**Distribution Agreement**

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Alison Zinsli Date

**Evaluation of agreement measures among groups of raters with an application to the interpretation of kidney obstruction.**

By:

Alison Zinsli

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

_____

Amita Manatunga, Ph.D
Committee Chair

_____

Renee Moore, Ph.D.
Committee Chair

**Evaluation of agreement measures among groups of raters with an application to the interpretation of kidney obstruction.**

By:

Alison Zinsli

B.S.

Wake Forest University

2016

Thesis Committee Chair: Amita Manatunga, Ph.D.

The abstract of
A thesis submitted to the Faculty of the
Rollins school of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2018

## Abstract

**Evaluation of agreement measures among groups of raters with an application to the interpretation of kidney obstruction.**

By:
Alison Zinsli

Kidney obstruction prevents the kidneys from properly draining which can lead to loss of function if left untreated. The Department of Nuclear Medicine at Emory University is developing a decision supporting software called RENEX to assist radiologists in limiting their errors and arriving at the correct diagnosis when interpreting the renal scans. In the absence of a gold standard, experts assessment of kidney obstruction is considered to be the best available standard. The objective of this study aims to quantify the agreement among experts and residents with and without the RENEX intervention and to address the question of whether the RENEX intervention helps the residents perform similar to the experts by quantifying the agreement between groups.

Three experts and three residents with and without the RENEX educational intervention interpreted data from 50 patients for both their left and right on degree of obstruction. They classified obstruction on a continuous scale from -1 to 1 and could be categorized into three groups: unobstructed [-1, -0.2), undetermined [-0.2, 0.2] and obstructed (0.2, 1]. Agreement was evaluated within groups using the concordance correlation coefficient (CCC) and weighted kappa. Further analysis was done to determine whether a resident can replace an expert in the interpretation of kidney obstruction. Since the same patient is evaluated multiple times the observations are correlated and a bootstrap methodology was used to calculate accurate standard error and confidence intervals.

The agreement index of CCC for experts for the left and right kidney are 0.819 (0.619, 0.937) and 0.866 (0.706, 0.935), respectively. Whereas, residents with and without the use of RENEX CCC agreement for the left kidney is 0.314 (0.132, 0.492) and 0.726 (0.504, 0.874), respectively; and 0.347 (0.107, 0.558) and 0.680 (0.435, 0.835) for the right kidney. There is a significant difference between the agreement of experts and residents ($p<0.001$, left and right kidney) but when residents use RENEX the difference in agreement is no longer significantly different ($p=0.505$, $p=0.050$; left and right kidney, respectively). A similar pattern can be seen when an expert is replaced by a resident; the agreement is better when the resident uses RENEX.

In conclusion, not only did residents agreement improve, but RENEX also helped improve the accuracy of the resident's classification compared to the expert's. The methodology that was used in determining whether a resident can replace an expert is generally applicable to other similar studies.

# Evaluation of agreement measures among groups of raters with an application to the interpretation of kidney obstruction.

By:

Alison Zinsli

B.S.

Wake Forest University

2016

Thesis Committee Chair: Amita Manatunga, Ph.D.

The thesis submitted to the Faculty of the
Rollins school of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2018

**Acknowledgements**

Thank you to the entire Biostatistics Department at Emory University. I have had the greatest support from a variety of professors in the department who have helped me develop as a biostatistician. More specifically I would like to thank my thesis advisor, Dr. Amita Manatunga, for her incredible mentorship as an advisor and as a professor. I have learned a great deal from working with her. I would also like to thank Jeong Jang for working with me on a variety of questions I had through-out this project. Finally, thank you to Dr. Renee Moore for helping me finalize my thesis to prepare it for submission.

I would not where I am or who I am today without the love and support of my Mom and Dad. Without their encouragement and generosity, I would not have had the opportunities I have been given. A special thanks to all my friends who have helped me and supported me through-out my educational career.

**Table of Contents**

## Introduction

Prevalence of chronic kidney disease has increased by 30% in the past decade and is now estimated to affect 27 million Americans and accounts for more than 24% of all Medicare costs (Taylor, 1997). Detecting renal disease early is imperative in order to effectively treat the disease and prevent progression. Therefore, effective strategies are urgently needed detect early renal disease, improve diagnostic accuracy, direct therapy and monitor the patient's response to treatment. One method that has played an important role in the management of patients with known or suspected renal disease is nuclear medicine renal scans. These method is particularly useful in patients that have suspected obstruction to drainage from their kidney's which can lead to loss of function for the affected kidney if left untreated. To perform renal scans, an intravenous injection of gamma emitting tracer (MAG3) is administered and is rapidly removed from the blood by the kidneys and then travels down the ureters to the bladder. The movement of MAG3 through the kidneys can be modeled by a time activity curve (renogram curve), generated by placing a region of interest (ROI) over each kidney and counting the photons detected in the kidney ROI at multiple intervals during an initial 20 to 24 minute period of data collection. In patients with suspected kidney obstruction an additional data collection period is conducted after the administration of a potent diuretic. Once all this data is acquired, interpretation of the renal scan is based on the analysis of images and the renogram curve. The interpretation is usually conducted by radiologist who could have as little as 4 months training in all nuclear medicine (Xu, 2009). With an estimated 590,000 renal scans performed annually and the lack of training radiologist receive allows for increased error rates in patient diagnosis of kidney obstruction. If a patient is falsely

classified as having a high degree of obstruction then they would have undergone an unnecessary treatment and if a patient is falsely classified as having little or no kidney obstruction then more health complications could arise.

The Department of Nuclear Medicine at Emory University is developing a method to address this problem by assisting radiologists in limiting their errors and arriving at the correct diagnosis when interpreting the renal scans and MAG3 data. The product is a decision supporting system (DSS) for kidney obstruction called RENEX which is a knowledge base system of heuristic rules based on the interpretations of kidney experts and quantitative variables extracted from renograms to conclude whether a kidney is obstructed (Taylor, 2012). However, evaluating the reliability and validity of RENEX has proven to be challenging due to the lack of a gold standard in kidney obstruction classification. The closest thing to the "gold standard" is the classification the experts determine but even between the experts there is variability and no perfect agreement between them.

A study was conducted at Emory University to assess the performance of residents from the Department of Radiology. Three residents and three experts rated patient scans for both their left and right kidney on a continuous scale for degree of obstruction. Additionally, residents rated the kidney obstruction while using the RENEX intervention. Receiver operating curves (ROC) are used to evaluate and compare the performance of new DSS methods when the diagnosis is definite. Since experts do not always agree there is variability among ratings. The concordance correlation coefficient (CCC) and weighted kappa can be used to assess the agreement within a group to account for the variability (Albert, 2007) However, if there is high agreement on degree of

obstruction between the residents, they could still be wrong and the agreement calculation is not designed to directly address that issue. This thesis addresses how to evaluate the residents with and without the use of RENEX compared to the experts and considers an empirical method to compare the performance of residents with and without RENEX to experts.

When it comes to determining agreement between observers or measurements on a continuous scale there are a variety of ways in which analysis can be approached. Methods such as the Pearson correlation coefficient, least squares, paired t-test, coefficient of variation and intraclass correlation coefficient all have strengths and weaknesses when it comes to determining agreement (Lin, 1989). The Pearson correlation coefficient is an appropriate method to measure a linear relationship but it cannot to detect any departure from the 45º line (Lin, 1989). The least square approach fails to detect departure from an y=x line if the data are scattered and if the data is highly agreeable there is a chance the least squares approach cannot detect it due to the small residual error (Lin, 1989). The paired t-test cannot assess poor agreement in paired data (Lin, 1989). The coefficient of variation and intraclass correlation both view duplicate measurements as random which is not necessarily the case and should not be viewed as such (Lin, 1989). The concordance correlation coefficient (CCC) is a method that addresses the weaknesses from the other methods by evaluating agreement between duplicate observations from the same sample by measuring the variation from the 45º line through the origin (the concordance line) (Lin, 1989). The CCC measures how far each observation deviates from the line fit to the data (precision) as well as how far the line

deviates from the 45º line through the origin (accuracy) (King, 2007). The CCC lies between -1 and 1 where -1 is prefect disagreement and 1 equates to perfect agreement.

Depending on the data, agreement can also be determined for categorical classification by using the kappa statistic. When considering ordinal categorical variables, weighted kappa is a popular method to use. The weighted kappa considers that classifications that are close to each other allow for better agreement than those that are further away. In the case of kidney obstruction, it is important to look at the categorization of the degree of kidney obstruction because if it is obstructed then it requires surgery. Therefore, even if the raters might not exactly agree on the continuous scale, if they come to the same conclusions on the categorical scale then the proper medical advice can be implemented.

Unfortunately, the measurements between the raters are not independent because the measurements are clustered within each of the patients. When there is a dependence between the raters, the standard error of the agreement statistic is inaccurate because it violates the independence assumption. In order to calculate the standard error between the observers a bootstrapping method was used.

This thesis aims to address the following questions (1) to quantify the agreement among experts and residents with and without the RENEX intervention and (2) to address the question of whether the RENEX intervention helps the residents perform similar to the experts by quantifying the agreement between groups.

## Methods

### The Data

Data was collected from three residents and three experts on their classification on degree of kidney obstruction in fifty patients for both their left and right kidney. Interpretation of renal scans and time activity analysis are used in classifying kidney obstruction and requires extensive training and experience in order to accurately assess the kidneys (Jang). Nuclear medicine residents with at least one year of training were asked to assess the same renal scans as the kidney radiologist experts. Furthermore, the residents used an educational software called RENEX to help assist with classification in hopes of improving their score and minimizing their variability. There are 3 independent repeated measures for each kidney for each patient from both the experts, residents and residents using RENEX. The degree of obstruction in a kidney is rated on a continuous scale from -1 to 1. A kidney is considered to be unobstructed if it is rated between -1 and -0.2, obstructed between 0.2 and 1 and is undeterminable between -0.2 and 0.2.

### Concordance Correlation Coefficient

Lin proposed the method of calculating concordance correlation coefficient as a measure of accuracy and precision in agreement measures. Assume that pairs of samples $(Y_{i1}, Y_{i2}), i = 1, 2, \dots, n$ are independently selected from a bivariate population with means $\mu_1$ and $\mu_2$ then the degree of concordance between $Y_1$ and $Y_2$ can be characterized by the expected values of the squared difference

$$E[(Y_1 - Y_2)^2] = (\mu_1 - \mu_2)^2 + (\sigma_1 + \sigma_2)^2 + 2(1 - \rho)\sigma_1\sigma_2$$

where $\rho$ is the Pearson correlation coefficient. Therefore, the CCC between two observers based on their variances, covariances and means can be defined as

$$\rho_{ccc} = \frac{\{2\sigma_{12}\}}{\{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2\}}$$

where $\sigma_{12}, \sigma_1^2, \sigma_2^2, \mu_1, \mu_2$ are the covariance, variances and means of the two observers, respectively. King extended this statistic to work for two observers that have *p* repeated measures for *n* patients given by the following equation

$$\rho_{ccc} = \frac{\sum_{j=1}^{p}\sum_{k=1}^{p} d_{jk}(\sigma_{12_{jk}} + \sigma_{21_{jk}})}{\sum_{j=1}^{p}\sum_{k=1}^{p} d_{jk}\left(\sigma_{11_{jk}} + \sigma_{22_{jk}}\right) + \sum_{j=1}^{p}\sum_{k=1}^{p} d_{jk}(\mu_{1_j} + \mu_{2_j})(\mu_{1_k} + \mu_{2_k})}$$

where $\sigma_{12_{jk}}$ is the covariance between the measurements of observer 1 at time *j* and observer 2 at time *k*; $\sigma_{21_{jk}}$ is the covariance between the measurements of observer 1 at time *k* and observer 2 at time *j*; $\sigma_{11_{jk}}$ is the covariance between observer 1 at time *j* and *k*; $\sigma_{22_{jk}}$ is the covariance between observer 2 at time *j* and *k*; and $\mu_1$ and $\mu_2$ are the means at time *j* and *k* and $d_{jk}$ is an arbitrary weight.

Carrasco and Jover developed a method to estimate the CCC using the intraclass correlation coefficient for repeated measures from the variance components model

$$Y_{ijt} = \mu + \alpha_i + \beta_j + \epsilon_{ijt}, t = 1, \ldots, p$$

where $Y_{ijt}$ is the measurement taken by observer *j* on subject *i* for *t* repeated measures; $\mu$ is the overall mean over subjects and observers; $\alpha_i$ is the subject random effect assumed to be distributed as $\alpha_i \sim N(0, \sigma_\alpha^2)$; $\beta_j$ is the mean deviation of observer *j* from the overall mean; and $\epsilon_{ijt}$ is the random error assumed to be distributed as $\epsilon_{ijt} \sim N(0, \sigma_e^2)$. This leads to the following formula to calculate CCC

$$\rho_{ccc} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_e^2}$$

Calculation of the CCC using variance components is included in the appendix.

Furthermore, the CCC can be calculated through U-statistics where the distributions of the estimators are asymptotically normal and consistent estimators for variances when using moderately large sample sizes. The CCC for repeated measures using U-statistics can be described as

$$\rho_{ccc} = \frac{(n-1)(V-U)}{U+(n-1)V}$$

where $nU = \sum_i (\boldsymbol{X}_i - \boldsymbol{Y}_i)' \boldsymbol{D} (\boldsymbol{X}_i - \boldsymbol{Y}_i)$ and $n(n-1)V = \sum_{i \neq j} (\boldsymbol{X}_i - \boldsymbol{Y}_j)' \boldsymbol{D} (\boldsymbol{X}_i - \boldsymbol{Y}_j)$ and $(\boldsymbol{X}_{ij}, \boldsymbol{Y}_{ij})$ are the measurements from the observers.

$\bar{Y}$ and $S^2$ can replace $\mu$ and $\sigma^2$ to approximate CCC and the variance for the estimator can be defined as

$$\sigma^2_{\widehat{\rho_c}} = \frac{1}{n-1}\left[\frac{(1-\rho^2)\rho_c^2(1-\rho_c^2)}{\rho^2} + \frac{4\rho_c^3(1-\rho_c)\mu^2}{\rho} - \frac{2\rho_c^4\mu^4}{\rho^2}\right]$$

The estimator of CCC can have an improved normal approximation by using the inverse hyperbolic tangent transformation

$$Z = \frac{1}{2}\ln\frac{1+\hat{\rho}_{ccc}}{1-\hat{\rho}_{ccc}}$$

from there the delta can be used to obtain its asymptotic distribution where $\hat{Z} \sim N(0, V(\hat{Z}))$ and $V(\hat{Z})$ is approximately

$$V(\hat{Z}) = \frac{\sigma^2_{\widehat{\rho_c}}}{(1-\rho^2)^2} = \frac{1}{n-2}\left[\frac{(1-\rho^2)\rho_c^2}{(1-\rho_c^2)\rho^2} + \frac{4\rho_c^3(1-\rho_c)\mu^2}{\rho(1-\rho^2)^2} - \frac{2\rho_c^4\mu^4}{\rho^2(1-\rho_c^2)^2}\right]$$

Using the Z-transformation for confidence intervals for $\widehat{\rho_c}$ keeps the confidence interval in between (-1, 1) and therefore provides a more realistic interval (Carrasco, 2013).

**Weighted Kappa**

Weighted kappa is an appropriate agreement measurement when considering ordinal categorical observations. In the case of kidney obstruction, the ordinal categories

that can be created from the continuous data are unobstructed [-1, -0.2), undeterminable [-0.2, 0.2] and obstructed (0.2, 1]. Similar to CCC, -1 indicates prefect disagreement and 1 indicates perfect agreement. Weighted kappa can be given by the equation

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij} p_{ij}}{\sum_{i,j} w_{ij} e_{ij}}$$

where $w_{ij}$ are the weights for how close to the diagonal (perfect agreement) the rates are, $p_{ij}$ are the observed probabilities and $e_{ij}$ is the expected probabilities. Therefore, the kappa statistic takes into account the proportion of agreement that is expected by chance (Williamson, 2000). The standard error for weight kappa can be given by

$$SE = \frac{1}{1 - p_{e(w)}} \sqrt{\frac{\sum_{i,j} p_{ij}\left[v_{ij} - u_{ij}(1-\kappa_w)\right]^2 - \left[\kappa_w - p_{e(w)}(1-\kappa_w)\right]^2}{n}}$$

where

$$v_{ij} = 1 - \frac{w_{ij}}{w_{max}}, p_{e(w)} = \sum_i \sum_j v_{ij} p_i q_j \,, u_{ij} = \sum_h q_h v_{ih} + \sum_h p_h v_{hj}$$

**Bootstrap Confidence Intervals**

In this study the same patient is evaluated by three experts and three residents with and without the use of RENEX. Therefore, the observations are correlated to each other. Additionally, when calculating agreement within the groups, the agreement statistic is calculated between all possible pairs and averaged to obtain the overall agreement. This means that the standard error and confidence intervals cannot be easily generated from the software. However, the bootstrap method helps determine the true variance for the CCC and weighted kappa. The fundamental idea of the bootstrap is to perform calculations on the data to estimate the variations of the statistics that are computed from that dataset.

Sampling is the first step to the bootstrap method – either with or without replacement. In the case of the kidney patients, sampling with replacement was conducted with the 50 patients included in the data for either the left or right kidney. Once the sample is collected, the agreement statistic is calculated from that sub sample. Then another random sample is collected from the dataset and the agreement statistic is recalculated. This process is repeated many times. Ultimately the bootstrap method can be summarized as:

1. $x_1, \dots, x_n \sim F$ is sampled with replacement from the data (F)

2. the agreement statistic $\widehat{\theta_n^{(1)}}$ is calculated from the sample

3. $x_1^*, \dots, x_n^* \sim F$ is resampled with replacement of the same sample size as step one

4. the agreement statistic $\widehat{\theta_n^{(2)}}$ is recalculated and stored.

Once this process is repeated K times the mean of $\widehat{\theta_n^{(B)}}, B = 1, \dots, K$ is the agreement between the observers and the variance of the agreement can be calculated as

$$s^2 = \frac{1}{B}\sum_{j=1}^{B}\left(\widehat{\theta_n^{(j)}}\right)^2 - \left(\frac{1}{B}\sum_{j=1}^{B}\widehat{\theta_n^{(j)}}\right)^2$$

and by the law of large numbers

$$s^2 \xrightarrow{F} E\left(\widehat{\theta_n}\right) - \left(E\left(\widehat{\theta_n}\right)\right)^2 = V_P\left(\widehat{\theta_n}\right) = S_n(F)$$

In other words, the variance of the data, F, can be approximated by repeatedly simulating *n* observations from *F*.

A bootstrap confidence interval can also be calculated based off of the repeated agreement statistics that were calculated. Once $\widehat{\theta_n^{(1)}}$ ... $\widehat{\theta_n^{(B)}}$ is calculated let

$$\hat{F}(t) = \frac{1}{B}\sum_{j=1}^{B} I\left(\sqrt{n}\left(\widehat{\theta_n^{(j)}} - \overline{\theta_n}\right) \le t\right)$$

then the bootstrap confidence interval is equal to

$$\left[\overline{\theta_n} - \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n}}, \overline{\theta_n} - \frac{t_{\frac{\alpha}{2}}}{\sqrt{n}}\right] \text{ where } t_{\frac{\alpha}{2}} = \widehat{F^{-1}}\left(\frac{\alpha}{2}\right) \text{ and } t_{1-\frac{\alpha}{2}} = \widehat{F^{-1}}\left(1-\frac{\alpha}{2}\right)$$

**Exploratory Analysis**

Investigating the differences in the categorization for each resident compared to the three experts was determined both continuously and categorically. First, it was determined which categorization each patient belonged in for kidney obstruction for both their left and right kidney a 'majority rule' view was taken. In other words, patients were classified into a category when all three or two out of the three assigned numbers that were in the same category. If all three of the experts placed a patient in different categories than that patient was placed in the undeterminable group. Next, the categorization of the resident for each kidney was determined and was compared to the final classification of the experts to see if it agreed or disagreed.

When considering the continuous categorization of the kidney obstruction the classification of the three experts was averaged. The resident's classification was subtracted from the experts. If the difference between the experts and the resident is close to zero then that means the resident was fairly accurate in their classification. If the difference is large then that corresponds to the residents being very inaccurate to the expert's classification of kidney obstruction. If the difference was very negative then the residents classified the kidney as being more obstructed than the experts concluded and if the different is very positive then the residents classified the kidney as being more unobstructed than the experts indicated. These differences were plotted in groups according to the categorization of kidney obstruction determined by the experts.

**Agreement for obstruction classification within groups**

Carrasco estimated the CCC in two ways: by variance components and by U-statistics. The CCC was calculated for the three groups for both the left and right kidney using these two methods but the variance needed to be calculated through bootstrapping since the observations were not independent. Additionally, the continuous classification of the kidney obstruction could be categorized into three groups: unobstructed [-1, -0.2), undetermined [-0.2, 0.2] and obstructed (0.2, 1]. Once these categorizations were created the weighted kappa could be used to determine the agreement between the raters.

*CCC Bootstrap*

The CCC only considers the agreement between two raters and N repeated observations for each patient. A bootstrap function was created that took a 30-patient random sample with replacement and patients that had missing data were not considered. The CCC was calculated for between raters 1, 2 and 1, 3 and 2, 3 for each of the three groups for both kidneys. The agreement between the 3 pairs of the raters were averaged and stored. This process was repeated 1000 times. The variance of the 1000 CCC values was calculated and 95% bootstrap confidence interval was calculated by taking the 0.025 and 0.975 percentiles.

*Weighted Kappa Bootstrap*

The weighted kappa only considers the agreement between two raters and N repeated observations for each patient. The continuous variables were transformed into categories and the data was reconfigured to the format of a nx2 matrix. A bootstrap function was created that took a 30-patient random sample with replacement and patients that had missing data were not considered. The weighted kappa was calculated for between raters 1, 2 and 1, 3 and 2, 3 for each of the three groups for both kidneys. The

agreement between the 3 pairs of the raters were averaged and stored. This process was repeated 1000 times. The variance of the 1000 weighted kappa values was calculated and 95% bootstrap confidence interval was calculated by taking the 0.025 and 0.975 percentiles.

**Replacement analysis to determine resident accuracy**

If the residents agree with each other that does not necessarily mean that they are correct when it comes to classifying the degree of kidney obstruction. In order to determine how accurate the residents are when it comes to classification, replacement of one of the experts with the resident allows us to see the difference in agreement compared to all three experts.

The agreement was calculated between two experts and one resident for both the left and right kidney before and after the use of RENEX using the CCC and weighted kappa. Similarly, the variance for these agreement values needed to be calculated through bootstrap functions like the ones described above. The bootstrap function was repeated 1000 times. The variance of the 1000 agreement values was calculated and 95% bootstrap confidence interval was calculated by taking the 0.025 and 0.975 percentiles.

**Comparing agreement between groups**

A bootstrap comparison function was created to compare the agreement between the groups. By comparing the resident's classification to their classification after using RENEX we can see how much they improved after using the software. Comparing the residents before and after they use RENEX to the experts shows how close or far their agreement compares to the ideal standard. Additionally, after replacing the residents with an expert in the accuracy analysis, a comparison to the three experts could show how the

residents classification impacted the agreement. A different comparison function was created for the different agreement methods used: CCC and weighted kappa.

The difference between the agreement of the two groups was calculated. In order to find the variance of the difference a bootstrap method was used. A bootstrap function took a random 30-patient sample with replacement from the 50 patients used. The CCC between the 3 raters for the two groups being compared for each kidney was calculated for that sample. The two CCC values from each group were calculated for 1000 repetitions and the difference between the CCC values was determined and stored. The variance of the difference was determined by calculated the variance of the 1000 CCC difference values. A 95% bootstrap confidence interval was calculated for the difference by taking the 0.025 and 0.975 percentiles. To determine if there was a significant difference between the groups a test statistic was calculated using the formula

$$T = \frac{CCC_{grp1} - CCC_{grp2}}{\sqrt{\sigma^2_{diff}}} \sim N(0,1)$$

This test the hypothesis $H_0: CCC_{grp1} = CCC_{grp2}\ vs.\ H_1: CCC_{grp1} \neq CCC_{grp2}$

The same procedure was conducted for the categorical classification using weighted kappa.

## Results

### Exploratory

For the left kidney the experts categorized 33 unobstructed kidneys, 6 undeterminable kidneys and 11 obstructed kidneys. In figure 1a you can see that residents 1 and 2 tend to have a positive difference so they tend to classify the kidney as being more unobstructed than the experts indicated but still agree with the categorization.

However, in the obstructed column you can see that the frequency of disagreement increases. Resident 3 has a negative difference and tends to disagree in the unobstructed categorization. In figure 1b, once the residents use RENEX, it is clear that the difference in their agreement gets closer to zero and the frequency of disagreement decreases. For the right kidney the experts categorized 33 unobstructed kidneys, 5 undeterminable kidneys and 8 obstructed kidneys. This pattern is the same for the right kidney and can be seen in figure 2a and 2b.

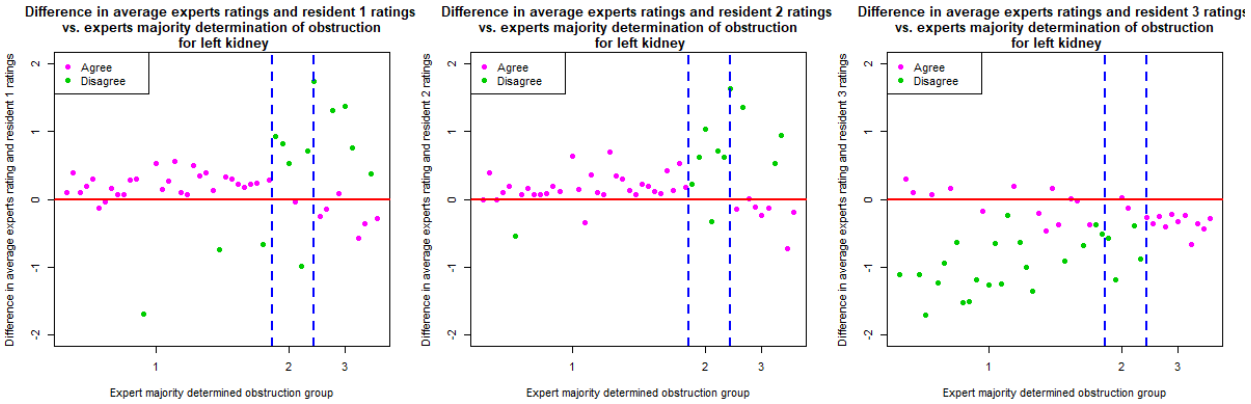



Figure 1a. The differences in classification of kidney obstruction between the three experts and residents one, two and three for the left kidney grouped by expert's majority rule categorization. The 1 represents the unobstructed group, the 2 represents the undetermined group and 3 represents the obstructed group. Purple points indicate the resident agreed categorically and the green indicates they disagreed categorically.

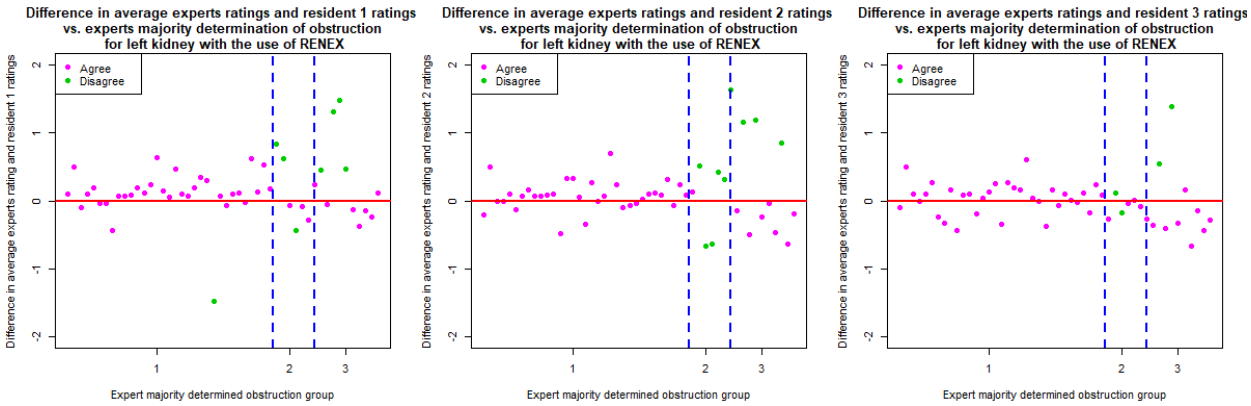



Figure 1b. The differences in classification of kidney obstruction between the three experts and residents one, two and three for the left kidney after the residents used RENEX grouped by expert's majority rule categorization. The 1 represents the unobstructed group, the 2 represents the undetermined group and 3 represents the obstructed group. Purple points indicate the resident agreed categorically and the green indicates they disagreed categorically.

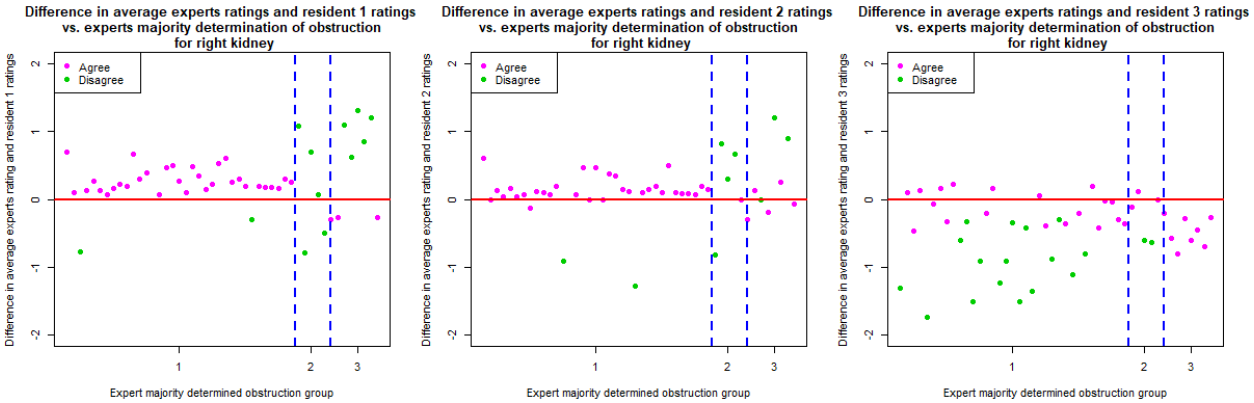



Figure 2a. The differences in classification of kidney obstruction between the three experts and residents one, two and three for the right kidney grouped by expert's majority rule categorization. The 1 represents the unobstructed group, the 2 represents the undetermined group and 3 represents the obstructed group. Purple points indicate the resident agreed categorically and the green indicates they disagreed categorically.

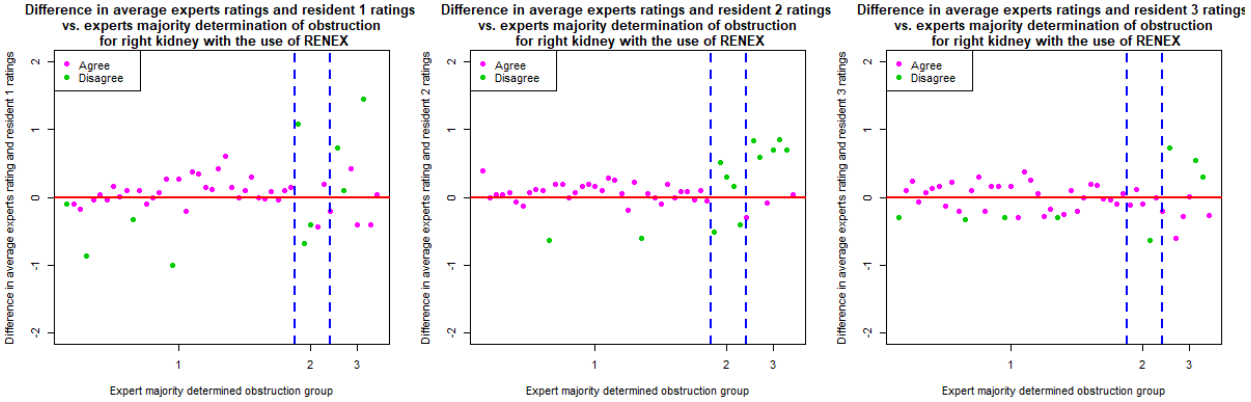



Figure 2b. The differences in classification of kidney obstruction between the three experts and residents one, two and three for the right kidney after the residents used RENEX grouped by expert's majority rule categorization. The 1 represents the unobstructed group, the 2 represents the undetermined group and 3 represents the obstructed group. Purple points indicate the resident agreed categorically and the green indicates they disagreed categorically.

## Agreement for obstruction classification within groups

Tables 1a indicates that the residents have poor agreement when it comes to classifying the degree of kidney obstruction on the continuous scale for the left (0.314 (0.132, 0.492)) and right (0.347 (0.107,0.558)) kidney. When using RENEX, the residents have improved agreement for the left (0.726 (0.504, 0.874)) and right (0.680 (0.435, 0.835)) kidney. The experts have the greatest agreement when it comes to classifying kidney obstruction for the left (0.819 (0.619, 0.937) and right (0.866 (0.706, 0.935)) kidney. Additionally, the variance for the experts CCC is smaller than the

variance for the residents with and without the use of RENEX. A similar pattern can be seen in Table 1b that demonstrates the agreement for the groups on the categorical scale using weighted kappa. The residents have poor agreement when it comes to classifying the degree of kidney obstruction for the left (0.267 (0.093, 0.447)) and right (0.308 (0.064,0.515)) kidney. When using RENEX, the residents have improved agreement for the left (0.752 (0.536, 0.905)) and right (0.573 (0.269, 0.812)) kidney. The experts have the greatest agreement when it comes to classifying kidney obstruction for the left (0.765 (0.515, 0.939) and right (0.774 (0.504, 0.932)) kidney.

**Agreement between raters using CCC**

| Group | Kidney | CCC | Bootstrap Variance | 95% Bootstrap Confidence Interval |
|---|---|---|---|---|
| Residents | Left | 0.314 | 0.0084 | (0.132, 0.492) |
| | Right | 0.347 | 0.0137 | (0.107, 0.558) |
| Residents with RENEX intervention | Left | 0.726 | 0.0093 | (0.504, 0.874) |
| | Right | 0.680 | 0.0103 | (0.435, 0.835) |
| Experts | Left | 0.819 | 0.0079 | (0.619, 0.937) |
| | Right | 0.866 | 0.0035 | (0.706, 0.935) |

Table 1a. The CCC, bootstrap variance and bootstrap confidence interval for the residents, residents using RENEX and experts for the left and right kidney. Due to missing data the CCC for the resident's agreement for the left kidney is calculated using 49 patients and 47 patients for the right kidney. When using RENEX, the agreement for the resident's left kidney is 50 patients and 47 for the right kidney. The expert's agreement for the left kidney is using 50 patients and 48 patients for the right kidney.

**Agreement between raters using Weighted Kappa**

| Group | Kidney | Weighted Kappa | Bootstrap Variance | 95% Bootstrap Confidence Interval |
|---|---|---|---|---|
| Residents | Left | 0.267 | 0.0087 | (0.093, 0.447) |
| | Right | 0.308 | 0.0134 | (0.064, 0.515) |
| Residents with RENEX intervention | Left | 0.752 | 0.0101 | (0.536, 0.905) |
| | Right | 0.573 | 0.0196 | (0.269, 0.812) |
| Experts | Left | 0.765 | 0.0118 | (0.515, 0.939) |
| | Right | 0.774 | 0.0122 | (0.504, 0.932) |

Table 1b. The weighted kappa, bootstrap variance and bootstrap confidence interval for the residents, residents using RENEX and experts for the left and right kidney. Due to missing data the weighted kappa for the resident's agreement for the left kidney is calculated using 49 patients and 47 patients for the right kidney. When using RENEX, the agreement for the resident's left kidney is 50 patients and 47 for the right kidney. The expert's agreement for the left kidney is using 50 patients and 48 patients for the right kidney.

When comparing the agreement between groups, it can be seen that the experts have a much better agreement than the residents for the left and right kidney on the continuous (CCC) and categorical (weighted kappa) scale (p-value <0.001) for the left and right kidney. Additionally, there is a significant difference in the resident's agreement before and after the use of RENEX on the continuous and categorical scale (p-value <0.001). Since the resident's agreement significantly improves after using RENEX there is no significant difference when compared to the experts on the continuous scale (left kidney p-value: 0.505; right kidney p-value: 0.050) and on the categorical scale (left kidney p-value 0.930; right kidney p-value 0.102).

**Difference in CCC between groups**

| Comparison | | Kidney | CCC difference | Bootstrap Variance | 95% Bootstrap Confidence Interval | Test Statistic | p-value |
|---|---|---|---|---|---|---|---|
| Experts | Residents | Left | 0.505 | 0.0135 | (0.264, 0.716) | 4.345 | <0.001 |
| | | Right | 0.519 | 0.0094 | (0.333, 0.722) | 5.355 | <0.001 |
| Experts | Residents with RENEX | Left | 0.093 | 0.0195 | (-0.188, 0.361) | 0.665 | 0.505 |
| | | Right | 0.187 | 0.0091 | (0.029, 0.398) | 1.957 | 0.050 |
| Residents with RENEX | Residents | Left | 0.412 | 0.0079 | (0.224, 0.574) | 4.635 | <0.001 |
| | | Right | 0.332 | 0.0071 | (0.142, 0.481) | 3.947 | <0.001 |

Table 2a. The difference in agreement on the continuous scale (CCC) between each combination of the three groups along with the bootstrap variance of the difference, the 95% bootstrap confidence interval for the difference and the p-value.

**Difference in weighted kappa between groups**

| Groups compared | | Kidney | Weighted Kappa difference | Bootstrap Variance | 95% Bootstrap Confidence Interval | Test Statistic | p-value |
|---|---|---|---|---|---|---|---|
| Experts | Residents | Left | 0.498 | 0.0019 | (0.205, 0.743) | 11.428 | <0.001 |
| | | Right | 0.465 | 0.0160 | (0.205, 0.709) | 3.679 | <0.001 |
| Experts | Residents with RENEX | Left | 0.014 | 0.0248 | (-0.316, 0.299) | 0.087 | 0.930 |
| | | Right | 0.200 | 0.0151 | (-0.038, 0.459) | 1.631 | 0.102 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Residents with RENEX | Residents | Left | 0.484 | 0.0113 | (0.266, 0.681) | 4.556 | <0.001 |
| | | Right | 0.265 | 0.0010 | (0.042, 0.448) | 8.374 | <0.001 |

Table 2b. The difference in agreement on the categorical scale (weighted kappa) between each combination of the three groups along with the bootstrap variance of the difference, the 95% bootstrap confidence interval for the difference and the p-value.

**Replacement analysis to determine resident accuracy**

In instances where the agreement between residents is high, it does not mean that they are correct in their diagnosis. In order to determine if the residents are accurate in their classification we can replace an expert with one of the residents and recalculate the agreement. After replacing the resident with each expert (3 combinations consisting of (resident x, expert 1, expert 2), (resident x, expert 1, expert 3) and (resident x, expert 2, expert 3)), the CCC can be averaged. In table 3a and 3b it can be seen how accurate each resident is by looking at how the agreement changed using CCC and weighted kappa. Each resident had a higher agreement after they used RENEX. The agreement between resident 1 and the experts was 0.634 (0.371, 0.811) but after using RENEX, the agreement increased to 0.715 (0.457, 0.884) for the left kidney and from 0.688 (0.405, 0.812) to 0.728 (0.504, 0.863) for the right kidney. For resident 2, their agreement with the experts was 0.695 (0.457, 0.872) for the left kidney before RENEX and 0.712 (0.460, 0.876) when using RENEX. For the right kidney, resident 2 had an agreement of 0.742 (0.482, 0.885) without RENEX and 0.780 (0.595, 0.889) with RENEX. The agreement between resident 3 and the experts was 0.573 (0.369, 0.737) but after using RENEX, the agreement increased to 0.811 (0.606, 0.922) for the left kidney and from 0.624 (0.437, 0.783) to 0.845 (0.697, 0.919) for the right kidney. Resident 3 had the best agreement with the experts after using RENEX but resident 2 seemed to have the most similar agreement before and after the use of RENEX. A similar pattern of increased agreement

when using RENEX was seen when considering the categorical scale using the weighted kappa.

**Agreement for group consisting of a resident and two experts using CCC**

| Group | Kidney | CCC | Bootstrap Variance | 95% Bootstrap Confidence Interval |
|---|---|---|---|---|
| R1 + two experts | Left | 0.634 | 0.0128 | (0.371, 0.811) |
| R1 + RENEX + two experts | Left | 0.715 | 0.0129 | (0.457, 0.884) |
| R1 + two experts | Right | 0.668 | 0.0107 | (0.405, 0.812) |
| R1 + RENEX + two experts | Right | 0.728 | 0.0085 | (0.504, 0.863) |
| R2 + two experts | Left | 0.695 | 0.0111 | (0.457, 0.872) |
| R2 + RENEX + two experts | Left | 0.712 | 0.0109 | (0.460, 0.876) |
| R2 + two experts | Right | 0.742 | 0.0106 | (0.482, 0.885) |
| R2 + RENEX + two experts | Right | 0.780 | 0.0060 | (0.595, 0.889) |
| R3 + two experts | Left | 0.573 | 0.0090 | (0.369, 0.737) |
| R3 + RENEX + two experts | Left | 0.811 | 0.0069 | (0.606, 0.922) |
| R3 + two experts | Right | 0.624 | 0.0079 | (0.437, 0.783) |
| R3 + RENEX + two experts | Right | 0.845 | 0.0033 | (0.697, 0.919) |

Table 3a. The average CCC when a resident replaces each expert. The combinations used to calculate the CCC are: (resident x, expert 1, expert 2), (resident x, expert 1, expert 3) and (resident x, expert 2, expert 3). The bootstrap variance and bootstrap confidence interval for the averaged CCC are also reported.

**Agreement for group consisting of a resident and two experts using weighted kappa**

| Group | Kidney | Weighted Kappa | Bootstrap Variance | 95% Bootstrap Confidence Interval |
|---|---|---|---|---|
| R1 + two experts | Left | 0.586 | 0.0176 | (0.287, 0.811) |
| R1 + RENEX + two experts | Left | 0.671 | 0.0143 | (0.406, 0.875) |
| R1 + two experts | Right | 0.587 | 0.0187 | (0.282, 0.826) |
| R1 + RENEX + two experts | Right | 0.684 | 0.0167 | (0.370, 0.873) |
| R2 + two experts | Left | 0.647 | 0.0181 | (0.341, 0.867) |
| R2 + RENEX + two experts | Left | 0.662 | 0.0160 | (0.377, 0.871) |
| R2 + two experts | Right | 0.655 | 0.0208 | (0.315, 0.887) |
| R2 + RENEX + two experts | Right | 0.543 | 0.0220 | (0.214, 0.796) |
| R3 + two experts | Left | 0.499 | 0.0109 | (0.292, 0.696) |
| R3 + RENEX + two experts | Left | 0.775 | 0.0112 | (0.529, 0.944) |
| R3 + two experts | Right | 0.523 | 0.0125 | (0.292, 0.734) |
| R3 + RENEX + two experts | Right | 0.761 | 0.0120 | (0.506, 0.926) |

Table 3b. The average weighted kappa when a resident replaces each expert. The combinations used to calculate the weighted kappa are: (resident x, expert 1, expert 2), (resident x, expert 1, expert 3) and (resident x, expert 2, expert 3). The bootstrap variance and bootstrap confidence interval for the average weighted kappa are also reported.

By comparing the agreement of groups that contained two experts and one resident to the agreement between all the experts you can determine which resident was

the most accurate to the experts when classifying the extent of obstruction in the kidneys. Using CCC, when the residents use RENEX their agreement is closer to the experts and are all not significantly different (p-value>0.05). There are some instances where the resident without using RENEX is not significantly difference from the experts but the agreement is still improved while using RENEX. However, this is different when considering the weighted kappa agreement. The only resident that was significantly different from the experts was resident 3 (p-value<0.001, 0.002 for left and right kidney) before the use of RENEX but after using RENEX the difference is not significant (p-value 0.8 for the left and right kidney).

**Difference in agreement between a group of three experts and a group of one resident and two experts using CCC**

| Group being compared to the three experts | Kidney | CCC difference | Bootstrap Variance | 95% Bootstrap Confidence Interval | Test Statistic | p-value |
|---|---|---|---|---|---|---|
| R1 + two experts | Left | 0.185 | 0.0105 | (0.012, 0.424) | 1.805 | 0.071 |
| R1 + RENEX + two experts | Left | 0.104 | 0.0072 | (-0.040, 0.279) | 1.225 | 0.220 |
| R1 + two experts | Right | 0.199 | 0.0067 | (0.069, 0.385) | 2.431 | 0.015 |
| R1 + RENEX + two experts | Right | 0.138 | 0.0072 | (0.004, 0.331) | 1.626 | 0.104 |
| R2 + two experts | Left | 0.124 | 0.0082 | (-0.029, 0.337) | 1.369 | 0.171 |
| R2 + RENEX + two experts | Left | 0.107 | 0.0078 | (-0.042, 0.283) | 1.211 | 0.226 |
| R2 + two experts | Right | 0.125 | 0.0050 | (0.011, 0.295) | 1.768 | 0.077 |
| R2 + RENEX + two experts | Right | 0.086 | 0.0028 | (-0.006, 0.201) | 1.625 | 0.104 |
| R3 + two experts | Left | 0.246 | 0.0053 | (0.099, 0.379) | 3.379 | 0.001 |
| R3 + RENEX + two experts | Left | 0.008 | 0.0036 | (-0.095, 0.136) | 0.133 | 0.894 |
| R3 + two experts | Right | 0.243 | 0.0035 | (0.128, 0.357) | 4.107 | <0.001 |
| R3 + RENEX + two experts | Right | 0.022 | 0.0011 | (-0.038, 0.094) | 0.663 | 0.507 |

Table 4a. The difference between the CCC of the experts and when a resident replaces an expert with and without the use of RENEX. The bootstrap variance and bootstrap confidence interval for the difference are also reported. The p-value indicates whether there is a significant difference between the resident and expert's agreement.

**Difference in agreement between a group of three experts and a group of one resident and two experts using weighted kappa**

| Group being compared to the three experts | Kidney | Weighted Kappa difference | Bootstrap Variance | 95% Bootstrap Confidence Interval | Test Statistic | p-value |
|---|---|---|---|---|---|---|
| R1 + two experts | Left | 0.179 | 0.0135 | (-0.027, 0.424) | 1.541 | 0.123 |
| R1 + RENEX + two experts | Left | 0.095 | 0.0107 | (-0.083, 0.304) | 0.918 | 0.358 |

| | | | | | | |
|---|---|---|---|---|---|---|
| R1 + two experts | Right | 0.187 | 0.0138 | (-0.022, 0.431) | 1.592 | 0.111 |
| R1 + RENEX + two experts | Right | 0.090 | 0.0100 | (-0.076, 0.298) | 0.900 | 0.368 |
| R2 + two experts | Left | 0.118 | 0.0108 | (-0.066, 0.318) | 1.135 | 0.256 |
| R2 + RENEX + two experts | Left | 0.104 | 0.0115 | (-0.086, 0.324) | 0.970 | 0.332 |
| R2 + two experts | Right | 0.119 | 0.0122 | (-0.062, 0.369) | 1.077 | 0.281 |
| R2 + RENEX + two experts | Right | 0.231 | 0.0134 | (0.0003, 0.465) | 1.99 | 0.046 |
| R3 + two experts | Left | 0.266 | 0.0062 | (0.101, 0.407) | 3.378 | <0.001 |
| R3 + RENEX + two experts | Left | -0.009 | 0.0050 | (-0.125, 0.148) | -0.127 | 0.899 |
| R3 + two experts | Right | 0.251 | 0.0069 | (0.071, 0.398) | 3.022 | 0.002 |
| R3 + RENEX + two experts | Right | 0.013 | 0.0041 | (-0.109, 0.146) | 0.203 | 0.839 |

Table 4b. The difference between the weighted of the experts and when a resident replaces an expert with and without the use of RENEX. The bootstrap variance and bootstrap confidence interval for the difference are also reported. The p-value indicates whether there is a significant difference between the resident and expert's agreement.

**Discussion**

Kidney obstruction is classified on a continuous scale from -1 to 1 where a patient's kidney is considered to be unobstructed if it is rated between -1 and -0.2, obstructed between 0.2 and 1 and is undeterminable between -0.2 and 0.2. Classification of kidney obstruction is quite difficult due to the ambiguity of the imaging and lab results and unfortunately there is no gold standard for classification. Therefore, the classification from the experts is the closest method to the gold standard. However, residents are the people that tend to make a majority of the classifications and they tend to have a wider variance between their classification. Luckily there is a software called RENEX that aids the resident in their classification to improve their agreement and accuracy.

The agreement between the residents and experts was calculated through the concordance correlation coefficient using both the variance component and the U-statistic. The continuous classification can be converted onto a categorical scale and the 3 ordinal categories can be used to calculate agreement using the weighted kappa.

Although these are accurate measures for measuring agreement between the raters, there are other methods that have been developed.

Jeong Jang et al. proposed using total deviation index (TDI) and coverage probability (CP) because they are intuitive for interpretation since the value is tied to the original measurement unit and rates can easily determine if there is good agreement by comparing the value of the index to a pre-determined coverage probability. Unfortunately, these methods only work when you are considering agreement between two raters or multiple raters that assume homogeneity of variance – something this study violates. However, they proposed several unscaled indices derived from toot mean square of pairwise differences.

When it comes to classifying kidney obstruction the experts have the highest agreement and smallest variance for the agreement of their classification across all agreement statistics. Comparatively, the residents have poor agreement when it comes classifying kidney obstruction on both the continuous and categorical scale. However, when the residents use RENEX their agreement does increase but the variability in their classification does not seem to approve across all statistics. Although the RENEX system improves the resident's classification of the kidney obstruction it is still not on par with the experts and therefore there is room for improvement.

Latent modeling could be conducted in order to form some sort of gold standard for classifying kidney obstruction by creating and ROC curve to determine the validity of the residents classification. Additionally, a more general approach of using generalized estimating equations (GEE) to model functions of a variety of agreement coefficients (Lin, 2002).

The bootstrap method is one of many simulation tests created to determine statistics from one dataset by creating multiple datasets from resampling. The permutation test is a similar simulation study to the bootstrap but has been seen to have larger power in the case of smaller sample sizes (Troendle, 2004). Comparing the standard error and confidence intervals generated from the bootstrap and permutation methods would be a study to consider. Additionally, this study was conducted using samples sizes of 30 patients resampled 1000 times. A study could be conducting comparing standard errors and confidence intervals from a variety of sample sizes and repetitions to see how they might affect the statistics.

One significant limitation to these estimates is that the RENEX software was developed using an algorithm that was based off the classification of the experts. This would suggest that the resident's classification using the RENEX system are not independent from the classification of the experts. Although the experts are considered the gold standard for classifying kidney obstruction, the software was developed based off of what these experts originally classified and therefore the values using RENEX are influenced by these observations.

**References**

Albert PS. Random effects modeling approaches for estimating roc curves from repeated ordinal tests without a gold standard. Biometrics, 63(2):593–602, 2007.

Banhart HX, Haber M, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. Biometrics. 2002;58(4):1020–1027.

Carrasco JL, Phillips BR, Puig-Martinez J, King TS, Chinchilli VM. Estimation of the concordance correlation coefficient for repated measures using SAS and R. Computer Methods and Program in Biomedicine. 2013;109:293-304.

Jang JH, Manatunga AK, Tayor AT, Long Q. Overall Unscaled Indices for Assessing Agreement Among Multiple Raters. Statistics in Medicine.

Lin LI, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: models, issues and tools. J Am Stat Assoc. 2002;97(457):257–270.

Lin LI, Hedayat AS, Wu W. A unified approach for assessing agreement for continuous and categorical data. J Biopharm Stat. 2007;17(4):629-652.

Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1):255– 268.

Taylor A, Thakore K, Folks R, Halkar R, and Manatunga A. Background subtraction in technetium-99m-mag3 renography. Journal of nuclear medicine: official publication, Society of Nuclear Medicine, 38(1):74–79, 1997.

Taylor A, Garcia E, Manatunga A, Binongo J, Folks R, Salman K, Moncayo V, Plaxton N, Halkar R, and Dubovsky E. Preliminary evaluation of an enhanced decision support system to interpret diuretic renograms. In Society of Nuclear Medicine Annual Meeting Abstracts, volume 53, page 597. Soc Nuclear Med, 2012.

Troendle JF, Korn EL, McShane LM. An Example of Slow Convergence of the Bootstrap

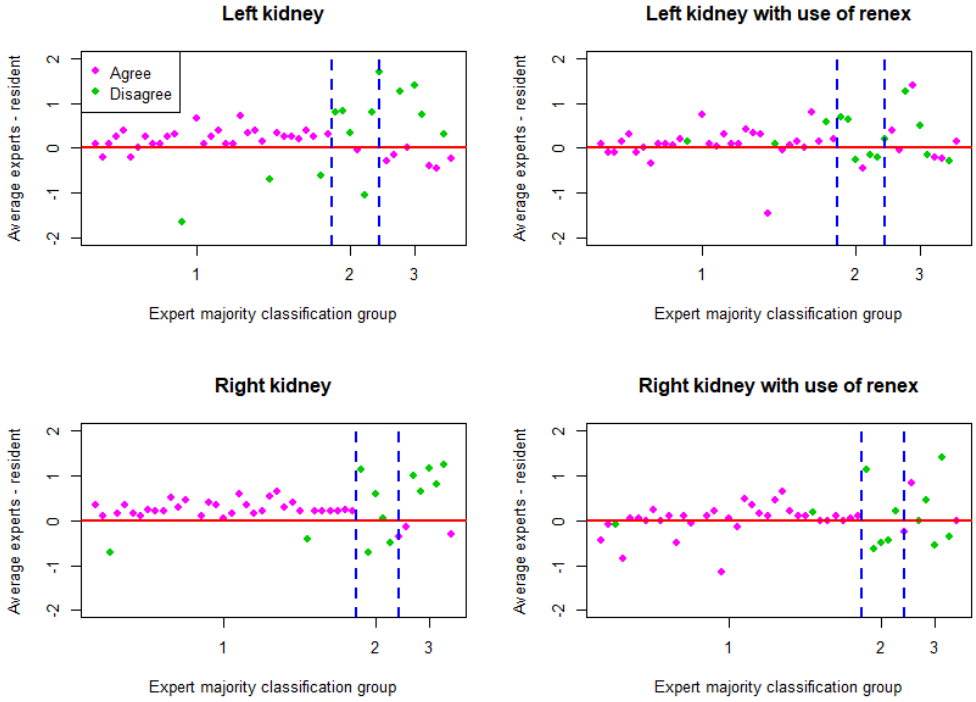
in High Dimensions. The American Statistician. 58(1): 25-29, 2004.

Williamson JM, Manatunga AK, Lipsitz SR. Modeling kappa for measuring dependent categorical agreement data. Biostatistics 2000;1(2):191-202.

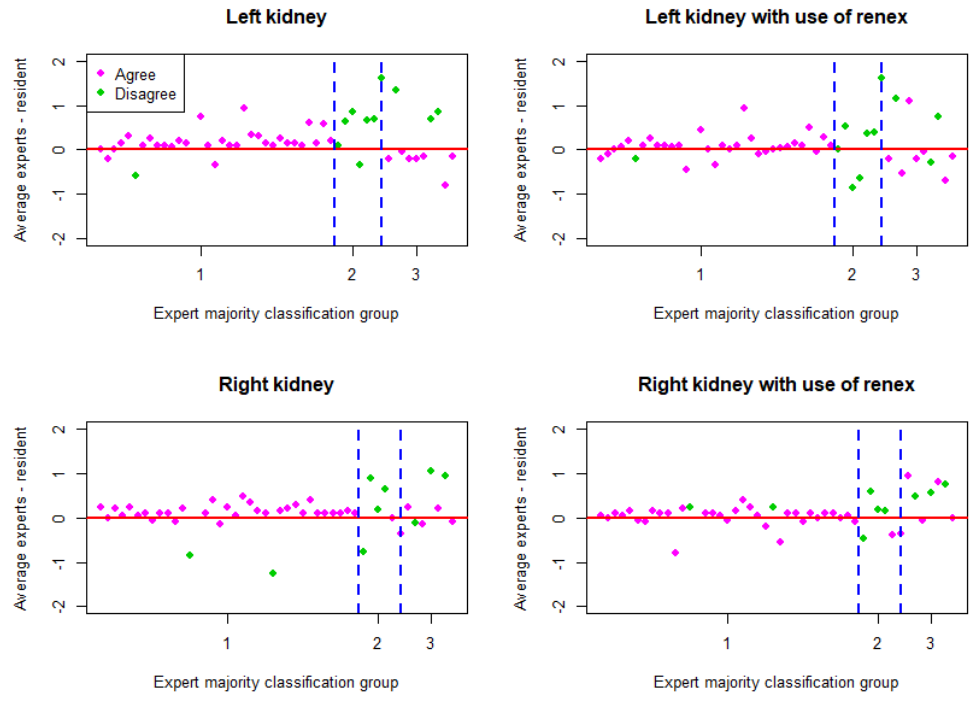
Xu H and Craig BA. A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. Biometrics, 65(4):1145–1155, 2009.

# Appendix

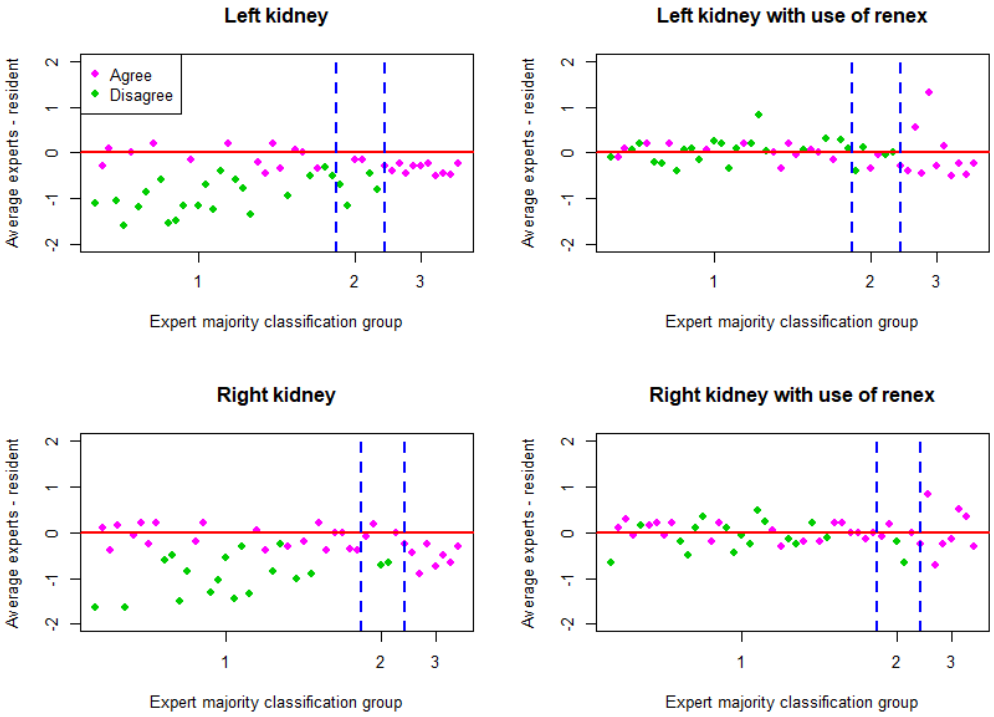
## Difference in classification between the average of expert 1 and 2 and resident 1 vs. expert's majority rule determination of obstruction



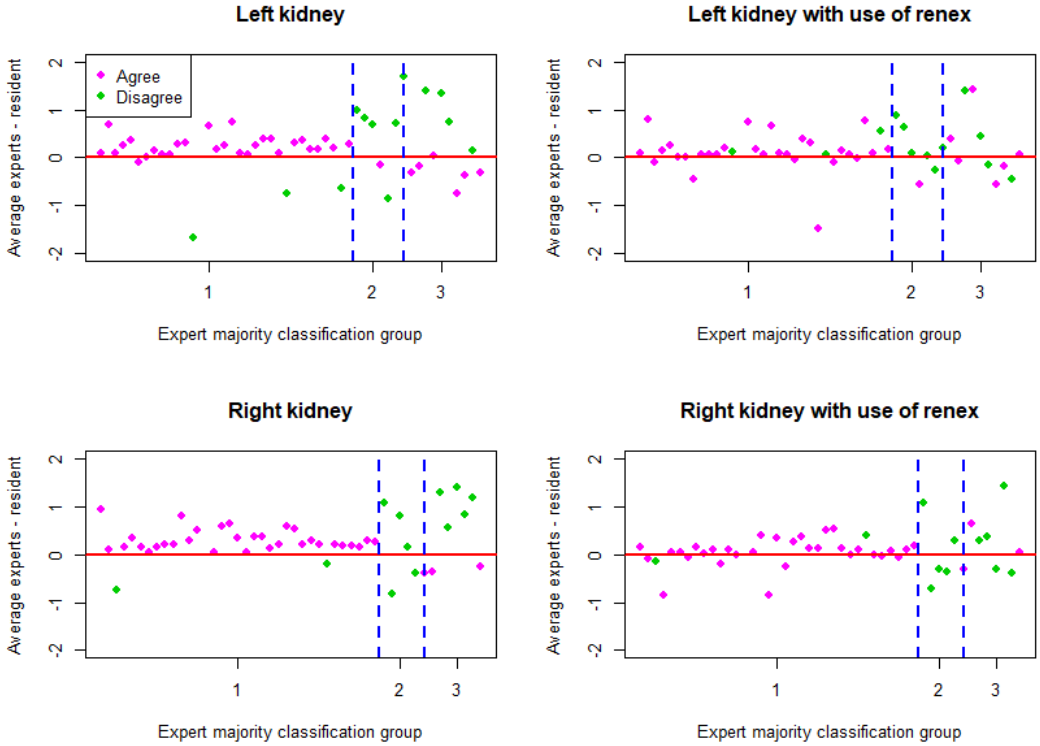

## Difference in classification between the average of expert 1 and 2 and resident 2 vs. expert's majority rule determination of obstruction



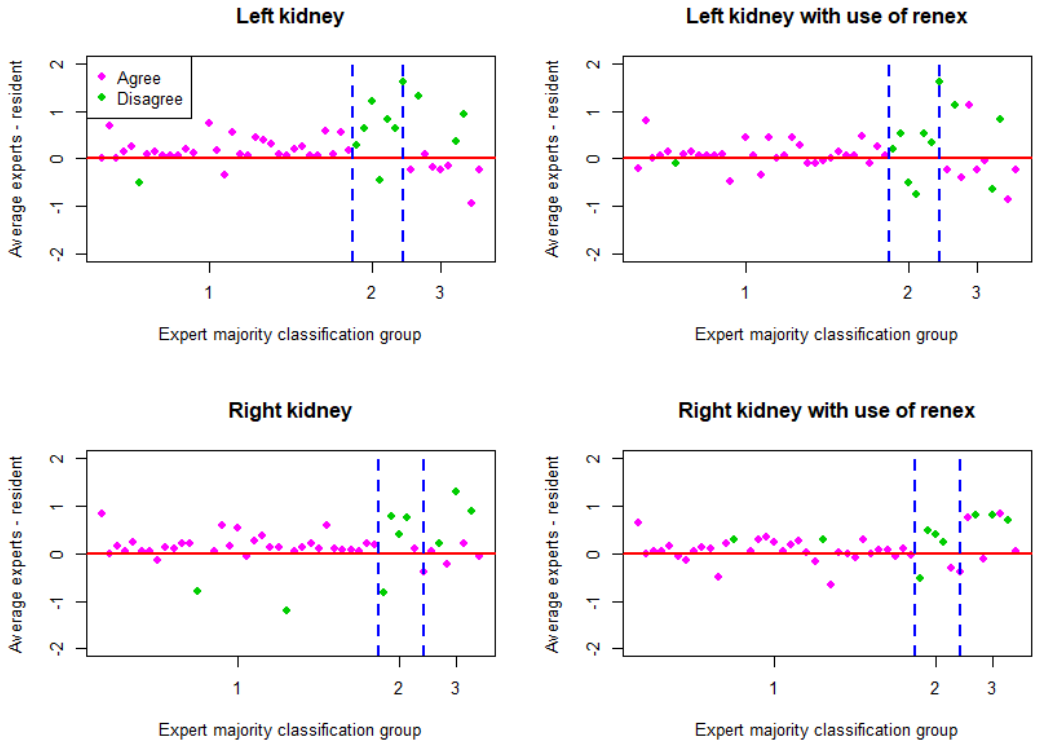
**Difference in classification between the average of expert 1 and 2 and resident 3 vs. expert's majority rule determination of obstruction**



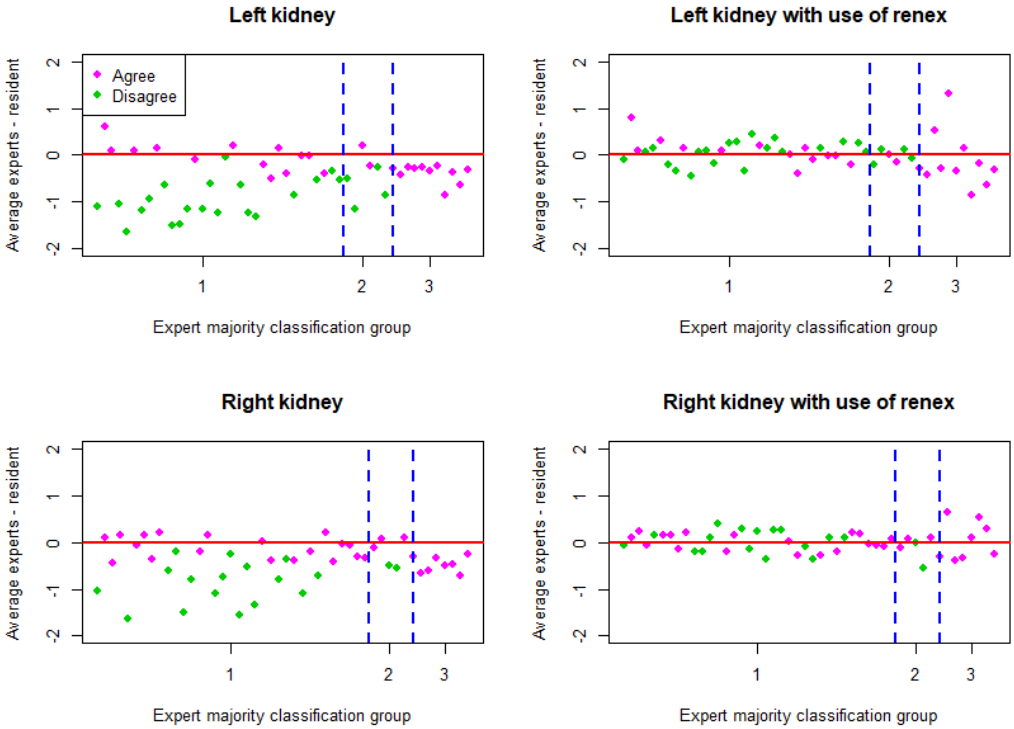

**Difference in classification between the average of expert 1 and 3 and resident 1 vs. expert's majority rule determination of obstruction**



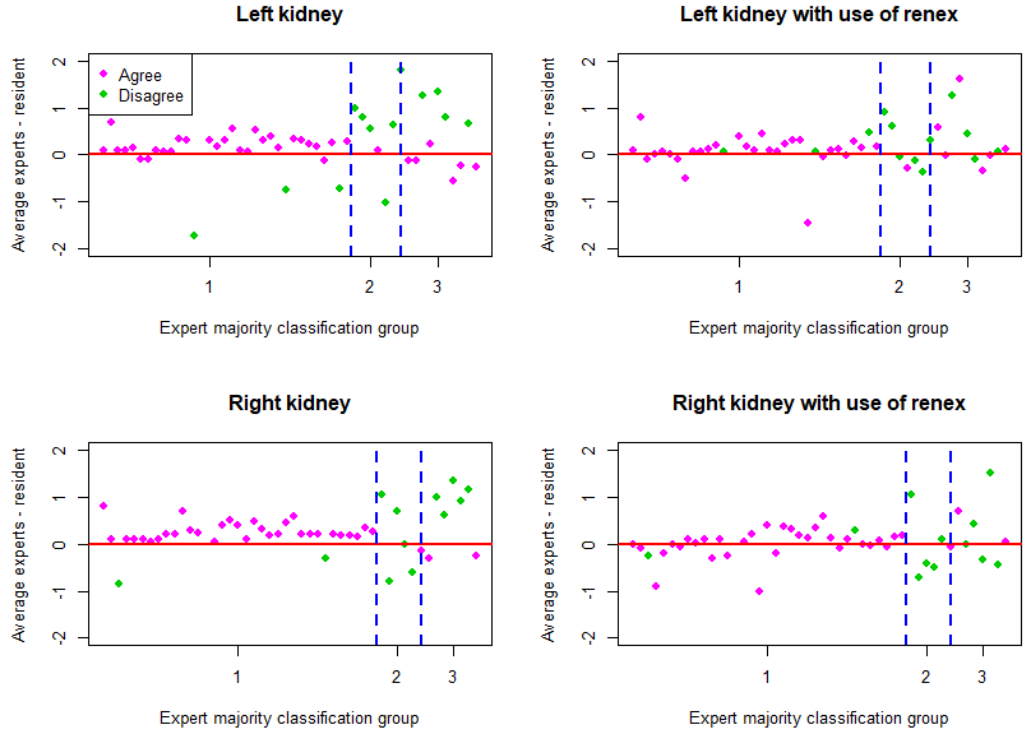
**Difference in classification between the average of expert 1 and 3 and resident 2 vs. expert's majority rule determination of obstruction**



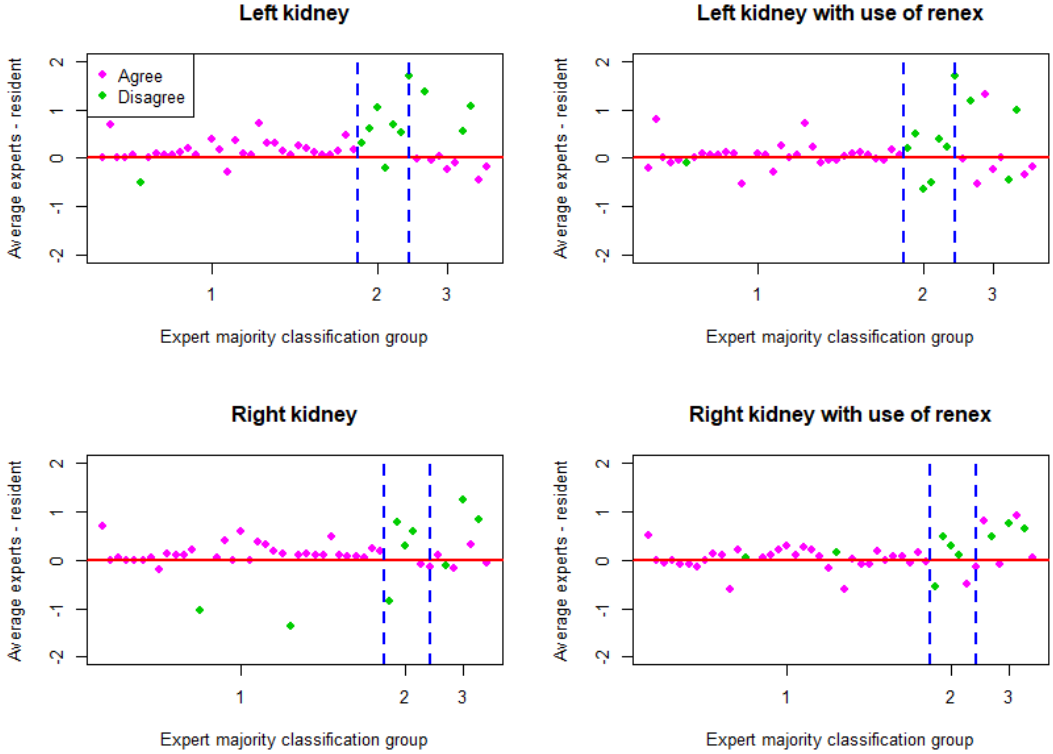

**Difference in classification between the average of expert 1 and 3 and resident 3 vs. expert's majority rule determination of obstruction**



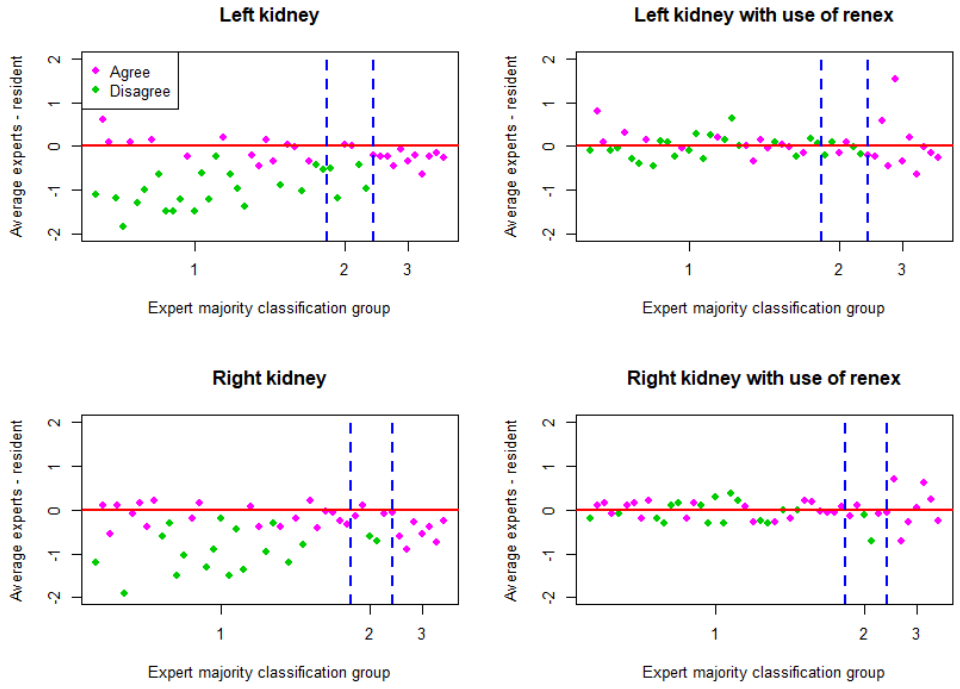
**Difference in classification between the average of expert 2 and 3 and resident 1 vs. expert's majority rule determination of obstruction**




**Difference in classification between the average of expert 2 and 3 and resident 2 vs. expert's majority rule determination of obstruction**



**Difference in classification between the average of expert 2 and 3 and resident 3 vs. expert's majority rule determination of obstruction**




*CCCVC Bootstrap*

The CCCVC function allows for X number of raters and N repeated observations for each patient. A bootstrap function was created that took a 30-patient random sample with replacement from the 50 patients. The CCC between the 3 raters from each group for each kidney was calculated for that sample. This was calculated for 1000 repetitions and each calculation was stored. Once all 1000 CCC values were calculated the variance was determined and a 95% bootstrap confidence interval was calculated by taking the 0.025 and 0.975 percentiles.

**Agreement between raters using CCC**

| Group | Kidney | CCC | Bootstrap Variance | 95% Bootstrap Confidence Interval |
|---|---|---|---|---|
| Residents | Left | 0.292 | 0.085 | (0.163, 0.501) |
| | Right | 0.327 | 0.110 | (0.153, 0.581) |
| Residents with RENEX intervention | Left | 0.732 | 0.082 | (0.573, 0.888) |
| | Right | 0.683 | 0.083 | (0.548, 0.869) |
| Experts | Left | 0.819 | 0.077 | (0.655, 0.945) |
| | Right | 0.867 | 0.052 | (0.748, 0.948) |

Table 1. The CCC calculated using variance components, bootstrap variance and bootstrap confidence interval for the residents, residents using RENEX and experts for the left and right kidney. Due to missing data the CCC for the resident's agreement for the left kidney is calculated using 49 patients and 47 patients for the right kidney. When using RENEX, the agreement for the resident's left kidney is 50 patients and 47 for the right kidney. The expert's agreement for the left kidney is using 50 patients and 48 patients for the right kidney

**Difference in CCC between groups**

| Comparison | | Kidney | CCC difference | Bootstrap Variance | 95% Bootstrap Confidence Interval | Test Statistic | p-value |
|---|---|---|---|---|---|---|---|
| Experts | Residents | Left | 0.527 | 0.0114 | (0.281, 0.706) | 4.934 | <0.001 |
| | | Right | 0.540 | 0.0079 | (0.338, 0.679) | 6.079 | <0.001 |
| Experts | Residents with RENEX | Left | 0.087 | 0.0132 | (-0.169, 0.296) | 0.755 | 0.449 |
| | | Right | 0.184 | 0.0073 | (-0.006, 0.329) | 2.152 | 0.031 |
| Residents with RENEX | Residents | Left | 0.440 | 0.0075 | (0.245, 0.583) | 5.081 | <0.001 |
| | | Right | 0.356 | 0.0074 | (0.178, 0.521) | 4.144 | <0.001 |

Table 2. The difference in agreement on the continuous scale (CCC) between each combination of the three groups along with the bootstrap variance of the difference, the 95% bootstrap confidence interval for the difference and the p-value.

**Agreement for group consisting of a resident and two experts using CCC**

| Group | Kidney | CCC | Bootstrap Variance | 95% Bootstrap Confidence Interval |
|---|---|---|---|---|
| R1 + two experts | Left | 0.652 | 0.0095 | (0.445, 0.827) |
| R1 + RENEX + two experts | Left | 0.742 | 0.0096 | (0.517, 0.902) |
| R1 + two experts | Right | 0.693 | 0.0093 | (0.478, 0.848) |
| R1 + RENEX + two experts | Right | 0.719 | 0.0063 | (0.582, 0.889) |
| R2 + two experts | Left | 0.705 | 0.0096 | (0.496, 0.882) |
| R2 + RENEX + two experts | Left | 0.729 | 0.0073 | (0.547, 0.882) |
| R2 + two experts | Right | 0.769 | 0.0083 | (0.537, 0.902) |
| R2 + RENEX + two experts | Right | 0.806 | 0.0044 | (0.642, 0.908) |
| R3 + two experts | Left | 0.574 | 0.0090 | (0.382, 0.751) |
| R3 + RENEX + two experts | Left | 0.834 | 0.0051 | (0.654, 0.933) |
| R3 + two experts | Right | 0.571 | 0.0102 | (0.392, 0.787) |
| R3 + RENEX + two experts | Right | 0.839 | 0.0026 | (0.732, 0.925) |

Table 3. The average CCC when a resident replaces each expert. The combinations used to calculate the CCC are: (resident x, expert 1, expert 2), (resident x, expert 1, expert 3) and (resident x, expert 2, expert 3). The bootstrap variance and bootstrap confidence interval for the averaged CCC are also reported.

**Difference in agreement between a group of three experts and a group of one resident and two experts using CCC**

| Group | Kidney | CCC difference | Bootstrap Variance | 95% Bootstrap Confidence Interval | Test Statistic | p-value |
|---|---|---|---|---|---|---|
| R1 + two experts | Left | 0.167 | 0.0094 | (0.006, 0.382) | 1.722 | 0.085 |
| R1 + RENEX + two experts | Left | 0.077 | 0.0062 | (-0.041, 0.264) | 0.977 | 0.328 |
| R1 + two experts | Right | 0.174 | 0.0051 | (0.068, 0.346) | 2.436 | 0.015 |

| | | | | | | |
|---|---|---|---|---|---|---|
| R1 + RENEX + two experts | Right | 0.148 | 0.0055 | (-0.021 0.263), | 1.996 | 0.046 |
| R2 + two experts | Left | 0.114 | 0.0067 | (-0.022, 0.296) | 1.393 | 0.164 |
| R2 + RENEX + two experts | Left | 0.090 | 0.0055 | (-0.045, 0.252) | 1.213 | 0.225 |
| R2 + two experts | Right | 0.098 | 0.0045 | (0.016, 0.269) | 1.461 | 0.144 |
| R2 + RENEX + two experts | Right | 0.061 | 0.0022 | (-0.012, 0.179) | 1.300 | 0.193 |
| R3 + two experts | Left | 0.245 | 0.0069 | (0.106, 0.426) | 2.949 | 0.003 |
| R3 + RENEX + two experts | Left | 0.015 | 0.0026 | (-0.090, 0.108) | 0.294 | 0.785 |
| R3 + two experts | Right | 0.296 | 0.0053 | (0.129, 0.405) | 4.066 | <0.001 |
| R3 + RENEX + two experts | Right | 0.028 | 0.0088 | (-0.042, 0.076) | 0.298 | 0.765 |

Table 4. The difference between the CCC of the experts and when a resident replaces an expert with and without the use of RENEX. The bootstrap variance and bootstrap confidence interval for the difference are also reported. The p-value indicates whether there is a significant difference between the resident and expert's agreement.

**Agreement after replacing an expert with a resident using CCC with variance components**

| Combination used to calculate CCC | Kidney | CCC using variance components | 95% Bootstrap Confidence Interval |
|---|---|---|---|
| E12R1 | Left | 0.668 | (0.471, 0.821) |
| E13R1 | Left | 0.617 | (0.401, 0.805) |
| E23R1 | Left | 0.665 | (0.434, 0.834) |
| E12R1 with RENEX | Left | 0.757 | (0.581, 0.899) |
| E13R1 with RENEX | Left | 0.695 | (0.472, 0.876) |
| E23R1 with RENEX | Left | 0.753 | (0.555, 0.914 |
| E12R1 | Right | 0.714 | (0.488, 0.857) |
| E13R1 | Right | 0.662 | (0.466, 0.824) |
| E23R1 | Right | 0.681 | (0.461, 0.841) |
| E12R1 with RENEX | Right | 0.756 | (0.579, 0.886) |
| E13R1 with RENEX | Right | 0.778 | (0.618, 0.890) |
| E23R1 with RENEX | Right | 0.747 | (0.569, 0.884) |
| E12R2 | Left | 0.721 | (0.534, 0.870) |
| E13R2 | Left | 0.678 | (0.475, 0.848) |
| E23R2 | Left | 0.742 | (0.544, 0.899) |
| E12R2 with RENEX | Left | 0.751 | (0.581, 0.878) |
| E13R2 with RENEX | Left | 0.711 | (0.526, 0.867) |
| E23R2 with RENEX | Left | 0.753 | (0.564, 0.901) |

| | | | |
|---|---|---|---|
| E12R2 | Right | 0.779 | (0.589, 0.915) |
| E13R2 | Right | 0.744 | (0.533, 0.899) |
| E23R2 | Right | 0.735 | (0.510, 0.891) |
| | | | |
| E12R2 with RENEX | Right | 0.814 | (0.662, 0.907) |
| E13R2 with RENEX | Right | 0.788 | (0.647, 0.892) |
| E23R2 with RENEX | Right | 0.789 | (0.599, 0.903) |
| | | | |
| E12R3 | Left | 0.596 | (0.423, 0.747) |
| E13R3 | Left | 0.549 | (0.348, 0.719) |
| E23R3 | Left | 0.565 | (0.371, 0.746) |
| | | | |
| E12R3 with RENEX | Left | 0.845 | (0.709, 0.929) |
| E13R3 with RENEX | Left | 0.805 | (0.628, 0.920) |
| E23R3 with RENEX | Left | 0.832 | (0.660, 0.945) |
| | | | |
| E12R3 | Right | 0.588 | (0.387, 0.772) |
| E13R3 | Right | 0.640 | (0.449, 0.813) |
| E23R3 | Right | 0.589 | (0.385, 0.771) |
| | | | |
| E12R3 with RENEX | Right | 0.849 | (0.732, 0.927) |
| E13R3 with RENEX | Right | 0.874 | (0.779, 0.928) |
| E23R3 with RENEX | Right | 0.844 | (0.706, 0.931) |

**Agreement after replacing an expert with a resident using CCC with U-statistics**

| Combination used to calculate CCC | Kidney | CCC using U-statistics | 95% Bootstrap Confidence Interval |
|---|---|---|---|
| E12R1 | Left | 0.635 | (0.419, 0.805) |
| E13R1 | Left | 0.589 | (0.362, 0.794) |
| E23R1 | Left | 0.631 | (0.382, 0.829) |
| | | | |
| E12R1 with RENEX | Left | 0.727 | (0.525, 0.883) |
| E13R1 with RENEX | Left | 0.663 | (0.400, 0.855) |
| E23R1 with RENEX | Left | 0.732 | (0.487, 0.908) |
| | | | |
| E12R1 | Right | 0.683 | (0.445, 0.832) |
| E13R1 | Right | 0.621 | (0.412, 0.796) |
| E23R1 | Right | 0.636 | (0.384, 0.818) |
| | | | |
| E12R1 with RENEX | Right | 0.713 | (0.491, 0.860) |
| E13R1 with RENEX | Right | 0.737 | (0.541, 0.876) |
| E23R1 with RENEX | Right | 0.697 | (0.475, 0.863) |
| | | | |
| E12R2 | Left | 0.691 | (0.484, 0.857) |
| E13R2 | Left | 0.645 | (0.395, 0.838) |
| E23R2 | Left | 0.713 | (0.463, 0.898) |
| | | | |
| E12R2 with RENEX | Left | 0.711 | (0.504, 0.876) |
| E13R2 with RENEX | Left | 0.672 | (0.444, 0.845) |
| E23R2 with RENEX | Left | 0.716 | (0.494, 0.883) |

| | | | |
|---|---|---|---|
| E12R2 | Right | 0.761 | (0.559, 0.897) |
| E13R2 | Right | 0.719 | (0.499, 0.870) |
| E23R2 | Right | 0.694 | (0.452, 0.866) |
| | | | |
| E12R2 with RENEX | Right | 0.789 | (0.636, 0.897) |
| E13R2 with RENEX | Right | 0.756 | (0.600, 0.877) |
| E23R2 with RENEX | Right | 0.755 | (0.569, 0.895) |
| | | | |
| E12R3 | Left | 0.600 | (0.439, 0.750) |
| E13R3 | Left | 0.542 | (0.346, 0.711) |
| E23R3 | Left | 0.554 | (0.358, 0.729) |
| | | | |
| E12R3 with RENEX | Left | 0.823 | (0.664, 0.919) |
| E13R3 with RENEX | Left | 0.772 | (0.561, 0.904) |
| E23R3 with RENEX | Left | 0.809 | (0.603, 0.935) |
| | | | |
| E12R3 | Right | 0.619 | (0.457, 0.779) |
| E13R3 | Right | 0.648 | (0.478, 0.808) |
| E23R3 | Right | 0.595 | (0.410, 0.774) |
| | | | |
| E12R3 with RENEX | Right | 0.829 | (0.711, 0.915) |
| E13R3 with RENEX | Right | 0.851 | (0.746, 0.917) |
| E23R3 with RENEX | Right | 0.822 | (0.682, 0.920) |

**Agreement after replacing an expert with a resident using weighted kappa**

| Combination used to calculate weighted kappa | Kidney | Weighted Kappa | 95% Bootstrap Confidence Interval |
|---|---|---|---|
| E12R1 | Left | 0.590 | (0.310, 0.826) |
| E13R1 | Left | 0.547 | (0.252, 0.786) |
| E23R1 | Left | 0.585 | (0.310, 0.813) |
| | | | |
| E12R1 with RENEX | Left | 0.659 | (0.369, 0.869) |
| E13R1 with RENEX | Left | 0.628 | (0.386, 0.849) |
| E23R1 with RENEX | Left | 0.688 | (0.434, 0.895) |
| | | | |
| E12R1 | Right | 0.617 | (0.311, 0.851) |
| E13R1 | Right | 0.551 | (0.302, 0.785) |
| E23R1 | Right | 0.537 | (0.232, 0.780) |
| | | | |
| E12R1 with RENEX | Right | 0.671 | (0.361, 0.890) |
| E13R1 with RENEX | Right | 0.703 | (0.464, 0.880) |
| E23R1 with RENEX | Right | 0.630 | (0.344, 0.841) |
| | | | |
| E12R2 | Left | 0.628 | (0.385, 0.859) |
| E13R2 | Left | 0.595 | (0.317, 0.845) |
| E23R2 | Left | 0.668 | (0.384, 0.873) |
| | | | |
| E12R2 with RENEX | Left | 0.639 | (0.372, 0.863) |
| E13R2 with RENEX | Left | 0.627 | (0.360, 0.839) |
| E23R2 with RENEX | Left | 0.687 | (0.449, 0.881) |

| | | | |
|---|---|---|---|
| E12R2 | Right | 0.697 | (0.404, 0.908) |
| E13R2 | Right | 0.637 | (0.356, 0.869) |
| E23R2 | Right | 0.580 | (0.236, 0.826) |
| | | | |
| E12R2 with RENEX | Right | 0.563 | (0.265, 0.828) |
| E13R2 with RENEX | Right | 0.508 | (0.207, 0.771) |
| E23R2 with RENEX | Right | 0.514 | (0.215, 0.789) |
| | | | |
| E12R3 | Left | 0.526 | (0.338, 0.717) |
| E13R3 | Left | 0.475 | (0.277, 0.666) |
| E23R3 | Left | 0.505 | (0.301, 0.702) |
| | | | |
| E12R3 with RENEX | Left | 0.761 | (0.533, 0.944) |
| E13R3 with RENEX | Left | 0.746 | (0.522, 0.913) |
| E23R3 with RENEX | Left | 0.792 | (0.583, 0.964) |
| | | | |
| E12R3 | Right | 0.517 | (0.311, 0.712) |
| E13R3 | Right | 0.559 | (0.359, 0.761) |
| E23R3 | Right | 0.475 | (0.262, 0.691) |
| | | | |
| E12R3 with RENEX | Right | 0.738 | (0.509, 0.924) |
| E13R3 with RENEX | Right | 0.796 | (0.588, 0.933) |
| E23R3 with RENEX | Right | 0.701 | (0.423, 0.909) |

**Difference in agreement between a group of three experts and a group of one resident and two experts using CCC**

| Combination used to calculate CCC | Kidney | CCC Difference using variance components | 95% Bootstrap Confidence Interval |
|---|---|---|---|
| E12R1 | Left | 0.167 | (-0.066, 0.377) |
| E13R1 | Left | 0.216 | (0.051, 0.408) |
| E23R1 | Left | 0.174 | (0.004, 0.381) |
| | | | |
| E12R1 with RENEX | Left | 0.076 | (-0.136, 0.290) |
| E13R1 with RENEX | Left | 0.139 | (0.011, 0.303) |
| E23R1 with RENEX | Left | 0.080 | (-0.054, 0.235) |
| | | | |
| E12R1 | Right | 0.159 | (0.015, 0.338) |
| E13R1 | Right | 0.217 | (0.097, 0.384) |
| E23R1 | Right | 0.193 | (0.069, 0.373) |
| | | | |
| E12R1 with RENEX | Right | 0.121 | (-0.016, 0.295) |
| E13R1 with RENEX | Right | 0.098 | (-0.027, 0.242) |
| E23R1 with RENEX | Right | 0.127 | (0.005, 0.289) |
| | | | |
| E12R2 | Left | 0.106 | (-0.083, 0.289) |
| E13R2 | Left | 0.152 | (0.0149, 0.332) |
| E23R2 | Left | 0.089 | (-0.0572, 0.249) |
| | | | |
| E12R2 with RENEX | Left | 0.085 | (-0.119, 0.261) |
| E13R2 with RENEX | Left | 0.126 | (0.005, 0.275) |

| | | | |
|---|---|---|---|
| E23R2 with RENEX | Left | 0.082 | (-0.065, 0.239) |
| | | | |
| E12R2 | Right | 0.094 | (-0.042, 0.267) |
| E13R2 | Right | 0.127 | (0.021, 0.267) |
| E23R2 | Right | 0.140 | (0.018, 0.308) |
| | | | |
| E12R2 with RENEX | Right | 0.057 | (-0.056, 0.176) |
| E13R2 with RENEX | Right | 0.088 | (0.006, 0.177) |
| E23R2 with RENEX | Right | 0.084 | (-0.004, 0.198) |
| | | | |
| E12R3 | Left | 0.240 | (0.042, 0.419) |
| E13R3 | Left | 0.282 | (0.134, 0.439) |
| E23R3 | Left | 0.270 | (0.108, 0.453) |
| | | | |
| E12R3 with RENEX | Left | -0.013 | (-0.179, 0.118) |
| E13R3 with RENEX | Left | 0.035 | (-0.036, 0.129) |
| E23R3 with RENEX | Left | 0.004 | (-0.100, 0.107) |
| | | | |
| E12R3 | Right | 0.279 | (0.141, 0.419) |
| E13R3 | Right | 0.227 | (0.095, 0.373) |
| E23R3 | Right | 0.285 | (0.135, 0.431) |
| | | | |
| E12R3 with RENEX | Right | 0.026 | (-0.049, 0.096) |
| E13R3 with RENEX | Right | 0.003 | (-0.067, 0.061) |
| E23R3 with RENEX | Right | 0.028 | (-0.029, 0.096) |

**Difference in agreement between a group of three experts and a group of one resident and two experts using CCC**

| Combination used to calculate CCC | Kidney | CCC difference using U-statistics | 95% Bootstrap Confidence Interval |
|---|---|---|---|
| E12R1 | Left | 0.171 | (-0.084, 0.390) |
| E13R1 | Left | 0.220 | (0.056, 0.442) |
| E23R1 | Left | 0.179 | (-0.004, 0.420) |
| | | | |
| E12R1 with RENEX | Left | 0.082 | (-0.135, 0.323) |
| E13R1 with RENEX | Left | 0.157 | (0.015, 0.345) |
| E23R1 with RENEX | Left | 0.087 | (-0.063, 0.283) |
| | | | |
| E12R1 | Right | 0.172 | (0.016, 0.364) |
| E13R1 | Right | 0.229 | (0.101, 0.396) |
| E23R1 | Right | 0.219 | (0.069, 0.418) |
| | | | |
| E12R1 with RENEX | Right | 0.145 | (-0.009, 0.343) |
| E13R1 with RENEX | Right | 0.116 | (-0.023, 0.289) |
| E23R1 with RENEX | Right | 0.155 | (0.015, 0.349) |
| | | | |
| E12R2 | Left | 0.119 | (-0.108, 0.338) |
| E13R2 | Left | 0.165 | (0.033, 0.359) |
| E23R2 | Left | 0.101 | (-0.074, 0.309) |

| | | | |
|---|---|---|---|
| E12R2 with RENEX | Left | 0.098 | (-0.128, 0.311) |
| E13R2 with RENEX | Left | 0.141 | (0.006, 0.315) |
| E23R2 with RENEX | Left | 0.094 | (-0.081, 0.305) |
| | | | |
| E12R2 | Right | 0.096 | (-0.054, 0.251) |
| E13R2 | Right | 0.135 | (0.019, 0.276) |
| E23R2 | Right | 0.155 | (0.024, 0.321) |
| | | | |
| E12R2 with RENEX | Right | 0.070 | (-0.067, 0.196) |
| E13R2 with RENEX | Right | 0.098 | (0.012, 0.206) |
| E23R2 with RENEX | Right | 0.100 | (0.002, 0.227) |
| | | | |
| E12R3 | Left | 0.210 | (-0.009, 0.374) |
| E13R3 | Left | 0.268 | (0.146, 0.404) |
| E23R3 | Left | 0.255 | (0.108, 0.386) |
| | | | |
| E12R3 with RENEX | Left | -0.009 | (-0.210, 0.139) |
| E13R3 with RENEX | Left | 0.036 | (-0.036, 0.159) |
| E23R3 with RENEX | Left | 0.002 | (-0.117, 0.137) |
| | | | |
| E12R3 | Right | 0.238 | (0.126, 0.358) |
| E13R3 | Right | 0.206 | (0.080, 0.325) |
| E23R3 | Right | 0.258 | (0.133, 0.367) |
| | | | |
| E12R3 with RENEX | Right | 0.027 | (-0.052, 0.099) |
| E13R3 with RENEX | Right | 0.004 | (-0.083, 0.073) |
| E23R3 with RENEX | Right | 0.037 | (-0.021, 0.109) |

**Difference in agreement between a group of three experts and a group of one resident and two experts using kappa**

| Combination used to calculate weighted kappa | Kidney | Weighted Kappa Difference | 95% Bootstrap Confidence Interval |
|---|---|---|---|
| E12R1 | Left | 0.170 | (-0.064, 0.425) |
| E13R1 | Left | 0.199 | (-0.022, 0.448) |
| E23R1 | Left | 0.171 | (-0.075, 0.432) |
| | | | |
| E12R1 with RENEX | Left | 0.094 | (-0.112, 0.323) |
| E13R1 with RENEX | Left | 0.130 | (-0.073, 0.336) |
| E23R1 with RENEX | Left | 0.068 | (-0.178, 0.295) |
| | | | |
| E12R1 | Right | 0.142 | (-0.131, 0.428) |
| E13R1 | Right | 0.205 | (0.009, 0.447) |
| E23R1 | Right | 0.217 | (0.027, 0.454) |
| | | | |
| E12R1 with RENEX | Right | 0.092 | (-0.117, 0.337) |
| E13R1 with RENEX | Right | 0.050 | (-0.153, 0.262) |
| E23R1 with RENEX | Right | 0.125 | (-0.031, 0.322) |
| | | | |
| E12R2 | Left | 0.129 | (-0.071, 0.346) |
| E13R2 | Left | 0.155 | (-0.042, 0.3760 |

| | | | |
|---|---|---|---|
| E23R2 | Left | 0.080 | (-0.171, 0.339) |
| | | | |
| E12R2 with RENEX | Left | 0.129 | (-0.077, 0.366) |
| E13R2 with RENEX | Left | 0.126 | (-0.091, 0.349) |
| E23R2 with RENEX | Left | 0.077 | (-0.172, 0.316) |
| | | | |
| E12R2 | Right | 0.061 | (-0.177, 0.310) |
| E13R2 | Right | 0.131 | (-0.075, 0.393) |
| E23R2 | Right | 0.182 | (0.006, 0.430) |
| | | | |
| E12R2 with RENEX | Right | 0.192 | (-0.053, 0.457) |
| E13R2 with RENEX | Right | 0.255 | (0.055, 0.485) |
| E23R2 with RENEX | Right | 0.251 | (0.047, 0.478) |
| | | | |
| E12R3 | Left | 0.225 | (0.029, 0.415) |
| E13R3 | Left | 0.284 | (0.134, 0.440) |
| E23R3 | Left | 0.250 | (0.070, 0.426) |
| | | | |
| E12R3 with RENEX | Left | -0.002 | (-0.158, 0.171) |
| E13R3 with RENEX | Left | 0.007 | (-0.141, 0.164) |
| E23R3 with RENEX | Left | -0.034 | (-0.244, 0.143) |
| | | | |
| E12R3 | Right | 0.236 | (0.078, 0.408) |
| E13R3 | Right | 0.197 | (-0.009, 0.385) |
| E23R3 | Right | 0.278 | (0.132, 0.419) |
| | | | |
| E12R3 with RENEX | Right | 0.023 | (-0.117, 0.174) |
| E13R3 with RENEX | Right | -0.039 | (-0.251, 0.113) |
| E23R3 with RENEX | Right | 0.051 | (-0.052, 0.179) |