Published in final edited form as:

Stat Med. 2011 June 30; 30(14): 1712–1721. doi:10.1002/sim.4203.

Combining multiple continuous tests for the diagnosis of kidney impairment in the absence of a gold standard

Binbing Yu^{1,*†}, Chuan Zhou², and Stefania Bandinelli³

¹Laboratory of Epidemiology, Demography and Biometry, National Institute on Aging, Bethesda, Maryland 20892, U.S.A.

²Department of Pediatrics, University of Washington Medical Center, Center for Child Health, Behavior and Development, Seattle Children's Hospital and Regional Medical Center, Seattle, WA 98195, U.S.A.

³Geriatric unit, Azienda, Sanitaria di Firenze, Italy

SUMMARY

Receiver Operating Characteristic (ROC) curves are commonly used to summarize the classification accuracy of diagnostic tests. It is not uncommon in medical practice that multiple diagnostic tests are routinely performed or multiple disease markers are available for the same individuals. When the true disease status is verified by a gold standard test, a variety of methods have been proposed to combine such potential correlated tests to increase the accuracy of disease diagnosis. In this article, we propose a method of combining multiple diagnostic tests in the absence of a gold standard. We assume that the test values and their classification accuracies are dependent on covariates. Simulation studies are performed to examine the performance of the combination method. The proposed method is applied to data from a population-based aging study to compare the accuracy of three screening tests for kidney function and to estimate the prevalence of moderate kidney impairment.

Keywords

Diagnostic test; Gold standard; Markov chain Monte-Carlo; Sensitivity; Specificity

1. INTRODUCTION

Chronic kidney disease (CKD) is a growing problem in the United States. Between 1990 and 2000, the number of people with kidney failure requiring dialysis or transplantation virtually doubled to 380,000. The annual cost of treating kidney failure in the United States has already topped 20 billion dollars. Experts estimate that 20 million Americans have significantly reduced kidney function, and even a small loss of kidney function can double a person's risk of developing cardiovascular disease. Many of these people will experience heart attacks or strokes before they become aware of their kidney disease. So identifying and treating CKD early can help prevent heart problems as well as postpone kidney failure.

Accurate assessment of kidney function in clinical practice plays a crucial role in screening for kidney disease, adjusting treatment plan, and following the evolution of known kidney disease. The glomerular filtration rate (GFR) is an excellent measure of the filtering capacity of the kidneys. A low or decreasing GFR is a good index of CKD. According to American Kidney Foundation's Kidney Disease Outcomes and Quality Initiative (K/DOQI), CKD is defined as kidney damage or $\text{GFR} < 60 \text{ mL/min/1.73m}^2$ for three months or more, irrespective of cause [1]. However, concerns have been expressed about the definition and classification, methods to estimate GFR and ascertainment of proteinuria [2]. Inulin clearance is widely regarded as the gold standard (GS) for measuring glomerular filtration rate. The classic method of inulin clearance requires an intravenous infusion and timed urine collections over a period of several hours, making it costly and cumbersome. As a result a number of alternative measures for estimating GFR have been devised. This motivates us to find an optimal measure of GFR, which may lead to a more accurate diagnosis of kidney impairment, given multiple available measures.

For adults, the best measure for estimating GFR is the Modification of Diet in Renal Disease (MDRD) Study formula, which is calculated using age, sex and serum creatinine [3]. The MDRD formula is recommended for routine use and is reported with all requests for serum creatinine. However, the diagnostic accuracy of the MDRD formula for the older population has not been validated [4]. Two other GFR estimation methods are also used for the older people. The Cockcroft-Gault (CG) formula is recommended for drug dosing decisions and the formula requires patient age, sex, weight or height, and serum creatinine. Measurement of 24-hour creatinine clearance (CC), which requires a serum creatinine measurement and a 24 hour urine collection, is the least reliable method in general but is valuable in extremes of body composition. For the older population, several recent studies [1, 5] show that the CC tends to overestimate the true GFR whereas the CG formula generally underestimates the true GFR, and the MDRD formula could both over and under estimate the true GFR. Because all three formulae are not perfect and error prone, the diagnosis, and thus the estimation of prevalence, of kidney impairment based on these methods are inaccurate as well. It is desirable to combine the three GFR formulae to improve the accuracy of diagnosis of kidney impairment. As the derivation of the three test values are dependent on age, sex, weight and other factors, the weights of the optimal combinations should depend on covariates as well.

The methods of combining multiple diagnostic markers (tests) are extensive but most articles only describe situations where a GS test is available. Choi et al. [6] and Wang et al. [7] presented the methods of estimating and comparing the ROC curves for several competitive diagnostic tests in the absence of a GS. They only considered the estimation of the ROC curve for each single test. As a continuation, we present a Bayesian model for combining multiple tests in the absence of a GS. The parameters are estimated using the Markov chain Monte-Carlo (MCMC) method. There are several advantages of the proposed method. First, by modeling covariate-dependent ROC, an individual-specific cutoff point for kidney impairment can be determined with fixed sensitivity and specificity. Second, a more accurate estimate of prevalence can be obtained by combining multiple tests. Third, the accuracies of the multiple tests and their combinations can be evaluated and compared easily using the posterior distribution based on the MCMC samples. Simulations are carried out to examine the performance of the proposed method of modeling multiple markers. The method is then applied to data from the InCHIANTI Study to estimate the prevalence of kidney impairment among the older population.

The rest of the paper is organized as follows. The Bayesian model for multiple correlated diagnostic tests is introduced and the method of combining multiple markers is described in Section 2. The simulation results are summarized in Section 3 and analysis of the study data

is presented in Section 4. We then conclude with discussions and future research in Section 5.

2. STATISTICAL METHOD

Here we describe the statistical model for the multiple correlated diagnostic tests without a GS and give the estimates of optimal weights of combining the tests. Because the GS test is not available, the actual disease status is latent. Johnson et al. [8] showed that, for two binary tests without gold standard, using two populations with different prevalences can circumvent the problem of non-identifiability between the parameters of test accuracy. Assuming covariate-dependent disease prevalence in the absence of GS thus makes the parameters identifiable. Furthermore, this also makes biological sense because the probability of having kidney impairment increases with age and comorbidity conditions. Thus, we assume that the disease prevalence depends on covariates.

Let D be the unknown binary indicator of disease, where $D = 1$ means diseased and $D = 0$ means non-diseased. In the absence of a GS, the true disease status is not observed. Let Y_k be the value of the k -th diagnostic test, $k = 1, \dots, K$ and let $\mathbf{Y} = (Y_1, \dots, Y_K)'$ denote the vector of multiple correlated tests. The covariate vector of an individual is denoted by $\mathbf{x}$. We assume that the probability (prevalence) of being healthy (non-diseased) follows a logistic model

$$P(D=0)=\pi=\frac{\exp(\gamma^T \mathbf{x})}{1+\exp(\gamma^T \mathbf{x})}, \quad (1)$$

where γ is the row vector of coefficients. If the latent disease status were observed, the test values, after suitable transformations, follow a multivariate normal (MVN) distribution:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_K \end{pmatrix} \sim \text{MVN}(\mu(\mathbf{x}, D), \Sigma_D), \quad (2)$$

where

$$\mu(\mathbf{x}, D) = \begin{pmatrix} \mu_1(\mathbf{x}, D) \\ \vdots \\ \mu_K(\mathbf{x}, D) \end{pmatrix} \text{ and } \Sigma_D = \begin{pmatrix} \sigma_{11D}^2 & \cdots & \sigma_{1KD} \\ \vdots & \ddots & \vdots \\ \sigma_{K1D} & \cdots & \sigma_{KKD}^2 \end{pmatrix}.$$

The mean of the k th test value is

$$\mu_k(\mathbf{x}, D) = \beta_0^{(k)} + \beta_x^{(k)} \mathbf{x} + \beta_D^{(k)} D + \beta_{xD}^{(k)} (\mathbf{x}D),$$

where $\beta^{(k)} = (\beta_0^{(k)}, \beta_x^{(k)}, \beta_D^{(k)}, \beta_{xD}^{(k)})$ is the corresponding row vector of regression coefficients. Here, we assume that the covariance matrix of the test scores, denoted by Σ_D , is only related to disease status. In general, it can be related to covariates as well. Let $\boldsymbol{\theta} = (\gamma, \beta^{(1)}, \dots, \beta^{(K)}, \Sigma_0, \Sigma_1)$ be the collection of all parameters to be estimated.

The ROC curve is a plot of sensitivity versus 1-specificity for all possible cutoff values. For individuals with covariates $\mathbf{x}$, the area under the ROC curve (AUC) for the k th test is

$$AUC_k(\mathbf{x}) = \Phi \left(\frac{\mu_k(\mathbf{x}, 1) - \mu_k(\mathbf{x}, 0)}{\sqrt{\sigma_{kk0}^2 + \sigma_{kk1}^2}} \right).$$

If the true disease statuses of the individuals were observed, let $\mathcal{C} = \{(\mathbf{x}_i, D_i, \mathbf{Y}_i), i = 1, \dots, n\}$ be the complete data, which include covariates, latent disease status and observed test values for n individuals. Let $\pi_i = P(D_i = 0)$ be the probability of being health of the i th individual. The likelihood function based on the complete data $\mathcal{C}$ is

$$L(\theta|\mathcal{C}) = \prod_{i=1}^n \left[(1 - \pi_i) \phi_K(\mathbf{Y}_i | \mu(\mathbf{x}_i, 1), \Sigma_1) \right]^{D_i} \left[\pi_i \phi_K(\mathbf{Y}_i | \mu(\mathbf{x}_i, 0), \Sigma_0) \right]^{1-D_i}, \quad (3)$$

where $\phi_K(\mathbf{Y}|\mu, \Sigma)$ is the K -dimension MVN density function with mean μ and variance Σ . The joint model of $(\mathbf{x}_i, D_i, \mathbf{Y}_i)$ can be written in the following hierarchical form

$$\begin{aligned} D_i | \mathbf{x}_i &\sim \text{Bernoulli}(1 - \pi_i), \\ \mathbf{Y}_i | \mathbf{x}_i, D_i &\sim \text{MVN}(\mu(\mathbf{x}_i, D_i), \Sigma_{D_i}). \end{aligned}$$

The observed test values $\mathbf{Y}_i$ actually follows a finite mixture distribution [9, p421]. Because the latent disease status $\{D_i, i = 1, \dots, n\}$ are not observed, one can use the MCMC method with Gibbs sampling to find the parameter estimates. Gibbs sampling [9] works appropriately for high-dimensional complex models based on conditional independence assumptions. We assume the elements of θ are independent of each other. The prior for each element is specified as follows:

$$\gamma \sim \text{MVN}(0, \mathbf{I}\sigma_\gamma^2), \beta^{(k)} \sim \text{MVN}(0, \mathbf{I}\sigma_\beta^2), k=1, \dots, K, \Sigma_D \sim \text{Wishart}(v, \Gamma),$$

where $\mathbf{I}$ is the identity matrix and $\text{Wishart}(v, \Gamma)$ is a Wishart distribution with v degrees of freedom and scale matrix Γ . The hyperparameters are chosen such that the priors of the parameters tend to be weakly informative. For example, we set $\sigma_\gamma^2 = \sigma_\beta^2 = 1000$, which implies mean 0 and variance 1000 for the regression coefficients $\gamma, \beta^{(1)}, \dots, \beta^{(K)}$. The initial variables for the parameters are selected from the range of prior mean ± 2 standard deviations. In order to examine the convergence of the MCMC samples, usually multiple independent chains should be run. Various methods of monitoring convergence have been proposed by Geweke et al. [10], Gelman and Rubin [11] and Heidelberger et al. [12]. The MCMC method is now implemented in publicly available software WinBUGS [13] and its open source OpenBUGS. Here, we use the R package **BRugs** [14], which calls OpenBUGS, to run MCMC and use R package coda to assess the convergence. For each MCMC run, initial burn-in samples are discarded and the posterior distributions of the parameters are based on additional MCMC samples.

After obtaining the parameter estimates using the MCMC method, multiple tests can be combined using a linear combination of the tests, where the composite test $Y^* = a'\mathbf{Y}$. Let $\Delta(\mathbf{x}) = \mu(\mathbf{x}, 1) - \mu(\mathbf{x}, 0)$. Su and Liu [15] showed that, the optimal vector of linear combination is

$$a=(\Sigma_0+\Sigma_1)^{-1}\Delta(\mathbf{x}) \quad (4)$$

and the corresponding $AUC(x)=\Phi\left(\sqrt{a'\Delta(\mathbf{x})}\right)$, which is the Mahalois distance between healthy and diseased populations. In addition, one can compare the AUCs of two tests at covariate level $\mathbf{x}$ by examining the difference $AUC_k(\mathbf{x}) - AUC_{k'}(\mathbf{x})$ and its CI.

The corresponding sensitivity (SE) and specificity (SP) for a given threshold value c and covariates $\mathbf{x}$ are

$$SE(c|\mathbf{x})=\Phi\left(\frac{a'\mu(\mathbf{x},1)-c}{\sqrt{a'\Sigma_1a}}\right) \text{ and } SP(c|\mathbf{x})=\Phi\left(\frac{c-a'\mu(\mathbf{x},0)}{\sqrt{a'\Sigma_0a}}\right).$$

The conditional ROC is calculated as

$$ROC(t|\mathbf{x})=\Phi\left(\frac{a'\Delta(\mathbf{x})+\Phi^{-1}(t)\sqrt{a'\Sigma_0a}}{\sqrt{a'\Sigma_1a}}\right), \text{ for } t \in (0,1). \quad (5)$$

Schisterman et al. [16] proposed a method of calculating covariate-dependent combination of markers. Their approach does not allow clustered observations and the CI of the AUC is calculated based on an approximation used in Behrens-Fisher problem. The Bayesian method can facilitate the computation and the CI of the AUC can be generated as a byproduct from the MCMC samples.

3. SIMULATION STUDY

We conduct a limited simulation to examine the performance of the proposed method. We consider two tests and the test scores $\mathbf{Y} = (Y_1, Y_2)^T$ follow a bivariate normal distribution as shown in Equation (2). For non-diseased subjects, $\mathbf{Y} \sim MVN(\mu_{\bar{D}}, \Sigma_{\bar{D}})$ and for diseased subjects, $\mathbf{Y} \sim MVN(\mu_D, \Sigma_D)$ and

$$\mu_{\bar{D}} = \begin{pmatrix} \mu_{1\bar{D}} \\ \mu_{2\bar{D}} \end{pmatrix}, \Sigma_{\bar{D}} = \begin{pmatrix} \sigma_{1\bar{D}}^2 & \sigma_{12\bar{D}} \\ \sigma_{12\bar{D}} & \sigma_{2\bar{D}}^2 \end{pmatrix}, \mu_D = \begin{pmatrix} \mu_{1D} \\ \mu_{2D} \end{pmatrix}, \Sigma_D = \begin{pmatrix} \sigma_{1D}^2 & \sigma_{12D} \\ \sigma_{12D} & \sigma_{2D}^2 \end{pmatrix}.$$

The parameters $(\rho_{\bar{D}}, \rho_D, \mu_{1\bar{D}}, \mu_{2\bar{D}})$ take values as shown in Table I. The other parameters are fixed as $\mu_{1\bar{D}}=\mu_{2\bar{D}}=0$, $\sigma_{1\bar{D}}^2=\sigma_{2\bar{D}}^2=1$ and $\sigma_{1D}^2=\sigma_{2D}^2=2$. Given the variances and the correlation coefficients in Table I, the covariances of the two tests are $\sigma_{12\bar{D}} = \rho_{\bar{D}}$ for the healthy population and $\sigma_{12D} = 2\rho_D$ for the diseased population, respectively.

We assume the disease prevalence is related to a binary covariate x ,

$$\text{logit}[P(D=1)] = \gamma_0 + \gamma_1 x,$$

where $\gamma_0 = 0$ and $\gamma_1 = 1$. Thus the disease prevalences are 50% for $x = 0$ and 73% for $x = 1$. The purpose of selecting the prevalence parameters is to investigate the accuracy of the underlying prevalence in the application.

We simulate $m = 500$ samples, each sample consists of 1000 individuals and half of them take value $x = 0$. The latent disease status is generated from the logistic model and the test values $Y = (Y_1, Y_2)$ are generated from a bivariate normal distribution using function `rmvnorm` in the R package `mvtnorm`. For each simulated dataset, we obtain 5000 MCMC samples for inference after a 1000 burn-in. The estimate of each parameter θ , denoted by $\hat{\theta}$, is obtained as the posterior mean of the 5000 MCMC samples. The standard error $S_{\hat{\theta}}$, and the 95% credible intervals (CIs) are also calculated based on the MCMC sample. Several statistics are calculated to assess the performance of the proposed model by summarizing the 500 simulated datasets. The bias is calculated as $\sum_i (\hat{\theta}_i - \theta)/m$ and the actual coverage rate (CR) is calculated as proportion of $\hat{\theta}$ belonging to the 95% CI. We also compare the mean standard error (MSE) and the empirical standard error (ESE). The MSE is the average of $S_{\hat{\theta}}$

and the ESE is $\sqrt{\sum_i (\hat{\theta}_i - \bar{\hat{\theta}})^2/m}$. The summary of bias, MSE, ESE and CR are shown in Table II.

To examine the performance for diseases with lower prevalences, we repeat the simulation with the same parameters except for the prevalence parameters $\gamma_0 = -1$ and $\gamma_1 = 1$. The results are essentially the same. Based on the simulations, we see that all the parameter estimates are empirically unbiased. The values of MSE and ESE are very close to each other for all cases and all parameters. The actual coverage rates of the CIs are reasonably close to the nominal level of 95%. There are almost no differences between Cases 1–3, indicating the small shift in location parameters μ_D does not change the estimation. To examine the effect of correlation, we compare the Cases 1, 4 and 7. As the correlation coefficient increases, the both standard error estimates, i.e., MSD and ESD, increase. But the biases for all parameters still remain small.

4. APPLICATION

The InCHIANTI study is a representative population-based study of older persons living in the Chianti geographic area (Tuscany, Italy). The goal of the study is to translate epidemiological research into geriatric clinical tools that makes possible more precise diagnosis and more effective treatment in older persons with mobility problems. A total of 1453 participants (age range 20–102 years) were selected from the populations of Greve in Chianti and Bagno a Ripoli (Tuscany, Italy), using a multistage sampling method. Baseline data were collected between September 1998 and March 2000. For the assessment of kidney function, participants underwent a first assessment at home, in which they were interviewed, taught the procedure for the 24-hour urine collection, and given the 24-hour urine collection container. They were asked to note the beginning and end time of the urine collection in a diary to allow for computation of length of collection. In a second assessment, those who had fasted for at least 8 hours came to the study clinic in the morning to undergo a peripheral blood collection. They brought with them their entire 24-hour urine collection, whose volume was measured. Weight and height were measured at the time of the clinical examination performed during a third visit, with participants wearing light clothes and no shoes.

In this analysis, we focus on the 1155 individuals with age 65 and older. The primary outcomes are the estimated GFR values based on three different methods, namely, the 24-hour CC (Y_1), the CG formula (Y_2) and the MDRD formula (Y_3). The disease of interest is moderate kidney impairment. The main goals of the analysis are to estimate the prevalence of kidney impairment by age and sex and to examine the relationship between the estimated GFR values and age, sex and BMI. The models for prevalence and the three test values (Y_1, Y_2, Y_3) are specified as follows:

$$\begin{aligned}\text{logit}\{P(D=0)\} &= \gamma_0 + \gamma_1 \text{Sex} + \gamma_2 \text{Age} \\ E(Y_k) = \mu_k(\mathbf{x}, D) &= \beta_0^{(k)} + \beta_1^{(k)} \text{Sex} + \beta_2^{(k)} \text{Age} + \beta_3^{(k)} \text{BMI} + \beta_4^{(k)} \text{Disease} + \beta_5^{(k)} \text{Disease} * \text{Sex} + \beta_6^{(k)} \text{Disease} * \text{Age}, k=1, 2, 3.\end{aligned}$$

The priors of the parameters are specified as shown in Section 2. The initial values are chosen randomly from the whole possible range of the value. In the analysis, we run 2 independent chains. For each chain, we discard the 10,000 burn-in sample and calculate the estimates using the next 10,000 MCMC samples. The convergence of the MCMC sampling is assessed by examining the plots of the MCMC runs or the Gelman and Rubin's statistics and Heidelberger and Welch's statistics. In Figure 1, we show the plots of the MCMC runs for the variances of Y_1 , Y_2 and Y_3 for the healthy population. It shows that the Gibbs samplers are mixing well. The patterns for the other parameters are similar. For actual statistics of convergence diagnosis, the point estimates of the Gelman and Rubin's statistics for the parameters γ , $\beta^{(1)}$, $\beta^{(2)}$, $\beta^{(3)}$ have a range of [1.00, 1.06]. The stationary tests of convergence [12] pass for all parameters except for the intercept term in the regression for MDRD. Overall, the convergence of the MCMC runs is confirmed.

The estimates of the regression coefficients for the prevalence and the three GFR measures are shown in Table III. Both sex and age are statistically significantly related to the prevalence of having moderate kidney impairment. The prevalence of kidney impairment is significantly higher among the females and increases significantly with age. The regression coefficients for three test values are also shown in Table III. We see that the all three test values are lower for the females and decrease as age increases. The test values for 24-hour CC and CG formula are positively related to BMI, while the MDRD formula is negatively related to BMI. This is consistent with previous finding that the MDRD formula is less affected by extremely body composition. The subjects with kidney impairment have significantly lower GFR values based on all three tests. There are no statistically significant interactions between disease and sex. For the interaction between disease and age, we found that the interaction is negative for the 24-hour CC and the MDRD formula, indicating that the difference of GFR measures between diseased and nondiseased subjects are more pronounced for older subjects. The prevalences by sex and age and their 95% CIs are shown in Table IV. The estimated prevalences are based on unobserved kidney disease status. For example, at age 65, the prevalences of kidney impairment are 39.2% (24.6%–52.4%) and 57.7% (45.1%–69.2%) for male and female, respectively.

Based on the K/DOQI definition, [17] reported that the prevalences of CKD stages 1–4 in the United States were 10.0% in 1988–1994 and 13.1% in 1999–2004, respectively. There seems to be a wide difference between the estimated prevalences of kidney impairment in Table IV and the prevalences of CKD in the US. There are two main reasons. First, the K/DOQI definition of CKD is based on a certain threshold for estimated GFR. For example, Stage 1 CKD is defined as persistent albuminuria with an estimated GFR higher than 90 mL/min/1.73m². While the status of kidney impairment is a latent indicator of diseases. Although conceptually both definitions indicates the impairment of kidney function, the prevalences estimated in Table IV cannot be simply interpreted as the prevalences of CKD. Second, the prevalences of CKD depend strongly on age. In the same paper, Coresh et al. [17] showed that the prevalence of CKD increased from less than 5% for 20–39 years old to around 45% for 70+ years. In our paper, the population under study are 65 years or older from south Europe. In addition, Pizzarelli et al. [18] found that the CKD prevalence for the southern European elderly population is high and varies widely according to the method of estimating GFR. In this sense, the estimated prevalences of kidney impairment are consistent with the CKD prevalence based on K/DOQI definition.

Because the multiple measures of GFR are available, we would like to examine whether combining the multiple tests can improve the disease diagnosis. The optimal weights of combining the three tests are shown in Table V. Because the weight of the test is based on the regression coefficients, it is not appropriate if the multiple tests have different scales. In this analysis, all three tests are indirect measurements of GFR, the scales, i.e., mean, standard deviation and range, for the three tests are very close. In this situation, the weights can be interpreted by their actual values. We can see that for both males and females, the CG formula has a higher weight for younger people and the MDRD formula has a higher weight for older people. The weights for the 24-hour CC are relatively smaller than the other two tests. For example, for a 70-year old man, the contribution (weight) of the CG formula is 1.067 while the weights for the other two tests are 0.106 and -0.173 , respectively. This indicates that the CG formula is a better measure of GFR for the 70-year old men. In general, however, when the tests have different scales, one solution is to create standardized z-score for each test so that all tests have the same scale.

The AUCs of the three diagnostic tests and their combination are shown in Figure 2 by age and sex. The composite test performs the best compared to the individual tests. Again, we do not see significant difference in test accuracy between males and females. The MDRD formula outperforms the 24-hour CC for both sexes and all ages. As age increases, the accuracies of the 24-hour CC and MDRD measures increase, while the accuracy of CG formula decreases. This shows clearly that the performance of each single test and their combination depends on age. In practice, it might be complicated to create a composite test score based on the three tests. Based on Figure 2, we see that the CG formula performs the best after subjects younger than 77 and the AUC for the CG formula is close to the combined test and the MDRD formula is better than the other two tests for age older than 77.

To examine whether the combination of three tests outperforms the pairwise combinations, we conduct additional analysis. The AUCs for the three tests combined and pairwise combination are shown in Figure 3. For both male and female, the pairwise combinations perform better than each single test of the pairs. For age younger than 73, the AUCs of the CC+CG and CG+MDRD combinations are close to that for the three-test combination. For age older than 73, the CC+MDRD combination outperforms the other two pairwise combinations. Overall, the three-test combination performs uniformly better than all three pairwise combinations.

5. DISCUSSION

In this article, we propose a method of estimating the accuracy of the multiple continuous diagnostic tests in the absence of gold standard, where the unobserved disease status is treated as a latent variable. The multiple tests are then combined to create a composite test, which could significantly increase diagnostic accuracy than each single test. The evaluation of diagnostic accuracy of multiple diagnostic tests is possible via a Bayesian latent disease model and the MCMC algorithm. The proposed method can incorporate other decision rules, for example, believe the negative or believe the positive rule. We can also extend this method to multiple ordinal diagnostic test without a gold standard.

However, the proposed method is based on the binormal assumption for test values, which may require a suitable transformation in order to conform with this assumption. For the situations with a GS, a Box-Cox transformation of the test scores conditional on the disease status can be used. For the situations in the absence of a GS, the transformation is less straightforward. If distributions of diagnostic test values have been previously reported in published studies when a GS is used, it is possible that a particular transformation is

recommended. If training data are available in conjunction with the current data, a suitable transformation could be determined by using them. However, expert scientific opinion might be the only available source of information. The opinion will likely be based on knowledge of other diagnostic tests like the one in current use, where some form of training data were available for determining a suitable transformation.

Acknowledgments

The research was supported in part by the Intramural Research Program of the National Institute on Aging.

REFERENCES

1. Levey AS, Coresh J, Balk E, Kausz AT, Levin A, Steffes MW, Hogg RJ, Perrone RD, Lau J, Eknoyan G, et al. National kidney foundation practice guidelines for chronic kidney disease: Evaluation, classification, and stratification. *Annals of Internal Medicine*. 2003; 139(2):137–147. +136. [PubMed: 12859163]
2. Levey AS, Eckardt KU, Tsukamoto Y, Levin A, Coresh J, Rossert J, Zeeuw DD, Hostetter TH, Lameire N, Eknoyan G. Definition and classification of chronic kidney disease: A position statement from kidney disease: Improving global outcomes (kdigo). *Kidney International*. 2005; 67:2089–2100. [PubMed: 15882252]
3. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. Modification of diet in renal disease study group. a more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Annals of Internal Medicine*. 1999; 130:461–470. [PubMed: 10075613]
4. Giannelli SV, Patel KV, Windham G, Pizzarelli F, Ferrucci L, Guralnik JM. Magnitude of underascertainment of impaired kidney function in older adults with normal serum creatinine. *Journal of the American Geriatrics Society*. 2007; 55:816–823. [PubMed: 17537080]
5. Stevens LA, Coresh J, Greene T, Levey AS. Assessing kidney function – measured and estimated glomerular filtration rate. *New England Journal of Medicine*. 2006; 354:2473–2483. [PubMed: 16760447]
6. Choi YK, Johnson WO, Collins MT, Garnder I. Bayesian inference for receiver operating characteristic curves in the absence of a gold standard. *Journal of Agricultural, Biological, and Environment Statistics*. 2006; 11:210–229.
7. Wang C, Turnbull BW, Grohn Y, Nielsen SS. Estimating receiver operating characteristic curves with covariates when there is no perfect reference test for diagnosis of johnes disease. *Journal of Dairy Science*. 2006; 89:3038–3046. [PubMed: 16840620]
8. Johnson WO, Gastwirth JL, Pearson LM. Screening without a “gold standard”: the hui–walter paradigm revisited. *American Journal of Epidemiology*. 2001; 153:921–924. [PubMed: 11323324]
9. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian Data Analysis*. second edn.. New York: Chapman and Hall; 2003.
10. J, G. *Bayesian Statistics 4* chap. Oxford, UK: Clarendon Press; 1992. Evaluating the accuracy of sampling-based approaches to calculating posterior moments.
11. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science*. 1992; 7:457–511.
12. Heidelberger P, Welch PD. Simulation run length control in the presence of an initial transient. *Operations Research*. 1983; 31:1109–1144.
13. Lunn DJ, Thomas A, Best N, Spiegelhalter D. Winbugs – a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*. 2000; 10:325–337.
14. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2009. URL <http://www.R-project.org>, ISBN 3-900051-07-0
15. Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*. 1993; 88:1350–1355.
16. Schisterman EF, Faraggi D, Reiser B. Adjusting the generalized roc curve for covariates. *Statistics in Medicine*. 2004; 23:3319–3331. [PubMed: 15490426]

17. Coresh J, Selvin E, Stevens LA, Manzi J, Kusek JW, Eggers P, Van Lente F, Levey AS. Prevalence of Chronic Kidney Disease in the United States. *JAMA*. 2007; 298(17):2038–2047. [PubMed: 17986697]
18. Pizzarelli F, Lauretani F, Bandinelli S, Windham GB, Corsi AM, Giannelli SV, Ferrucci L, Guralnik JM. Predictivity of survival according to different equations for estimating renal function in community-dwelling elderly subjects. *Nephrology Dialysis Transplantation*. 2009; 24(4):1197–1205.

Figure 1.
MCMC runs of the variances of Y_1 , Y_2 and Y_3 for the healthy population

Figure 2.
AUCs of the three diagnostic tests and the three-test combination by age and sex

Figure 3.
AUC of the three-test combination and pairwise combinations by age and sex

Table I

Parameter values used in the simulation

Case	ρ_D	ρ_D	μ_{1D}	μ_{2D}
1	0.0	0.0	5	4
2	0.0	0.0	5	3
3	0.0	0.0	4	3
4	0.4	0.4	5	4
5	0.4	0.4	5	3
6	0.4	0.4	4	3
7	0.4	0.8	5	4
8	0.4	0.8	5	3
9	0.4	0.8	4	3

Table II

Summary of the parameter estimates

Case	Parameter											
	γ		μ_D		Σ_D		μ_D		Σ_D			
	γ_0	γ_1	μ_{1D}	μ_{2D}	σ_{1D}^2	σ_{2D}^2	σ_{12D}	μ_{1D}	μ_{2D}	σ_{1D}^2	σ_{2D}^2	σ_{12D}
1	Bias	-0.01	0.02	0.00	-0.00	0.03	0.01	0.01	-0.00	0.00	0.01	0.00
	MSE	0.09	0.14	0.05	0.08	0.08	0.08	0.05	0.05	0.08	0.12	0.08
	ESE	0.09	0.14	0.05	0.08	0.08	0.07	0.05	0.05	0.08	0.12	0.09
	CR	95.4	93.8	95.6	95.7	95.9	96.5	95.4	94.5	95.6	95.2	95.3
2	Bias	-0.01	0.02	0.00	-0.01	0.03	0.01	0.01	-0.00	0.01	0.01	0.01
	MSE	0.09	0.14	0.05	0.08	0.08	0.08	0.06	0.05	0.08	0.13	0.09
	ESE	0.09	0.14	0.05	0.08	0.08	0.08	0.05	0.05	0.08	0.13	0.09
	CR	96.1	95.3	94.6	94.6	95.3	95.9	95.6	95.1	94.4	94.2	94.3
3	Bias	-0.01	0.02	0.00	-0.01	0.03	0.02	0.01	-0.00	0.01	0.01	0.00
	MSE	0.09	0.14	0.06	0.08	0.09	0.08	0.06	0.06	0.08	0.13	0.09
	ESE	0.09	0.14	0.06	0.08	0.08	0.08	0.06	0.06	0.08	0.13	0.10
	CR	95.2	94.0	96.6	95.0	97.0	94.3	95.4	94.8	95.4	95.0	94.1
4	Bias	-0.00	0.01	0.00	-0.01	0.03	0.03	0.02	0.00	-0.00	0.01	0.01
	MSE	0.09	0.14	0.06	0.08	0.09	0.08	0.07	0.06	0.08	0.13	0.10
	ESE	0.09	0.14	0.06	0.08	0.09	0.08	0.07	0.05	0.08	0.13	0.10
	CR	95.1	94.0	95.5	95.8	95.9	96.2	95.0	95.5	96.1	94.0	96.0
5	Bias	-0.01	0.02	0.01	-0.02	0.04	0.02	0.02	0.01	-0.01	0.02	0.01
	MSE	0.09	0.14	0.06	0.08	0.09	0.08	0.07	0.06	0.08	0.14	0.10
	ESE	0.09	0.15	0.17	0.33	0.10	0.09	0.07	0.11	0.21	0.15	0.10
	CR	95.0	94.2	96.4	95.5	96.2	94.6	93.5	95.1	94.9	95.4	95.0
6	Bias	-0.01	0.01	0.01	-0.01	0.04	0.02	0.02	0.01	-0.01	0.02	0.02
	MSE	0.10	0.15	0.07	0.08	0.10	0.09	0.08	0.06	0.08	0.16	0.12

Parameter																		
Case	γ		μ_D				Σ_D				μ_D				Σ_D			
	γ_0	γ_1	μ_{1D}	μ_{2D}	σ_{12D}^2	σ_{1D}^2	σ_{2D}^2	σ_{12D}^2	μ_{1D}	μ_{2D}	σ_{12D}^2	σ_{1D}^2	σ_{2D}^2	σ_{12D}^2	σ_{1D}^2	σ_{2D}^2		
ESE	0.10	0.15	0.07	0.08	0.10	0.09	0.07	0.07	0.06	0.08	0.16	0.14	0.13					
CR	94.8	94.0	96.2	96.2	96.6	95.1	96.4	96.4	95.9	94.9	94.7	94.3	94.0					
7	Bias	-0.00	0.02	0.01	-0.01	0.03	0.02	0.03	0.01	-0.01	0.01	0.01	0.01	0.01				
MSE	0.09	0.14	0.06	0.08	0.09	0.08	0.07	0.07	0.06	0.08	0.14	0.14	0.13					
ESE	0.10	0.16	0.17	0.33	0.10	0.08	0.08	0.08	0.14	0.26	0.14	0.14	0.13					
CR	93.0	93.6	95.3	94.2	95.2	96.1	95.3	95.3	94.7	94.9	94.6	96.6	94.7					
8	Bias	-0.01	0.01	0.01	-0.01	0.03	0.01	0.02	0.01	-0.01	0.02	0.01	0.02	0.01	0.02			
MSE	0.09	0.14	0.06	0.08	0.10	0.08	0.06	0.06	0.05	0.08	0.15	0.13	0.13					
ESE	0.10	0.15	0.06	0.08	0.09	0.07	0.06	0.06	0.05	0.08	0.15	0.13	0.13					
CR	93.7	94.5	95.4	94.6	95.9	95.9	95.4	95.4	94.2	95.5	95.4	95.2	96.6					
9	Bias	-0.01	0.02	0.02	-0.03	0.04	0.02	0.03	0.02	-0.03	0.03	0.02	0.02	0.02				
MSE	0.11	0.15	0.08	0.08	0.11	0.08	0.08	0.08	0.06	0.09	0.18	0.17	0.16					
ESE	0.11	0.18	0.20	0.37	0.11	0.09	0.08	0.15	0.29	0.19	0.18	0.18	0.18					
CR	95.3	95.3	95.3	94.9	96.6	95.5	96.2	95.8	95.3	95.1	95.3	94.7	94.7					

Table III

Estimates of regression coefficients for prevalence and three GFR measures

Variable	Mean	Median	SD	MC error	95% CI	p-value
1. Prevalence model						
Intercept	0.448	0.433	0.316	0.021	(-0.094, 1.121)	0.157
Sex	-0.761	-0.773	0.276	0.016	(-1.278, -0.178)	0.006
Age	-0.040	-0.040	0.018	0.001	(-0.076, -0.006)	0.025
2. Regression for 24-hour serum creatinine clearance						
Intercept	102.300	102.400	2.799	0.163	(96.300, 107.500)	0.000
Sex	-8.848	-8.871	3.094	0.164	(-14.890, -2.575)	0.004
Age	-0.980	-0.986	0.233	0.013	(-1.411, -0.505)	0.000
BMI	1.019	1.019	0.165	0.004	(0.694, 1.344)	0.000
Disease	-11.540	-11.670	4.104	0.244	(-18.970, -2.528)	0.005
Disease*Sex	-2.517	-2.522	3.902	0.214	(-10.270, 5.108)	0.519
Disease*Age	-0.494	-0.494	0.272	0.016	(-1.051, 0.020)	0.069
3. Regression for the Cockcroft Goult formula						
Intercept	91.640	91.720	1.827	0.121	(87.780, 95.190)	0.000
Sex	-5.261	-5.308	1.873	0.119	(-8.861, -1.233)	0.005
Age	-1.574	-1.571	0.128	0.008	(-1.843, -1.324)	0.000
BMI	1.711	1.709	0.098	0.003	(1.522, 1.908)	0.000
Disease	-14.980	-14.990	2.330	0.152	(-19.300, -9.650)	0.000
Disease*Sex	-0.738	-0.679	2.182	0.139	(-5.093, 3.347)	0.735
Disease*Age	0.293	0.290	0.146	0.009	(0.016, 0.588)	0.045
2. Regression for the MDRD formula						
Intercept	88.640	88.600	1.995	0.128	(84.750, 92.590)	0.000
Sex	-2.257	-2.306	2.321	0.146	(-6.773, 2.799)	0.331
Age	-0.258	-0.257	0.158	0.010	(-0.581, 0.059)	0.102
BMI	-0.341	-0.340	0.123	0.003	(-0.582, -0.099)	0.006

Variable	Mean	Median	SD	MC error	95% CI	p-value
Disease	-11.830	-11.870	2.412	0.151	(-16.590, -7.087)	0.000
Disease*Sex	-2.601	-2.533	2.673	0.167	(-7.941, 2.266)	0.331
Disease*Age	-0.283	-0.290	0.179	0.011	(-0.630, 0.088)	0.113

Table IV

Estimated prevalences and 95% CI of kidney impairment in the absence of gold standard

Age	Male	Female
65	39.2% (24.6%, 52.4%)	57.7% (45.1%, 69.2%)
70	43.9% (30.5%, 57.0%)	62.4% (52.7%, 71.4%)
75	48.7% (35.8%, 62.5%)	66.9% (58.4%, 74.7%)
80	53.5% (40.3%, 68.5%)	71.0% (61.9%, 79.1%)
85	58.3% (43.2%, 74.1%)	74.8% (64.2%, 84.0%)

Table V

Optimal weights of combining the three diagnostic tests

Sex	Age	CC	CG	MDRD
Male	65	0.015	1.531	−0.547
	70	0.106	1.067	−0.173
	75	0.219	0.489	0.292
	80	0.364	−0.250	0.886
Female	65	0.060	1.239	−0.299
	70	0.151	0.766	0.083
	75	0.264	0.182	0.554
	80	0.406	−0.555	1.149