an estimate or the p-value of a hypothesis testing will be incorrect.

To see how this critical issue for analysis impacts the results, we will look at two data structures for analysis:

1. Clustering not considered at any level, all slice data used across all dogs (n=163 total slices, treated as independent observations)
2. Clustering considered, slices nested within dogs
   (dog1,slices=24; dog2, slices=28; dog3, slices=20; dog4, slices=21; dog5, slices=19; dog6, slices=23; dog7, slices=28)

C. Data

Analyses were based on data that summarized pixel intensity measures from the heart of seven dogs. The experimental protocol involved producing an infarct in the heart of each dog. Perfusion was measured in both infarct and remote regions of the heart of each dog after production of the infarct. Details of the method are provided by George et al (2006).

The data structure consisted of 4 nested levels: dog, region, ring and slice. Two regions were identified in each dog's heart, "infarct" and "remote". Each region encompassed either 4 or 5 rings. Within each ring, infarct and remote measures were collected from between 1 and 6 slices (mean = 3 slices). In The structure is illustrated for a single dog, showing only 2 rings, in Table 1.

This analysis used measures of perfusion by multi-detector computed tomography ("MDCT"), measured by signal density (SD) in Hounsfeld Units as determined by software. Because SD in the entire left ventricular (LV) blood pool varied across dogs, myocardial signal densities were normalized by signal density in the entire LV cavity. Specifically, the analyses focused on: myocardial SD ratio = myocardial signal density / signal density in the LV blood pool. An additional set of MDCT values were simulated ("SIM-MDCT") to illustrate equivalence analyses. The intent was to illustrate a case of "substantial equivalence," therefore the values of SIM-MDCT were generated from a normal distribution with a mean and standard deviation equal to the mean and standard deviation of the actual MDCT values across all slices for a given dog, region and ring.

Table 1. Illustration of the raw data structure.

Dog	Region	Ring	Slice	MDCT (actual)	SIM-MDCT (simulated)
1	Infarct	1	1	0.24035	0.23250
1	Infarct	1	2	0.23342	0.21010
1	Infarct	1	3	0.20382	0.25605
1	Infarct	1	4	0.24127	0.21320
1	Infarct	1	5	0.22389	0.23377
1	Infarct	2	1	0.21941	0.22810
1	Infarct	2	2	0.23024	0.22347
1	Infarct	2	3	0.21900	0.22299
1	Infarct	2	4	0.22293	0.22443
1	Infarct	2	5	0.23118	0.22565
1	Remote	1	1	0.40672	0.43051
1	Remote	2	1	0.37978	0.39532
1	Remote	2	2	0.41657	0.42641

D. Results

Descriptive statistics (mean, standard deviation, median, and sample size = total number of slices) are shown in Table 2.

Table 2. Mean (SD), median, and sample size for MDCT and SIM-MDCT, by dog

Dog	MDCT		SIM-MDCT	
	Infarct	Remote	Infarct	Remote
1	0.22 (0.01), 0.22, n=15	0.39 (0.02), 0.39, n=9	0.22 (0.02), 0.22, n=15	0.39 (0.03), 0.38, n=9
2	0.25 (0.02), 0.26, n=17	0.42 (0.02), 0.42, n=11	0.25 (0.02), 0.25, n=17	0.42 (0.02), 0.42, n=11
3	0.31 (0.01), 0.31, n=12	0.38 (0.01), 0.37, n=8	0.30 (0.02), 0.31, n=12	0.37 (0.01), 0.37, n=8
4	0.31 (0.01), 0.31, n=13	0.37 (0.01), 0.37, n=8	0.31 (0.01), 0.31, n=13	0.37 (0.01), 0.37, n=8
5	0.15 (0.01), 0.15, n=12	0.20 (0.01), 0.19, n=7	0.15 (0.01), 0.14, n=12	0.20 (0.02), 0.19, n=7
6	0.12 (0.02), 0.11, n=18	0.15 (0.02), 0.14, n=5	0.12 (0.02), 0.11, n=18	0.14 (0.02), 0.13, n=5
7	0.16 (0.01), 0.16, n=18	0.21 (0.01), 0.21, n=10	0.16 (0.01), 0.16, n=18	0.21 (0.01), 0.21, n=10

Superiority

An initial analysis of the strength of clustering revealed an intra-class correlation (ICC) of 0.81 and a design effect (DEFF) of 18.97. This indicates a strong, positive clustering effect. The DEFF was computed using SUDAAN 9.01 (this was estimated using the Taylor series method, other methods such as jackknife could yield different estimates).

Using data illustrated in Table 1, the Wilcoxon Signed Rank Test was performed. For example, let's say we are only interested in whether there is a significant difference in median MDCT between the infarct and remote regions when using MDCT (the simulated values, SIM-MDCT, were used only in the equivalence analyses, not in the superiority analyses). Then Table 3 shows the results of such an analysis with the two data structures for analysis.

Table 3. Descriptive statistics and test results for comparison of median MDCT across regions (infarct vs. remote), based on different approaches for dealing with multiple measurements within dogs

Median MDCT by region and statistical test results	**Ignore clustering (treat observations as independent)**	**Design effects adjustment for clustering**
Descriptive statistics		
Median MDCT, infarct	0.3688	0.3688
Median MDCT, remote	0.2140	0.2140
Test results		
Wilcoxon signed rank test statistic	-6.624	-1.521
Asymptotic p-value	< 0.0001	0.12830
Exact p-value	< 0.0001	*

* StatXact was used for computation of the exact Wilcoxon signed rank test; the exact Wilcoxon signed rank test adjusted for clustering is not available.

Let's define statistical significance as a p-value<0.05. These results show that there is a significant difference in median MDCT between the infarct and remote regions (ignoring DEFF), but when the estimates were adjusted for clustering there was a non-significant result. Note that there is essentially no difference in the asymptotic and exact p-values (for examples of how asymptotic and exact confidence intervals and p-values can differ, see McCarthy and Gable, 1999). The use of the Wilcoxon Signed Rank Test is appropriate for such an analysis that focuses on the comparison of two correlated regions. If more than one time point is to be considered and/or more than two correlated regions are to be compared, the Wilcoxon Signed Rank Test would not be the best test to use.

In the modeling approach, the basic superiority model was: $\text{MDCT}_{ijkl} = \gamma_0 + \gamma_1 * \text{Region}_{ijkl} + \varepsilon_{ijkl}$

Random effects were estimated in some models (see descriptions in Table 4). The null hypothesis is that MDCT does not differ between regions (infarct vs. remote), that is, H_0: $\gamma_1 = 0$. If the null hypothesis is rejected, this is evidence in support of superiority.

Two broad modeling approaches to account for clustering were considered: *cluster-specific* approaches (cluster-specific models are estimated and pooled and the correlation structure is modeled; Bowman & Waller, 2004) and *population average* approaches (the correlation is treated as a nuisance -- the model estimates are adjusted for it, but the degree and structure of correlation and the variance among clusters is not estimated; Lavange, Koch & Schwartz, 2001). Several variants of each approach to account for clustered data were estimated (denoted "CS" for cluster-specific or "PA" for population average).

Table 4. Description of models used in Superiority and Equivalence analyses, with model fit indices (AIC, BIC)

Model	**Description**	**Superiority**		**Equivalence**	
		AIC †	**BIC**	**AIC**	**BIC**
Models failing to account for clustering					
Ignore clustering	Model raw (detailed) data as independent observations (n = 163)	-343	-333	-865	-856
Cluster-specific (mixed) models					
Cluster-specific 1 (*Superiority only*)	A model with a random intercept (i.e., model differences among dogs in average MDCT) but no other random effects.	-626	-627	--	--
Cluster-specific 2 (*Superiority only*)	A model with a random intercept and a random difference between regions (i.e., model dog-to-dog variation in the MDCT difference between the Infarct and Remote regions).	-826	-827	--	--
Cluster-specific 3 (*Superiority only*)	A model with a random intercept and a random difference between regions, and the residual spatial correlation among slices within rings (after accounting for the dog-specific intercepts and slopes) is also modeled, assuming that measures from slices adjacent in space are more highly correlated than those further separated in space.	-829	-829	--	--
Cluster-specific 1(E) (*Equivalence only*)	Dog-to-dog variation in intercept and slope were not modeled; the residual spatial correlation among slices within rings (after accounting for the dog-specific intercepts and slopes) is also modeled, assuming that measures from slices adjacent in space are more highly correlated than those further separated in space.	--	--	-880	-872
Population average (marginal) models					
Population average 1	A model with the same coefficients as in the model ignoring clustering, but standard errors adjusted for design effects, taking non-independence into account through adjustment of variance estimates.	N/A ‡	N/A ‡	N/A ‡	N/A ‡
Population average 2	A generalized estimating equations (GEE) model assuming autoregressive(1) as the working correlation structure. This assumes that observations closer together in space (defined in terms of region, ring and slice) are more highly associated.	N/A ‡	N/A ‡	N/A ‡	N/A ‡
Population average 3	A generalized estimating equations (GEE) model assuming exchangeable working correlations. The assumption here is that there is a constant correlation between observations within dogs that does not depend on spatial proximity among measures in the heart.	N/A ‡	N/A ‡	N/A ‡	N/A ‡

† For the fit indices (AIC and BIC), the smaller the value, the better the model fit.

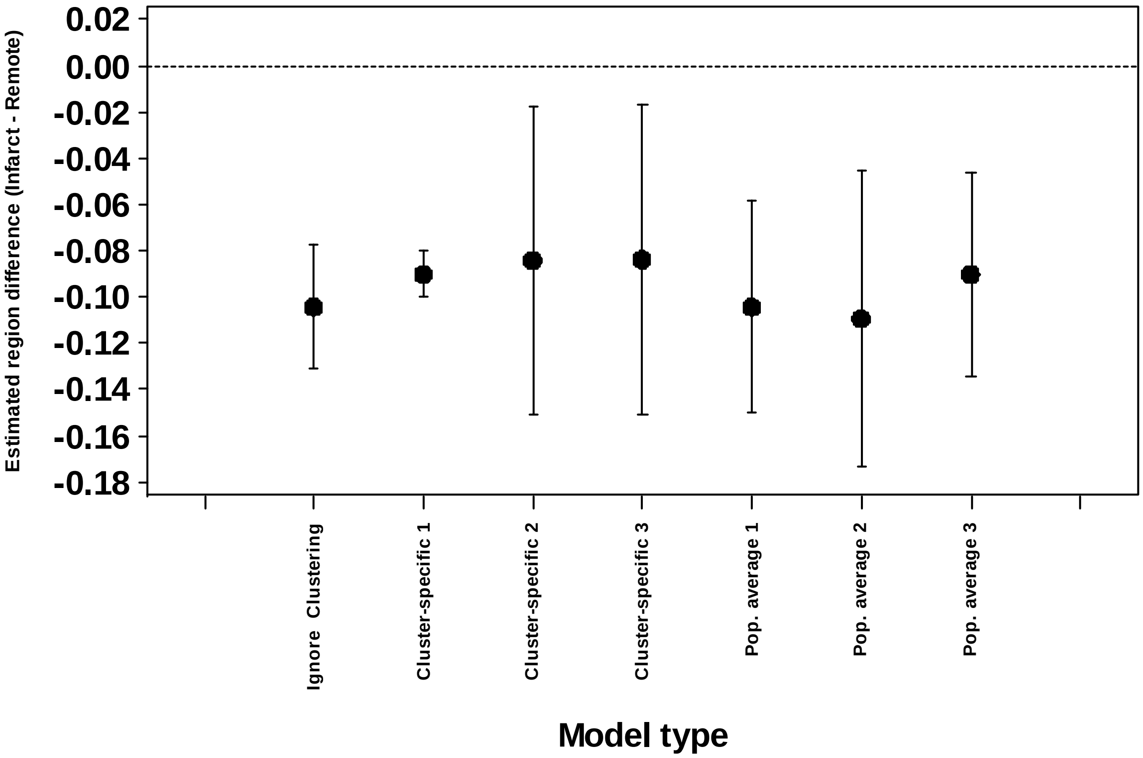
‡ Because estimation for the population average models involves pseudo- or quasi-likelihood methods, likelihood-based fit indices that were not available.

All models were estimated in SAS 9.1 (PROC MIXED for the cluster-specific models; PROC SURVEYREG for Population average 1; and PROC GENMOD for all other models).

The two general approaches (cluster-specific vs. population average) were compared based on 1) whether they lead to the same conclusions regarding superiority (hypothesis tests); 2) the magnitude of the estimates (betas and confidence intervals); 3) bias of estimates; and 4) risk of overfitting, i.e., tendency to capitalize on "noise" in a specific sample, reducing the likelihood that results will generalize to other, similar samples the same population.

Model results are shown in Figure 1.

Figure 1. Estimated regional difference (infarct minus remote) and 95% confidence interval estimated by regression models.



This displays the model-estimated difference between regions (MDCT for infarct minus MDCT for remote) with 95% confidence interval. In all methods, MDCT for the Infarct region was significantly less than that in the Remote region. Both the estimated difference and CI differed among regions. All methods except the mixed model with a random intercept only had wider confidence intervals than the model ignoring clustering.

Method Comparison (Equivalence)

As described above, the two methods (MDCT and SIM-MDCT) were compared using the Bland-Altman techniques, mountain plots, Deming regression and Passing-Bablok regression. All of these techniques were examined using individual slice measurements but treating them as independent (n = 163). In addition, to take clustering into account, confidence intervals for Deming and Passing-Bablok regression were adjusted for design effects.

MDCT and SIM-MDCT measures were also compared using generalized linear modeling with clustering ignored, as well as regression taking clustering into account, specifically using mixed or "cluster-specific" models (for example, Difference$_{ijkl}$ = $\gamma_0 + u_{1i} + \gamma_1$ * Sum$_{ijkl}$ + u_{2i} * Sum$_{ijkl}$ + $\varepsilon_{ijkl,}$ where Difference$_{ijkl}$ is the difference between MDCT and SIM-MDCT measures of perfusion in slice l within ring k within region j within dog k's heart, and Sum$_{ijkl}$ is the sum of these same measures) and population average or marginal models (for example, Difference$_{ijkl}$ = $\gamma_0 + \gamma_1$ * Sum$_{ijkl}$ + ε_{ijkl}). Variants of these models were described above (see Table 4). Evidence of equivalence is provided when the estimated Intercept and Sum coefficient do not differ from 0 (i.e., H_0: $\gamma_0 = 0$ and $\gamma_1 = 0$ are *not* rejected).

The results of method comparison techniques using Bland-Altman, mountain plots and Deming and Passing-Bablok regression are shown in Figures 2-5. Intercept and slope estimates for Deming and Passing-Bablok regression are also shown in Table 5. Using the software package SUDAAN, we computed the ICC and Design Effect for the method comparison analysis, ICC= -0.03 and the DEFF = 0.33. The negative ICC (in contrast to the superiority analysis) is due to the method for simulating the data; when a large value was randomly generated within a given cluster, the other values within a cluster were unlikely to also be large.

Table 5. Intercept and slope estimates with 95% confidence intervals for Deming and Passing-Bablok regression, using two approaches (treat slices as independent vs. design effects adjustment for clustering).

Technique and estimated parameter	**Ignore clustering (treat observations as independent)**		**Design effects adjustment for clustering**	
	Estimate	95% CI	Estimate	95% CI
Deming				
Intercept	0.0015	-0.0046 to 0.0075	0.0015	-0.0026 to 0.0043
Slope	0.9972 *	0.9729 to 1.0214	0.9972 †	0.9833 to 1.0111
Passing-Bablok				
Intercept	0.0026	-0.0031 to 0.0085	0.0026	-0.0018 to 0.0049
Slope	0.9902	0.9681 to 1.0145	0.9902	0.9775 to 1.0042

* No significant difference from linearity (p's > 0.10) by the csum test.
† The design effects approach was only used to adjust the confidence intervals of the intercept and slope estimates; there was no csum test of linearity adjusted for design effects.

Figure 2. Bland Altman Plot for analysis treating individual slice measures as independent (n = 163).

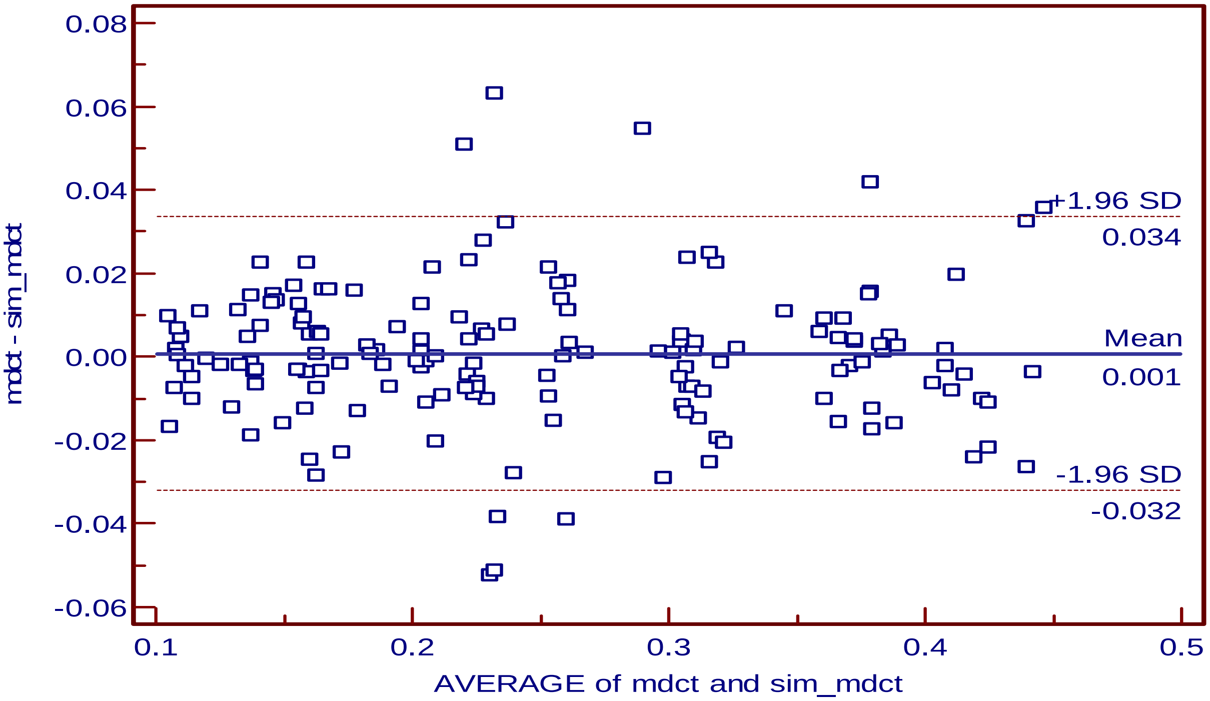


Figure 3. Mountain Plots (Folded Empirical Cumulative Distribution Plot), for analysis treating individual slice measures as independent (n = 163).

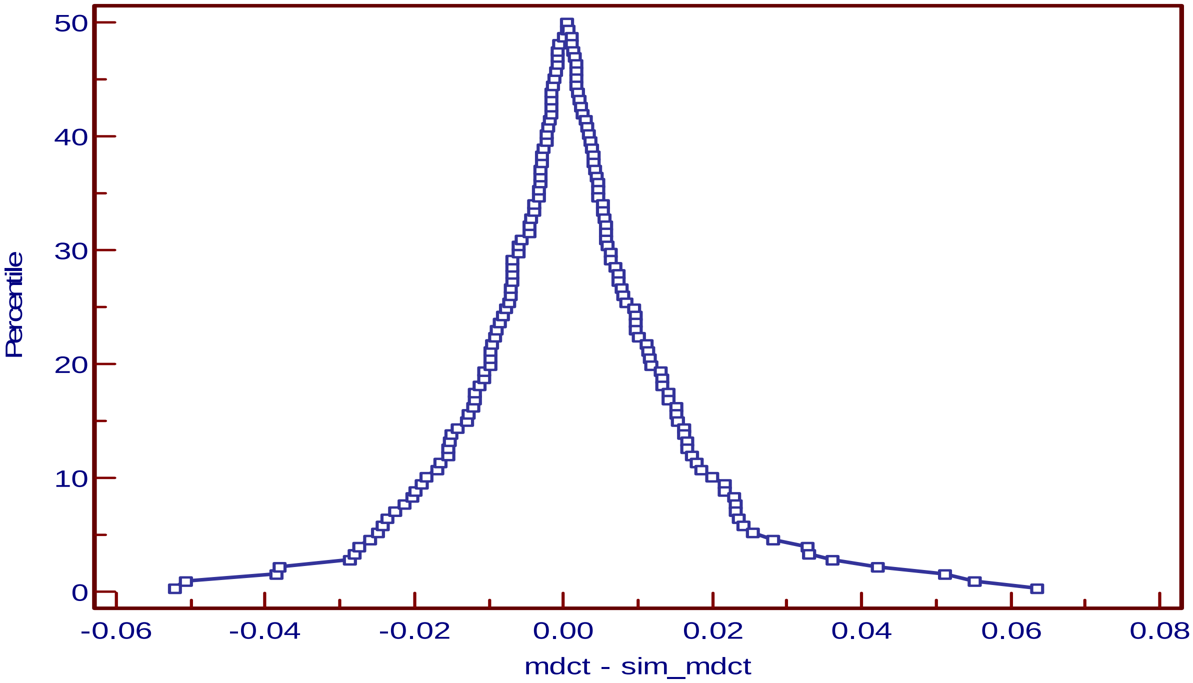


Figure 4. Deming Regression for analysis treating individual slice measures as independent (n = 163).

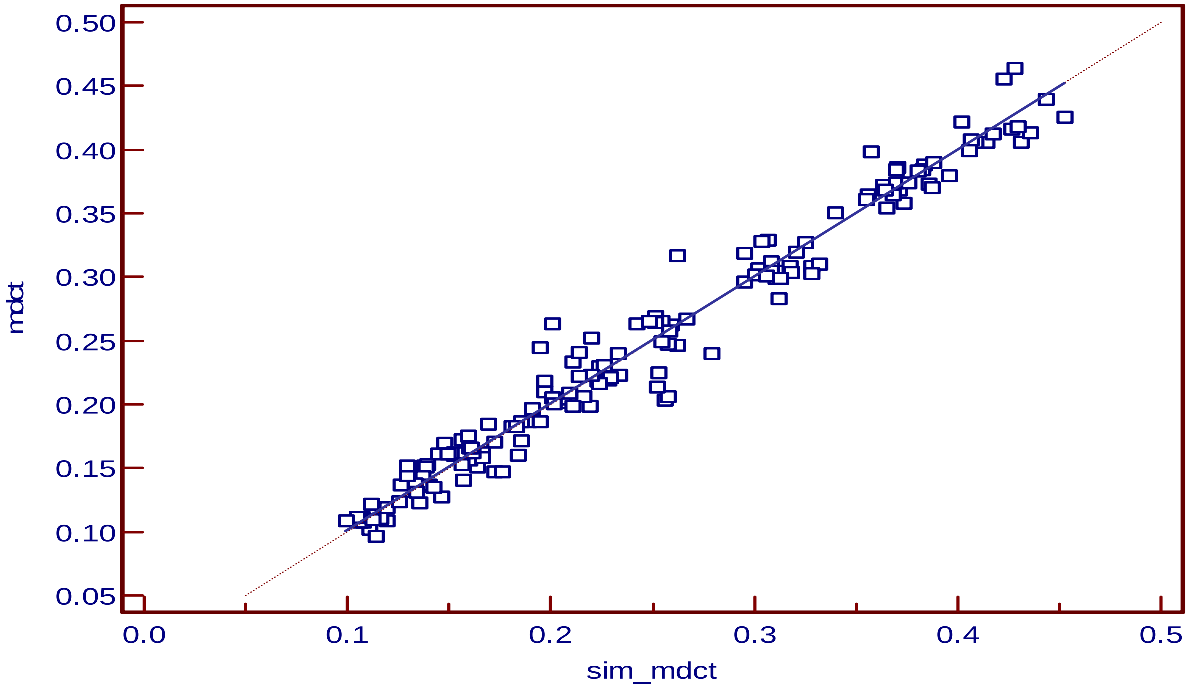


Figure 5. Passing-Bablok Regression for for analysis treating individual slice measures as independent (n = 163).

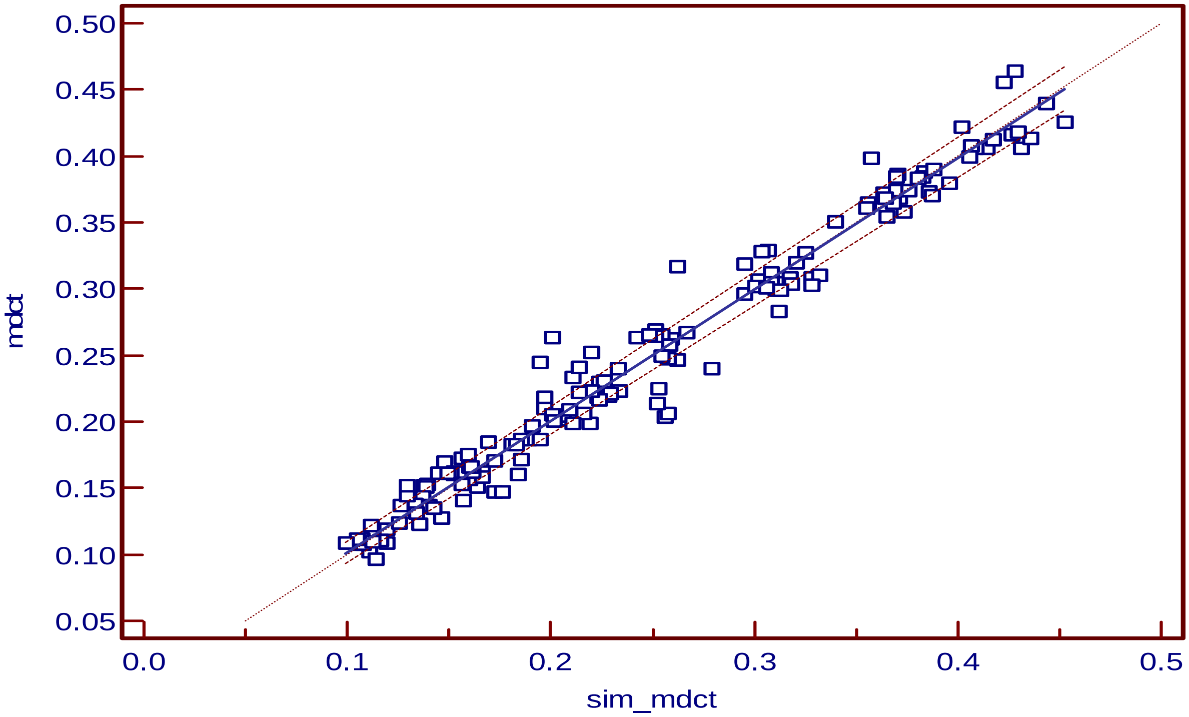
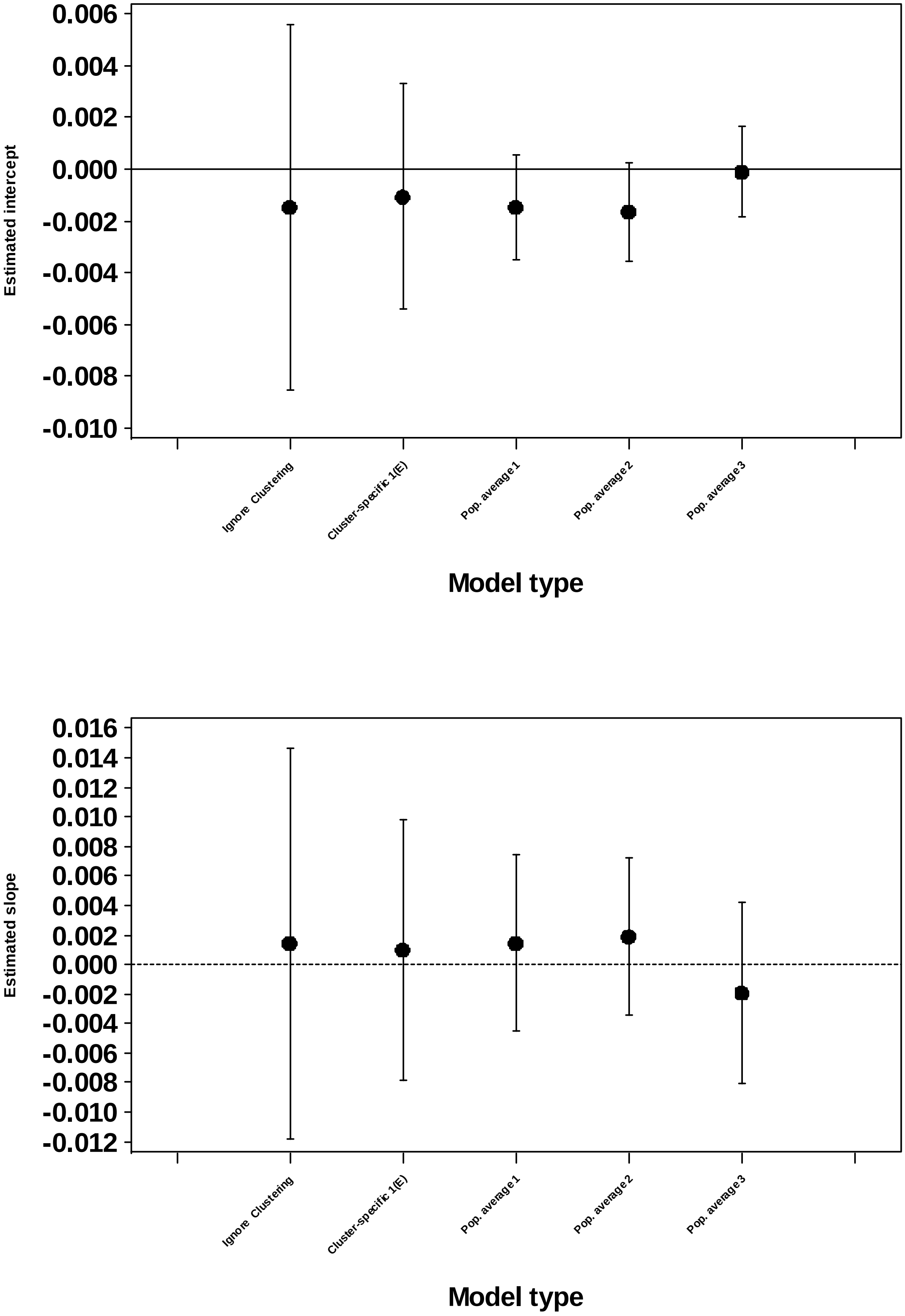


Figure 6. Estimated intercept (top panel) and slope (bottom panel), with 95% confidence intervals, estimated for method comparison using regression models.



The estimates in Figure 6 show that all methods support the conclusion of equivalence (Intercept = 0, Slope = 0). In contrast to the superiority example, where the design effect was greater than 1 (positive intra-class correlation), in the equivalence dataset the design effect was less than 1 (negative intraclass correlation), therefore the methods taking clustering into account had tighter confidence intervals than the method ignoring clustering.

E. Conclusions

Cardiac imaging data typically involve a set of nested structures (e.g., dog, region, ring, slice). This poster illustrates how the method of handling the nested data structure may impact both point and variance estimates. This is true of equivalence analyses as well as superiority analyses. Ignoring the nested data structure can lead to biased estimates; whether the estimates are biased upward or downward is situation-specific (e.g., ignoring clustering when there is a positive intra-class correlation can lead to confidence intervals that are too tight, while ignoring clustering when there is a negative intra-class correlation can lead to confidence intervals that are too wide). In clinical situations, such biases could lead to treatment decisions that are either too aggressive or not aggressive enough.

In the examples given here, different methods of taking the nested structure (clustering) into account (e.g., mixed models, design effects adjustments and GEE) generally produced similar results. This may not always be the case, perhaps especially in situations when there are outlier clusters and/or small samples. In such situations, due to the shrinkage toward the mean used in mixed models, mixed model estimates may be less affected by a few leverage points. In many analyses, it may be useful to do a sensitivity analysis, examining the results based on several different methods of handling the clustering. If the results differ greatly based on the method of handling clustering, it is important to delve into the reasons for such differences.

F. References

Altman DG. (1991). Practical Statistics for Medical Research. Chapman and Hall.

Armitage P, Berry G.(1994). Statistical Methods in Medical Research (3rd edition). Blackwell.

Bowman FD, Waller LA. Modelling of cardiac imaging data with spatial imaging. Stat Med. 2004 Mar 30;23(6):965-85.

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.

Bland JM and Altman DG (1986). Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. Lancet, Feb 8; 1 (8476): 307-10.

Altman, D.G. and Bland, J.M. (1983). Measurement in Medicine: The Analysis of Method Comparison Studies, The Statistician, 32, 307-317.

Bland, J.M. and Altman, D.G. (1999). Measuring Agreement in Method Comparison Studies, Statistical Methods in Medical Research, 8, 135-160.

Combleet PJ, Gochman N (1979) Incorrect least-squares regression coefficients in method-comparison analysis. Clinical Chemistry, 25:432-438.

Conover W.J. (1980). Practical Nonparametric Statistics. 2nd edn. John Wiley and Sons,New York.

George, R.T., C. Silva, M.A.S. Cordeiro, D. R. Thompson, W. F. McCarthy, T. Ichihara, J.A.C. Lima, A.C. Lardo (2006). Multi-Detector Computed Tomography Myocardial Perfusion Imaging During Adenosine Stress. Journal of the American College of Cardiology, Vol. 48, No. 1: 153-160.

Hawkins DM (2002). Diagnostics for conformity of paired quantitative measurements, Statistics in Medicine, 21: 1913-1935.

Krouwer JS, Monti KL (1995) A simple, graphical method to evaluate laboratory assays. Eur J Clin Chem Clin Biochem, 33:525-527.

Lavange, LM, Koch, GG, Schwartz, TA. (2001). Applying sample survey methods to clinical trials data. Statistics in Medicine, 2609-2623.

McCarthy WF and Gable JM (1999). A comparison of two methods for making inferences about the parameters of the logistic regression model. Presented at the International Biometric Society Eastern North American Region (1999 Spring Meeting in Atlanta, Georgia).

Passing H, Bablok W (1983) A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part I. J. Clin. Chem. Clin. Biochem., 21:709-720.

Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods. Biometrics, 1, 80-83.