Published in final edited form as:

Stat Med. 2008 September 30; 27(22): 4489–4501. doi:10.1002/sim.3298.

A Nonparametric Likelihood Test for Detecting Discordance Between Two Measurements with Application to Censored Viral Load Determinations

Zonghui Hu^{1,*}, Dean Follmann¹, Jing Qin¹, Robin L. Dewar², and Phumelele Sangweni³

¹Biostatistics Research Branch, National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, MD, USA

²Virus Isolation and Serology Laboratory, SAIC-Frederick, INC., NCI-Frederick, Frederick, MD, USA

³Project Phidisa, 1 Military Hospital, Pretoria, SA

Abstract

Patient management frequently involves quantitative evaluation of a patient's attributes. For example in HIV studies, a high viral load can be a trigger to initiate or modify an antiretroviral therapy. At times, a new method of evaluation may substitute for an established one, provided that the new method does not result in different clinical decisions as compared to the old method. Traditional measures of agreement between the two methods are inadequate for deciding if a new method can replace the old. Especially when the data are censored by a detection limit, estimates of agreement can be biased unless the distribution for the censored data is correctly specified; this is usually not feasible in practice. We propose a nonparametric likelihood test which seamlessly handles censored data. We further show that the proposed test is a generalization of the test on nominal measurement concordance to continuous measurement. An exact permutation procedure is proposed for implementing the test. Our application is an HIV study to determine whether one method of processing plasma samples can safely substitute for the other. The plasma samples are used to determine viral load and a large portion of data are left censored due to a lower detection limit.

Keywords

Detection limit; likelihood ratio; maximum likelihood estimation; McNemar test; nonparametric likelihood

1 Introduction

Some clinical studies involve two methods for measuring a continuous variable. The measured variable may be clinically important and affect treatment choice. It is thus

*Correspondence to: 6700A Rockledge Drive, Bethesda MD 20892-7609. Phone: (301)4512434, Fax: (301)4802319, E-mail: huzo@niaid.nih.gov.

important to assure that clinical decisions will not change when considering the substitution of an established method for a new one.

In the management of HIV infected patients, the concentration of HIV in the blood, or viral load, is used to guide therapeutic decisions. A patient's viral load strongly indicates HIV disease severity and predicts disease progression and infectiousness. For example, it is believed that HIV RNA suppression below 50 copies/mL should become the standard efficacy endpoint for HARTT therapies [1], thus patients not achieving < 50 copies/mL should be prescribed a change in therapy. It is also found that patients with viral loads < 1500 copies/mL are rarely infectious ([2] and [3]).

An ongoing collaboration in South Africa, Project Phidisa, is establishing a clinical trial network to study HIV disease. Central to the endeavor is reliable viral load measurement. The standard method for collecting and transporting plasma samples is by EDTA tubes, where blood is drawn into a tube coated with anticoagulant EDTA. The EDTA tube is centrifuged and plasma is pipetted out. Plasma is then frozen and shipped to a central lab for viral load determination. However, this standard EDTA method is not easily implemented in South Africa, where it can be hard to transport frozen samples in the bushes. A more convenient method of plasma processing is by PPT tubes, which have a barrier that rides between blood cells and plasma after centrifugation. PPT tubes can then be shipped at ambient temperature. Prior to shipping, plasma can be either separated from the cells or left unseparated, we call them the PPT separated and the PPT unseparated method respectively. For a reference on these methods, see [4]. Of interest in Project Phidisa is whether the PPT methods can replace EDTA. Holodniy et al. studied these tubes and concluded that both PPT methods are equivalent to EDTA [5], whose study was based on HIV infected patients with viral loads above 5,000 RNA copies/mL. However, it is the performance of these methods over low viral loads (< 5000 copies/mL) that is of the greatest importance. For example, a viral load reading below (above) 50 copies/mL would indicate that the current therapy works (fails) to inhibit viral loads and the patient should stay (switch) therapy. If PPT is to substitute EDTA, we would at least require that a patient's the viral load reading is equally likely to be below 50 copies/mL from PPT and EDTA.

In Project Phidisa, patients with a wide range of viral loads were involved, from 750,000 RNA copies/mL down to below 50 RNA copies/mL, with a lower detection limit of 50 RNA copies/mL. Figure 1 presents the log₁₀ viral load readings of each patient from PPT versus EDTA, the left panel for the PPT separated and the right panel for the PPT unseparated. It seems that the readings from the PPT separated are close to those from the EDTA, while the readings from the PPT unseparated are likely higher than the respective readings from EDTA, at least for low viral loads. In particular, there are 92 subjects with PPT unseparated readings above 50 copies/mL and the respective EDTA readings below 50 copies/mL, but only 6 subjects with EDTA readings above 50 copies/mL and the respective PPT unseparated below 50 copies/mL.

There are methods for evaluating the agreement between two methods over the same subjects, for example, the descriptive difference plot of Bland and Altman ([6], [7], [8]) and the agreement coefficients like the intraclass correlation coefficient (ICC; [9], [10], [11])

and the concordance correlation coefficient (CCC; [12]). The difference plot is a scatter plot of the difference between the paired measurements against the average of the two. If the two methods agree, all points should be randomly scattered around the horizontal line through the origin.

Though there are different types of ICC and CCC in literature, the basic concepts are simple. The ICC originates from ANOVA model of the observed readings, and basically calculates the percentage of variation not caused by the use of two different methods. The ICC ranges from 0 to 1. The CCC measures the closeness of the paired measurements to the concordance line (the 45° line through the origin) and ranges from -1 to 1. Larger values of both agreement coefficients indicate better agreement. A drawback with ICC and CCC is that they can indicate stronger agreement for more heterogeneous populations [13]. Another drawback is that ICC and CCC may indicate increasing agreement as the marginal distributions become more different even if the true agreement decreases or stays the same [14].

In applications like Project Phidisa, a graphical or a quantitative evaluation of the agreement between the two methods (one is EDTA, the other is PPT separated or PPT unseparated) is not enough. We need to construct a test for making a decision about whether one method can substitute for the other. It is hard to develop a test from agreement coefficients for this purpose. First, it is hard to decide a threshold in the null hypothesis, such that the agreement coefficient above the threshold corresponds to substitutability due to the above mentioned drawbacks. Second, if the data are censored, agreement coefficient estimates can be biased, unless correctly specified parametric distributions are assumed [15]. In practice, it is not possible to know if the assumed parametric distribution applies below the detection limit. The presence of a detection limit also hampers the use of Bland-Altman's difference plot.

In this paper, we propose a test based on nonparametric likelihood. In HIV studies, a clinical criterion for assessing disease severity is usually based on whether a patient's viral loads are below a certain level s [1-3]. To determine if a new method can substitute for the other, we especially care whether a patient's viral load measurements from the two methods are equally likely to be below s . This is equivalent to the concordance between the two measurements in terms of the binary event "viral load reading below/above s ".

Nonparametric likelihood can be applied to evaluate the concordance at s , as we will show in the next section. For continuous measurements within a region Ω , we can let the value of s vary over Ω and develop a global nonparametric likelihood (NPL) test statistic to test on the overall concordance between the two measurements. The NPL test can thus be viewed as a generalization of tests for nominal measurement concordance to continuous measurements. If the overall concordance is not rejected by the NPL test, we would conclude that the observed data does not contradict the equivalence between the two methods. As a nonparametric approach, the NPL test does not require any distributional assumptions on the two measurements even with censoring. It thus obviates the need for correct model specification and is attractive where little is known about the measurement distributions.

The remainder of this paper is organized as follows: § 2 develops the NPL test and describes an exact procedure for implementation, § 3 presents simulation results, § 4 applies the method to paired viral load measurements, while § 5 concludes with discussion.

2 Nonparametric Likelihood Test

We have paired measurements (X_k, Y_k) from subject k , where X_k is from one method, and Y_k from the other, for $k = 1, \dots, n$. The bivariate variables (X_k, Y_k) are iid, and follow a joint distribution F_{XY} with marginal distributions F_X and F_Y .

At a given cutpoint s , dichotomization of X and Y leads to binary variables X_s and Y_s , with $X_s = 1$ if $X > s$ and 0 otherwise, similarly for Y_s . Thus four categories $\mathcal{C}_{ij}(s)$, $i, j = 0$ and 1, are formed as shown by the 2×2 contingency table below,

$$\begin{array}{cc} & Y_s=0 & Y_s=1 \\ X_s=0 & n_{00}(s) & n_{01}(s) \\ X_s=1 & n_{10}(s) & n_{11}(s) \end{array}$$

where $n_{ij}(s)$ denotes the number of subjects in category $\mathcal{C}_{ij}(s)$, e.g.,

$n_{01}(s) = \sum_{k=1}^n I(X_k \leq s, Y_k > s)$ is the number of subjects with first measurement below s and the second above s . Let $p_{ij}(s) = P\{(X, Y) \in \mathcal{C}_{ij}(s)\}$, thus $p_{00}(s) + p_{11}(s)$ is the probability of concordance between X_s and Y_s , or the concordance between X and Y around s , and $p_{01}(s) + p_{10}(s)$ the probability of discordance around s .

The nonparametric log likelihood function for $\{p_{ij}(s), i, j = 0, 1\}$ is

$$t(s) = \sum_{i,j} n_{ij}(s) \log \{p_{ij}(s)\}. \quad (1)$$

We are interested in whether the two measurements have the same probability of being below s , that is,

$$H_0^s: F_X(s) = F_Y(s)$$

Note that since $F_X(s) = p_{00}(s) + p_{01}(s)$ and $F_Y(s) = p_{00}(s) + p_{10}(s)$, H_0^s is equivalent to $p_{01}(s) = p_{10}(s)$.

The log likelihood ratio with respect to H_0^s is thus,

$$\begin{aligned} \lambda(s) &= 2 \left\{ l(s; \hat{p}_{00}, \hat{p}_{01}, \hat{p}_{10}, \hat{p}_{11}) - l(s; \hat{p}_{00}^0, \hat{p}_{01}^0, \hat{p}_{10}^0, \hat{p}_{11}^0) \right\} \\ &= 2 \left\{ n_{01}(s) \log \left(\frac{2n_{01}(s)}{n_{01}(s) + n_{10}(s)} \right) + n_{10}(s) \log \left(\frac{2n_{10}(s)}{n_{01}(s) + n_{10}(s)} \right) \right\} \end{aligned} \quad (2)$$

where $\hat{p}_{ij}(s)$ is the maximum likelihood estimate (MLE) of $p_{ij}(s)$, and $\hat{p}_{ij}^0(s)$ is the MLE under H_0^s . It is easy to show that $\hat{p}_{ij}(s) = n_{ij}(s)/n$, and $\hat{p}_{ii}^0(s) = \hat{p}_{ii}(s)$ for $i = 0, 1$,

$\hat{p}_{01}^0(s) = \hat{p}_{10}^0(s) = \{\hat{p}_{01}(s) + \hat{p}_{10}(s)\} / 2$. Equation (2) follows by plugging these MLE's into the two respective log likelihood functions.

Define $\hat{p}(s) = n_{01}(s) / \{n_{01}(s) + n_{10}(s)\}$. A 2nd order Taylor expansion of the log likelihood ratio around $\hat{p}(s) = 0.5$ is,

$$\lambda(s) = 2 \{n_{01}(s) + n_{10}(s)\} [\hat{p}(s) \log \{2\hat{p}(s)\} + \{1 - \hat{p}(s)\} \log \{2 - 2\hat{p}(s)\}] \approx \{n_{01}(s) + n_{10}(s)\} \{2\hat{p}(s) - 1\}^2 \quad (3)$$

Under H_0^s and conditional on $n_{01}(s) + n_{10}(s)$, $n_{01}(s) \sim \text{Binomial}(n_{01}(s) + n_{10}(s), 0.5)$. Thus $\hat{p}(s)$ asymptotically follows a normal distribution with mean 0.5 and variance $0.5^2 / \{n_{01}(s) + n_{10}(s)\}$, and $\lambda(s)$ asymptotically follows a χ_1^2 distribution. The asymptotic form (3) of $\lambda(s)$ is in fact the McNemar statistic. Thus, at each fixed s , the nonparametric likelihood ratio is asymptotically equivalent to the McNemar statistic.

For continuous measurements X and Y with common support Ω , we have a sequence of such nonparametric likelihood ratios $\lambda(s)$ with s varying within Ω . For the overall concordance between X and Y , we require that for any $s \in \Omega$, there is the same probability for the two measurements to be below s . Therefore, the null hypothesis is,

$$H_0: F_X(s) = F_Y(s), \text{ for all } s \in \Omega.$$

In other words, the global null H_0 is that the two measurements follow the same marginal distribution. It is easy to see that, if H_0 is rejected, $F_{XY}(x, y) \neq F_{XY}(y, x)$, i.e., the two measurements are not exchangeable. In the following, we sometimes use “exchangeability” for whether two methods can substitute for each other.

We define a global statistic Λ for the test of H_0 , which is the supremum of the likelihood ratios over Ω . We call Λ the NPL test statistic.

$$\Lambda = \sup_{s \in \Omega} \lambda(s). \quad (4)$$

Each $\lambda(s)$ evaluates the discordance between X and Y around a specific cutpoint s , with larger values indicating bigger discordance between X_s and Y_s . Λ uses the information at the location where X and Y disagree the most. This idea of taking the supremum of a sequence of statistics is like that of a Kolmogorov-Smirnov test [16]. Since $\lambda(s)$ only changes when s changes from one observed value of X and Y to another, Λ is simply

$$\Lambda = \max_{s \in (x_1, \dots, x_n, y_1, \dots, y_n)} \lambda(s).$$

If Λ_0 is the critical value satisfying $P(\Lambda \geq \Lambda_0) = 1 - \alpha$ at the nominal significance level α , we reject H_0 when Λ is greater than or equal to Λ_0 . Rejection of H_0 means that the two measurements X and Y are not exchangeable and one method cannot substitute the other. Λ is the maximum of a sequence of correlated statistics (shown in the appendix), whose

distribution is not of any known family [17], though there are the simulation results [18]. We thus propose an exact permutation procedure for performing the test.

As shown in the appendix, $F_{XY}(x, y) = F_{YX}(y, x)$ implies H_0 . Therefore, we can obtain the null distribution of the NPL test statistic Λ by the following permutation procedure. Within each subject k , randomly decide whether or not to switch X_k with Y_k according to a Bernoulli(0.5), and thus obtain a permuted data point. For the r -th permuted data, compute the corresponding NPL test statistic which we denote as $\Lambda^{<r>}$ following Table 1.

In the table, $\{s_m, m = 1, \dots, M\}$ is the sequence of cutpoints, with M as the number of distinct observed values of X and Y ; $\lambda^{<r>}(s_m)$ is the nonparametric likelihood ratio evaluated at s_m from the r -th permuted data. The value of the NPL test statistic from the r -th permutation is

$$\Lambda^{<r>} = \max_{1 \leq m \leq M} \lambda^{<r>}(s_m)$$

The p-value for H_0 is the proportion of permutations with $\Lambda^{<r>}$ greater than or equal to Λ^{obs} , the value of the NPL test statistic from the observed data. That is,

$$\text{P-value} = \frac{\sum_r I(\Lambda^{<r>} \geq \Lambda^{\text{obs}})}{R}$$

where R is the number of permutations, with $R = 2^n$ for exhaustive permutation and a sufficiently large number for Monte-Carlo permutation. The permutation procedure approaches the exact p-value as the number of permutations gets sufficiently large [19].

In the case of censoring by a detection limit, for example, by the lower detection limit in HIV viral load determination, the NPL test statistic evaluates the discordance between the two measurements over the observation range as well as at the detection limit. That is, $\Lambda^C = \sup_{(s \in \Omega: s \geq C)} \lambda(s) = \max_{s \in (C, x_1, \dots, x_n, y_1, \dots, y_n)} \lambda(s)$, where C is the lower detection limit and $\lambda(C)$ reflects the discordance around the detection limit. Λ^C would be the same as the Λ from the completely observed data without censoring if the maximum of the nonparametric likelihood ratio occurs at some $s \geq C$, i.e., $\sup_{(s \in \Omega)} \lambda(s) = \sup_{(s \in \Omega, s \geq C)} \lambda(s)$.

3 Simulation Results

As shown in § 2, to determine the concordance between two variables X and Y around a point s , instead of the nonparametric likelihood ratio (2), we can also use McNemar statistic (3). The consequent global test statistic is of the same form as (4) except that the supremum of the McNemar statistics over $s \in \Omega$ is used. We call it the generalized McNemar test statistic, which is the generalization of McNemar test to continuous variables. We investigate the performance of the NPL and the generalized McNemar test for detecting discordance between two continuous measurements.

We simulate data to mimic that of Project Phidisa. Let $\{Z_k, k = 1, \dots, n\}$ be the underlying true values (for example, the true viral loads) for the n subjects, and assume

$$Z_k \sim iid \quad N\left(\mu_z, \sigma_z^2\right). \quad (5)$$

Let (X_k, Y_k) be the two measurements of Z on subject k (for example, X_k the reading of viral load from EDTA, and Y_k from PPT). Conditional on Z , the paired measurements follow a bivariate normal distribution,

$$\begin{pmatrix} X_k \\ Y_k \end{pmatrix} | Z_k=z \sim N\left\{\begin{pmatrix} \mu_x(z) \\ \mu_y(z) \end{pmatrix}, \Sigma(z)\right\}, \quad (6)$$

independently for $k = 1, \dots, n$. The conditional covariance $\Sigma(z)$ is assumed diagonal with $\text{var}(X|Z=z) = \sigma_x^2(z)$ and $\text{var}(Y|Z=z) = \sigma_y^2(z)$. Unconditionally, X has mean $E\{\mu_x(Z)\}$ and variance $E\{\sigma_x^2(Z)\} + \text{var}\{\mu_x(Z)\}$, Y has mean $E\{\mu_y(Z)\}$ and variance $E\{\sigma_y^2(Z)\} + \text{var}\{\mu_y(Z)\}$, X and Y are correlated with $\text{cov}(X, Y) = E\{\mu_x(Z)\mu_y(Z)\} - E\{\mu_x(Z)\}E\{\mu_y(Z)\}$.

We set $\mu_z = 10$, $\sigma_z = 2$ in (5). For the conditional distribution of X and Y in (6), we consider 4 cases. In case (I), $\mu_x(z) = \mu_y(z) = z$ and $\sigma_x^2(z) = \sigma_y^2(z) = 1$; In case (II), $\mu_x(z) = \mu_y(z) = z$, $\sigma_x^2(z) = 1$ and $\sigma_y^2(z) = 3$; In case (III), $\mu_x(z) = z$, $\mu_y(z) = z$ for $z > 9$ and $\mu_y(z) = z + 3$ for $z \leq 9$, $\sigma_x^2(z) = \sigma_y^2(z) = 1$; In case (IV), $\mu_x(z) = z$, $\mu_y(z) = z$ for $z > 9$ and $\mu_y(z) = z + 2$ for $z \leq 9$, $\sigma_x^2(z) = 1$ and $\sigma_y^2(z) = 3$. Therefore, X and Y agree in case (I) which is likely the situation for PPT separated versus EDTA, while X and Y disagree in the other 3 cases with (III) or (IV) likely the situation for PPT unseparated versus EDTA. There are 1000 simulations, each with $n = 80$ subjects.

To demonstrate the performance of the tests with respect to censoring, three thresholds for left censoring are considered, $C = 8.5, 9.5, 10$.

First, we evaluate the performances of the NPL test and the generalized McNemar test. Since H_0 implies equal marginal means, and the paired t -test and the Wilcoxon signed rank test are conventional for testing equal means of paired measurements, we compare the NPL test with them. For the t and Wilcoxon tests, values of X and Y below C are set equal to C , a naive (ad hoc) approach to deal with left censored data [15]. The empirical powers and significance levels are in Table 2. As a likelihood based test, the NPL test is generally more powerful than the t and the Wilcoxon tests, and slightly more powerful than the generalized McNemar test. Additionally, the NPL and the generalized McNemar tests retain nominal significance as their rejection rates are all below 0.05 in case (I). The most important observation is that the NPL test maintains stable powers in case of censoring. In case (III), the NPL test keeps over 75% power to detect the discordance caused by a shift in the conditional means even at $C = 10$ with 50% X censored and 29% Y censored; while the power of t and Wilcoxon tests drop below 70%. In case (II), the powers of the NPL test under the censoring cases are all close to the power from complete data in detecting the discordance caused by a difference in conditional variances, while results from t and

Wilcoxon tests can be misleading, e.g., at $C = 10$, the t -test has over 70% power to claim difference in the mean while in fact $E(X) = E(Y)$.

For the above simulation data, we compute the agreement coefficient and the linear regression coefficients in the difference plot. In the difference plot, if two variables agree, the least square (LS) linear regression of the paired differences versus the paired averages should have both intercept and slope as zero [9]. For the correlation coefficient, we use the modified concordance correlation coefficient [14], which corrects the second drawback of CCC's that traditional CCC's may indicate higher agreement for two measurements with more different marginal distributions. With good agreement, the modified CCC should be close to 1. Table 3 presents these estimates. It shows that both the difference plot and the agreement coefficient could be misleading with censoring. For example in case (III), the complete data would have a difference plot indicating disagreement with positive intercept (1.100, 95% confidence interval [0.680, 1.519]) and negative slope (-0.402, [-0.621, -0.189]), while the censored data at $C = 10$ corresponds to a difference plot with both slope and intercept near 0 indicating little disagreement. Similarly in case (III), the modified CCC for the censored data at $C = 10$ indicates a stronger agreement than it should be (0.554 versus 0.308). To correct for the biases in difference plot and agreement coefficient, a parametric model assumption for the censored data would be required. However, it is not always possible to obtain the correct model specification especially with large proportion of data censored.

We also performed simulations where Z follows a Gamma distribution. The conclusions were quite similar to Table 2 and 3 and were omitted here for brevity.

4 Case Study

Our data for Project Phidisa consists of 346 HIV infected patients under antiretroviral therapy. Each patient had one plasma sample processed by EDTA, one by PPT separated, and one by PPT unseparated. Among these 346 subjects, 194 had both the PPT separated and EDTA readings below the detection limit of 50 copies/mL, and 118 had both the PPT unseparated and EDTA readings below the detection limit. Overall, more than half of the subjects had either PPT (separated or unseparated) or EDTA readings below the detection limit. This substantial censoring makes parametric approaches problematic.

The goal of this study is to judge whether PPT tubes could safely substitute for EDTA tube. We investigate this separately for the PPT separated and the PPT unseparated by the NPL test.

In Figure 2 are the results from the NPL test. The null distribution of the test statistic Λ is estimated by the empirical distribution of $\Lambda^{<r>}$, $r = 1, \dots, R$, from the permutation under H_0 . Here, Monte Carlo permutation with $R = 2000$ is used. The nonparametric likelihood ratios $[\lambda(s) : s \in \{C, x_1, \dots, x_n, y_1, \dots, y_n\}]$ from the observed data, where $C = \log_{10}(50)$, are plotted against s . Imposed over the nonparametric likelihood ratios are the respective means and 95% upper percentiles from the permuted datasets, and for clarity, the smoothed curves are plotted. Since the nonparametric likelihood ratio $\lambda(s)$ at each fixed s is asymptotically

distributed as $\chi^2(1)$, we thus observe the mean curve around the horizontal line of intercept 1. For the 95% upper percentiles, if at some s_0 , the observed $\lambda(s_0)$ is above the 95% upper percentile at s_0 , it indicates $F_X(s_0) > F_Y(s_0)$. However, even with existence of such s_0 , it does not necessarily lead to the rejection of the null hypothesis H_0 . This is because the NPL test uses the global statistic $\Lambda = \sup_{s \in \Omega} \lambda(s)$, and H_0 is only rejected when Λ^{obs} from the observed data is larger than the 95% upper percentile of the null distribution of Λ .

For the PPT separated, the null distribution density of the NPL test statistic Λ is presented in Figure 2(a), and the NPL test statistic from the observed data, Λ^{obs} , is about 8. The p-value of the NPL test from the permutation procedure of §2 is accordingly $p = 0.432$ and thus failing to reject the hypothesis that PPT separated and EDTA are exchangeable.

For the PPT unseparated, the null distribution density of the NPL test statistic is presented in Figure 2(c). The NPL test statistic from the observed data is about 75, which is literally off the chart. The p-value from permutation is $p < 0.0005$, indicating we should reject the null hypothesis and conclude that PPT unseparated cannot substitute for EDTA. Figure 2(d) further shows that discordance between the PPT unseparated and EDTA occurs around low viral loads, consistent with Figure 1.

If we try to evaluate the agreement by an agreement coefficient, its value depends on the assumed distributions for readings below the detection limit. For the PPT separated with EDTA, if the undetectable readings are assigned the value of the detection limit $\log_{10} 50$, the modified CCC estimate is 0.980 (with 95% confidence interval [0.975, 0.984]); if uniform distribution over interval $[0, \log_{10} 50]$ is assumed for the undetectable readings, the modified CCC estimate is 0.903 ([0.881, 0.920]). For the PPT unseparated with EDTA, if the undetectable readings are assigned the value of the detection limit, the modified CCC estimate is 0.958 ([0.949, 0.967]); if uniform distribution over interval $[0, \log_{10} 50]$ is assumed, the modified CCC estimate is 0.859 ([0.835, 0.879]). The CCC estimates under the two distribution assumptions are quite different. We cannot judge whether PPT can substitute for EDTA based on these estimates, since any assumed distribution for undetectable readings is not examinable.

In conclusion, no discordance is detected between the PPT separated and the EDTA. However, discordance is detected between the PPT unseparated and EDTA, thus the PPT unseparated tube should not be used as a substitute for EDTA.

5 Discussions

We proposed a nonparametric likelihood test for determining the substitutability of two methods relating to a continuous variable. It was shown that the nonparametric likelihood test is asymptotically a generalization of McNemar test, in the sense that the nonparametric likelihood ratio for evaluating the discordance between the two measurements around each point is asymptotically a McNemar statistic. An appealing aspect of the nonparametric approach is its seamless handling of censored data.

For the nonparametric likelihood test discussed in this paper, the criterion for whether one method can substitute for the other is based on the equivalence of the marginal distributions,

$F_X(s) = F_Y(s)$ for $s \in \Omega$, which agrees with the goals specified in Project Phidisa. However, there can be other criteria for substitutability depending on the purpose of the clinical study, and the nonparametric likelihood approach can lead to other types of nonparametric tests.

If at each point s , the agreement between the dichotomized measurements X_s and Y_s is evaluated by the probability of concordant pairs, we have the Kappa statistic [20]. The Kappa statistic at s is $\mathcal{K}(s) = \{p_o(s) - p_e(s)\} / \{1 - p_e(s)\}$, where $p_o(s) = p_{00}(s) + p_{11}(s)$ is the observed probability of concordance, and $p_e(s) = \{p_{00}(s) + p_{01}(s)\} \{p_{00}(s) + p_{10}(s)\} / \{p_{11}(s) + p_{10}(s)\} \{p_{11}(s) + p_{01}(s)\}$ is the expected probability of concordance by chance. Therefore, $\mathcal{K}(s) = 0$ indicates concordance by pure chance and $\mathcal{K}(s) = 1$ indicates perfect concordance where X_s matches Y_s . At each s , a nonparametric log likelihood ratio $\lambda^{\mathcal{K}}(s)$ can be computed with respect to $H_0: \mathcal{K}(s) = 0$. Then, similar to the NPL test, a global statistic—the Kappa based nonparametric likelihood test statistic—can be defined over $\lambda^{\mathcal{K}}(s)$ for $s \in \Omega$, and the test can be performed by a permutation procedure.

ACKNOWLEDGMENT

This project has been funded in whole or in part by federal funds from the National Cancer Institute, National Institute of Health, under contract NOI-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government.

Appendix

1. Proof: exchangeability $F_{XY}(x, y) = F_{YX}(y, x)$ implies H_0

If $F_{XY}(x, y) = F_{YX}(y, x)$, then for any s , $F_X(s) = F_{XY}(s, +\infty) = F_{YX}(+\infty, s)$, which is $F_Y(s)$, so H_0 is true.

In fact, H_0 also implies exchangeability. By Sklar's theorem, the bivariate distribution function F_{XY} can be expressed by the Archimedean copula Ψ as,

$$F_{XY}(x, y) = \Psi^{-1} \{ \Psi(F_X(x)) + \Psi(F_Y(y)) \}.$$

If H_0 is true, $F_X(s) = F_Y(s)$ for all s , then $F_{XY}(x, y) = \Psi^{-1} \{ \Psi(F_Y(x)) + \Psi(F_X(y)) \} = \Psi^{-1} \{ \Psi(F_X(y)) + \Psi(F_Y(x)) \} = F_{YX}(y, x)$, i.e., X and Y are exchangeable.

2. Proof: under H_0 , nonparametric likelihood ratios at two points are correlated

Define $\lambda_{s_1}^{1/2} = \lambda^{1/2}(s_1)$, and similarly $\lambda_{s_2}^{1/2} = \lambda^{1/2}(s_2)$, as the square root of the nonparametric likelihood ratios at two points s_1 and s_2 . Denote $n_{s_1} = n_{01}(s_1) + n_{10}(s_1)$ and similarly for n_{s_2} , denote $\hat{p}_{s_1} = \hat{p}(s_1)$ and similarly for $\hat{p}_{s_2}$. Based on the Taylor approximation of the NPL ratios in (3), the covariance between $\lambda_{s_1}^{1/2}$ and $\lambda_{s_2}^{1/2}$ under H_0 has the following expression,

$$\begin{aligned}
\text{cov}(\lambda_{s_1}^{1/2}, \lambda_{s_2}^{1/2}) &= \text{cov}\{n_{s_1}^{1/2}(2\hat{p}_{s_1} - 1), n_{s_2}^{1/2}(2\hat{p}_{s_2} - 1)\} \\
&= E\{n_{s_1}^{1/2}n_{s_2}^{1/2}(2\hat{p}_{s_1} - 1)(2\hat{p}_{s_2} - 1)\} \\
&= E\{n_{s_1}^{1/2}n_{s_2}^{1/2}(4\hat{p}_{s_1}\hat{p}_{s_2} - 1)\} \\
&\approx n \frac{4D_1(s_1)D_1(s_2) - D_2(s_1)D_2(s_2)}{D_2^{1/2}(s_1)D_2^{1/2}(s_2)}
\end{aligned}$$

where $D_1(s) = F_X(s) - F_{XY}(s, s)$ and $D_2(s) = F_X(s) + F_Y(s) - 2F_{XY}(s, s)$.

As shown in § 2, conditional on $n_s, n_s^{1/2}(2\hat{p}_s - 1)$ is asymptotically $N(0, 1)$ under H_0 . Since $n_s^{1/2}(2\hat{p}_s - 1)$ has asymptotic mean 0 at both $s = s_1$ and $s = s_2$, and we have the 2nd equation. Since $E\{n_{s_1}^{1/2}n_{s_2}^{1/2}(\hat{p}_{s_1} + \hat{p}_{s_2})\} = E(n_{s_1}^{1/2}n_{s_2}^{1/2})$ under H_0 , we get the 3rd equation. In addition, $\text{cov}(\lambda_{s_1}^{1/2}, \lambda_{s_2}^{1/2})$ is actually the correlation coefficient between $\lambda_{s_1}^{1/2}$ and $\lambda_{s_2}^{1/2}$, as $n_s^{1/2}(2\hat{p}_s - 1)$ has asymptotic variance 1 at both $s = s_1$ and $s = s_2$. This correlation coefficient is generally not zero, so $\lambda_{s_1}^{1/2}$ and $\lambda_{s_2}^{1/2}$ are correlated, thus the nonparametric likelihood ratios at two points, $\lambda(s_1)$ and $\lambda(s_2)$, are also correlated.

REFERENCE

- Hill A, Miralles D, Vangeneugden T. Should we now adopt the HIV-RNA ;50 copy endpoint for clinical trials of antiretroviral - experienced as well as naive patients? *AIDS* 21. 2007; 12:1651–1653.
- Quinn TC, Wawer MJ, Sewankambo N. Viral load and heterosexual transmission of human immunodeficiency virus type 1. *New England Journal of Medicine*. 2000; 342:921–929. [PubMed: 10738050]
- Gray RH, Wawer MJ, Brookmeyer R. Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1 discordant couples in Rakai, Uganda. *Lancet*. 2001; 357:1149–1153. [PubMed: 11323041]
- Salimnia H, Moore EC, Crane LR, MacArthur RD, Fairfax MR. Discordance between viral loads determined by Roche COBAS AMPLICOR human immunodeficiency virus type 1 monitor (version 1.5) standard and ultrasensitive assays caused by freezing patient plasma in centrifuged Becton-Dickinson vacutainer brand plasma preparation tubes. *Journal of Clinical Microbiology*. 2005; 43:4635–4639. [PubMed: 16145119]
- Holodniy M, Rainen L, Herman S, Yen-Lieberman B. Stability of plasma human immunodeficiency virus load in vacutainer PPT plasma preparation tubes during overnight shipment. *Journal of Clinical Microbiology*. 2000; 38:323–326. [PubMed: 10618109]
- Bland JM, Altman DG. Measurement in Medicine: The analysis of method comparison studies. *The Statistician*. 1983; 32:307–317.
- Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986; 32:307–317. [PubMed: 2868172]
- Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*. 1999; 8:135–160. [PubMed: 10501650]
- Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*. 1966; 19:3–11. [PubMed: 5942109]
- McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*. 1996; 1:30–46.
- Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*. 1979; 86:420–428. [PubMed: 18839484]

12. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989; 45:255–268. [PubMed: 2720055]
13. Atkinson G, Nevill A. Comment on the use of concordance correlation to assess the agreement between two variables. *Biometrics*. 1997; 53:775–777.
14. Fay M. Random marginal agreement coefficients: rethinking the adjustment for chance when measuring agreement. *Biostatistics*. 2005; 6:171–180. [PubMed: 15618535]
15. Barnhart HX, Song J, Lyles RH. Assay validation for left-censored data. *Statistics in Medicine*. 2005; 24:3347–3360. [PubMed: 15977288]
16. Smirnov NV. Estimate of derivation between empirical distribution functions in two independent samples. *Bulletin of Moscow University*. 1939; 2:3–16.
17. Miller R, Siegmund D. Maximally selected Chi square statistics. *Biometrics*. 1982; 38:1011–1016.
18. Halpern J. Maximally selected chi square statistics for small samples. *Biometrics*. 1982; 38:1017–1023.
19. Efron, B.; Tibshirani, RJ. *An introduction to the Bootstrap*. Chapman and Hall; New York: 1993.
20. Landis JR, Koch GG. Observer agreement for categorical data. *Biometrics*. 1977; 33:159–174. [PubMed: 843571]

Figure 1.

Viral load readings from PPT and EDTA plasma samples in $\log_{10}$ scale. The line is the line of equity, the empty circles stand for subjects with one detectable reading and the other below the detection limit of 50 copies/mL. There are 194 subjects with both PPT separated and EDTA readings below the detection limit, and 118 subjects with both PPT unseparated and EDTA readings below the detection limit.

Figure 2.

NPL test for PPT versus EDTA based on 2000 permutations, the top panels are for the PPT separated and the bottom panels for the PPT unseparated. On the left are the estimated null distribution densities of the NPL test statistic Λ . On the right are the nonparametric likelihood ratios $\lambda(s)$ over the viral load reading s : the points are from the observed data, the upper curves are the 95% upper percentiles of the permuted distribution of $\lambda(s)$, and the lower curves are the means of the permuted distribution of $\lambda(s)$.

Table 1

Arrangement of the permuted NPL test statistics

permutation	nonparametric likelihood ratios					NPL test statistic
1	$\lambda^{<1>}(s_1)$	...	$\lambda^{<1>}(s_m)$	...	$\lambda^{<1>}(s_M)$	$\Lambda^{<1>}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	$\lambda^{<r>}(s_1)$	...	$\lambda^{<r>}(s_m)$	...	$\lambda^{<r>}(s_M)$	$\Lambda^{<r>}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R	$\lambda^{<R>}(s_1)$	...	$\lambda^{<R>}(s_m)$	...	$\lambda^{<R>}(s_M)$	$\Lambda^{<R>}$

Table 2

The empirical powers and significance levels at $\alpha = 0.05$ of the NPL, the generalized McNemar (G-McNemar), the paired t and the Wilcoxon signed rank tests. C is the threshold for left censoring, the 1st line under each case corresponds to no censoring, the 3rd and 4th columns are the censoring proportions. Each simulation study involves $n = 80$ subjects, and 1000 studies are simulated for each case

C	censoring %		empirical power or significance level			
	X	Y	NPL	G-McNemar	t	Wilcoxon
case (I)	0	0	4.4	4	4.7	4.7
	8.5	25	4.7	4.9	5	4.6
	9.5	41	4.4	3.8	4.1	5.3
case (II)	10	50	2	2.2	4.2	3.6
	0	0	84.5	79.2	4.9	5.4
	8.5	25	34	70.7	42.9	29.9
case (III)	9.5	41	45	71.8	66.2	63.2
	10	50	50	68.9	64.4	70.8
	0	0	97.4	97.5	98.8	96.3
case (IV)	8.5	25	6	93.1	94.6	93
	9.5	41	19	85.2	81.3	76.9
	10	50	29	78.6	75.8	64.3
	0	0	81.6	79	36.4	34.1
	8.5	25	26	80.8	78.6	66.4
	9.5	41	37	84.5	80.8	82.7
	10	50	43	85.6	83	85.8

Table 3

Estimated intercepts and slopes of the LS linear fitting in the difference plot with 95% confidence intervals, and the estimates of the modified CCC with 95% confidence intervals. Each simulation study involves $n = 80$ subjects, and 1000 studies are simulated for each case

	<u>C</u>		<u>censoring %</u>		<u>difference plot</u>		<u>CCC</u>
	X	Y	intercept	slope			
case (I)	0	0	-0.008 [-0.325, 0.299]	-0.002 [-0.137, 0.153]		0.796 [0.704, 0.866]	
	8.5	25	0.001 [-0.245, 0.240]	-0.004 [-0.160, 0.143]		0.774 [0.666, 0.864]	
	9.5	41	0.004 [-0.177, 0.170]	0.0003 [-0.190, 0.189]		0.751 [0.606, 0.857]	
	10	50	0.002 [-0.156, 0.169]	-0.003 [-0.217, 0.221]		0.742 [0.589, 0.853]	
case (II)	0	0	-0.017 [-0.601, 0.605]	0.613 [0.395, 0.857]		0.443 [0.284, 0.584]	
	8.5	25	0.187 [-0.209, 0.602]	0.512 [0.236, 0.801]		0.409 [0.212, 0.576]	
	9.5	41	-0.009 [-0.326, 0.307]	0.590 [0.275, 0.903]		0.369 [0.170, 0.559]	
	10	50	-0.206 [-0.507, 0.086]	0.642 [0.294, 0.985]		0.341 [0.114, 0.549]	
case (III)	0	0	1.100 [0.680, 1.519]	-0.402 [-0.621, -0.189]		0.308 [0.102, 0.492]	
	8.5	25	0.726 [0.374, 1.076]	-0.159 [-0.382, 0.0478]		0.466 [0.255, 0.639]	
	9.5	41	0.457 [0.175, 0.753]	-0.051 [-0.265, 0.165]		0.532 [0.329, 0.707]	
	10	50	0.205 [-0.039, 0.487]	0.006 [-0.244, 0.289]		0.554 [0.316, 0.726]	
case (IV)	0	0	0.440 [-0.254, 1.127]	0.568 [0.257, 0.873]		0.306 [0.100, 0.482]	
	8.5	25	0.382 [-0.084, 0.839]	0.562 [0.276, 0.858]		0.275 [0.073, 0.478]	
	9.5	41	0.097 [-0.237, 0.445]	0.648 [0.301, 0.979]		0.260 [0.044, 0.484]	
	10	50	0.149 [-0.504, 0.188]	0.712 [0.314, 1.110]		0.243 [-0.001, 0.476]	