STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:3347–3360
Published online 24 June 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2225

Assay validation for left-censored data

Huiman X. Barnhart[1,*,†], Jingli Song[2,‡] and Robert H. Lyles[3,§]

[1]*Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University, P.O. Box 17969, Durham, NC 27715, U.S.A.*
[2]*Eli Lilly and Company, Lilly Corporate Center, DC 6134 Indianapolis, IN 46285, U.S.A.*
[3]*Department of Biostatistics, The Rollins School of Public Health, Emory University, Atlanta, GA 30322, U.S.A.*

SUMMARY

In laboratory validation studies, it is often important to assess agreement between two assays, based on different techniques. Oftentimes, both assays have lower limits of detection and thus measurements are left censored. For example, in studies of Human Immunodeficiency Virus (HIV), the branched DNA (bDNA) assay was developed to quantify HIV-1 RNA concentrations in plasma. Validation of newer assays, such as the RT-PCR (reverse transcriptase polymerase chain reaction) involves assessing agreement of measurements obtained using the two techniques. Both bDNA and RT-PCR assays have lower limits of detection and thus new statistical methods are needed for assessing agreement between two left-censored variables. In this paper, we present maximum likelihood and generalized estimating equations approaches to evaluate agreement between two assays that are subject to lower limits of detection. The concordance correlation coefficient is used as an agreement index. The methodology is illustrated using HIV RNA assay data collected as part of a long-term HIV cohort study. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: agreement; assay validation; left censoring; limit of detection; generalized estimating equations

*Correspondence to: Huiman X. Barnhart, Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University, P.O. Box 17969, Durham, NC 27715, U.S.A.
†E-mail: huiman.barnhart@duke.edu
‡E-mail: songji@lilly.com
§E-mail: rlyles@sph.emory.edu

Contract/grant sponsor: National Institute of Allergy and Infectious Disease
Contract/grant sponsor: National Cancer Institute
Contract/grant sponsor: Emory University Quadrangle Fund
Contract/grant sponsor: NIH; contract/grant number: R01 MH070028-01A1

Received 2 June 2004
Accepted 24 February 2005

1. INTRODUCTION

In assay validation, two assays based on different techniques are often compared. Assay validation usually involves several fundamental parameters such as (1) accuracy, (2) precision, (3) selectivity, (4) sensitivity, (5) reproducibility, and (6) stability [1]. We focus on assessing agreement between two assays via parameters representing accuracy, precision and reproducibility. Although these parameters can be assessed separately, the concordance correlation coefficient [2, 3], is a popular agreement index for assessing accuracy and precision simultaneously that can also be used to assess reproducibility. In practice, assays often have lower limits of detection (LOD), denoting the lowest concentration of an analyte that the bioanalytical procedure can reliably differentiate from the background noise. Thus, the data collected are left censored. For example, in Human Immunodeficiency Virus (HIV) research, the branched DNA (bDNA) assay was developed to quantify HIV-1 RNA concentrations in plasma, prior to more recent assays based on the reverse transcriptase polymerase chain reaction (RT-PCR) technique. Both assays are subject to different lower limits of detection. To deal with left censored data, a naive (*ad hoc*) approach is to assign a constant value, say the lower limit or half of the lower limit, to subjects who have left censored data and to conduct the analysis as if we have complete data. This *ad hoc* approach is obviously biased and tends to produce larger mean estimates and smaller variance estimates than would have been obtained without lower limits of detection. Lyles *et al.* [4] proposed a maximum likelihood (ML) method based on normality assumptions to estimate the correlation coefficient between two left censored variables. Their method can be extended to estimate the concordance correlation coefficient.

In this paper, we describe the ML approach and present an alternative generalized estimating equations (GEE) approach to evaluate agreement between two assays that are both subject to lower limits of detection. In Section 2, the estimation and inference procedure for the concordance correlation coefficient are presented. A simulation study is conducted to compare the *ad hoc*, ML and GEE approaches in Section 3. In Section 4, we present an example from a HIV study for illustration. We conclude with a brief discussion in Section 5.

2. ASSESSING ASSAY AGREEMENT FOR LEFT CENSORED DATA

Let X and Y be continuous random variables representing the two assay readings on the same subject based on two different techniques. Due to lower limits of detection, we do not observe X and Y directly. Instead, we observe X_L and Y_L, the left-censored variables corresponding to X and Y with LODs L_x and L_y, where

$$X_L = \begin{cases} x & \text{if } X = x \geqslant L_x, \\ x_0 & \text{if } X = x < L_x, \end{cases} \qquad Y_L = \begin{cases} y & \text{if } Y = y \geqslant L_y \\ y_0 & \text{if } Y = y < L_y \end{cases}$$

with x_0 and y_0 as fixed constants. In practice, the value of LOD or half of LOD is typically used for x_0 and y_0. We assume that $(X, Y)'$ has a mean of $(\mu_x, \mu_y)'$ and a covariance

matrix of

$$\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

To assess the agreement of the two assay readings, we use the concordance correlation coefficient (CCC) [2] index expressed as

$$\rho_c = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} = \rho\chi^a \tag{1}$$

where ρ and χ^a are the precision (degree of variation) and accuracy (degree of location or scale shift) components of ρ_c, respectively. Note that ρ_c is a function of $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2)'$ and ρ. A naive estimate (the *ad hoc* method) for ρ_c is to replace $\boldsymbol{\theta}$ and ρ by using the sample means and sample covariance matrix based on the X_L's and Y_L's. However, because the X_L's and Y_L's are left censored versions of (X, Y), these sample estimates are biased for the mean and covariance matrix of (X, Y). An ML approach under normality assumptions [4] for estimating ρ in the presence of left censored data can be easily extended to estimate the concordance correlation coefficient if we insert the ML estimates for $\boldsymbol{\theta}$ and ρ into equation (1). Let $\{x_{Li}, y_{Li}\}, i = 1, \ldots, N$, be a random sample from random variables (X_L, Y_L). The ML estimates can be obtained by maximizing the observed data likelihood, based on N pairs of (X_L, Y_L)

$$L = \prod_{i=1}^{N} \left[\frac{1}{\sigma_{y|x_{Li}}\sigma_x} \phi\left(\frac{y_{Li} - \mu_{y|x_{Li}}}{\sigma_{y|x_{Li}}}\right) \phi\left(\frac{x_{Li} - \mu_x}{\sigma_x}\right) \right]^{d_{1i}} \left[\frac{1}{\sigma_x} \phi\left(\frac{x_{Li} - \mu_x}{\sigma_x}\right) \Phi\left(\frac{L_y - \mu_{y|x_{Li}}}{\sigma_{y|x_{Li}}}\right) \right]^{d_{2i}}$$

$$\times \left[\frac{1}{\sigma_y} \phi\left(\frac{y_{Li} - \mu_y}{\sigma_y}\right) \Phi\left(\frac{L_x - \mu_{x|y_{Li}}}{\sigma_{x|y_{Li}}}\right) \right]^{d_{3i}} \left[\int_{-\infty}^{L_y} \frac{1}{\sigma_y} \phi\left(\frac{y - \mu_y}{\sigma_y}\right) \Phi\left(\frac{L_x - \mu_{x|y}}{\sigma_{x|y}}\right) \mathrm{d}y \right]^{d_{4i}}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard univariate normal density and cumulative distribution functions, respectively, $\mu_{y|x} = \mu_y + (\rho\sigma_y/\sigma_x)(x - \mu_x), \sigma_{y|x} = \sigma_y\sqrt{1 - \rho^2}$, and $d_{1i}, d_{2i}, d_{3i}, d_{4i}$ are indicator variables for the following four conditions, respectively: (a) $x_{Li} \geqslant L_x$ and $y_{Li} \geqslant L_y$, (b) $x_{Li} \geqslant L_x$ and $Y_{Li} = y_0$, (c) $x_{Li} = x_0$ and $y_{Li} \geqslant L_y$, (d) $x_{Li} = x_0$ and $Y_{Li} = y_0$.

As an alternative, we now present two sets of generalized estimating equations for estimating the $\boldsymbol{\theta}$ and ρ parameters. We then construct the estimate for the agreement index ρ_c to assess agreement of two assays with lower limits of detection.

We first estimate the parameter $\boldsymbol{\theta} = (\mu_x, \mu_y, \sigma_x, \sigma_y)'$ by modelling the marginal mean of $\mathbf{Z}_i = (x_{Li}, y_{Li}, x_{Li}^2, y_{Li}^2)$ with $E(\mathbf{Z}_i) = \mathbf{U}(\boldsymbol{\theta})$ in the first set of estimating equations

$$\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{Z}_i - \mathbf{U}(\boldsymbol{\theta})) = \mathbf{0} \tag{2}$$

where $\mathbf{V}_i$ is the working covariance matrix for $\mathbf{Z}_i$, and the expressions for $\mathbf{U}(\boldsymbol{\theta}) = (\mathbf{U}[1], \mathbf{U}[2], \mathbf{U}[3], \mathbf{U}[4])'$ and the derivative matrix $\mathbf{D}_i = \partial\mathbf{U}/\partial\boldsymbol{\theta}$ are presented in the Appendix. The GEE approach uses empirical covariance estimates to adjust for a mis-specified covariance structure without the loss of much efficiency [5, 6]. For convenience, we take $\mathbf{V}_i$ as a diagonal matrix

with diagonal entries as the variances of X_L, Y_L, X^2 and Y^2 that we obtain under normality. Thus, we have

$$\mathbf{V}_i = \text{diag}(\mathbf{U}[3] - \mathbf{U}[1]^2, \mathbf{U}[4] - \mathbf{U}[2]^2, 2\sigma_x^4 + 4\mu_x^2\sigma_x^2, 2\sigma_y^4 + 4\mu_y^2\sigma_y^2)$$

as the working covariance matrix.

A second set of estimating equations based on modelling the conditional mean of $x_{Li}|y_{Li}$ is used to estimate the correlation parameter ρ. Note that the conditional distribution of $x_{Li}|y_{Li}$ is defined by

$$X_L|Y_L = \begin{cases} X_L|y & \text{if } Y = y \geqslant L_y \\ X_L|y_0 & \text{if } Y = y < L_y \end{cases}$$

with

$$X_L = \begin{cases} x & \text{if } X = x \geqslant L_x \\ x_0 & \text{if } X = x < L_x \end{cases}$$

Under normality assumptions, the distribution of $X|Y = y$ is $N(\mu_x + \rho\sigma_x(y-\mu_y)/\sigma_y, (1-\rho^2)\sigma_x^2)$. Let $v_y(Y) = (Y - \mu_y)/\sigma_y$. Thus, for $y_{Li} = y_i \geqslant L_y$, we have $\gamma_i(\rho, \boldsymbol{\theta}) = E(X_{Li}|Y_{Li} = y_{Li}) = \int_{L_x}^{\infty} x f(x|y_{Li})\,\mathrm{d}x + x_0 \int_{-\infty}^{L_x} f(x|y_{Li})\,\mathrm{d}x = x_0\Phi(\omega_{xy}(y_{Li})) + (\mu_x + \rho\sigma_x v_y(y_{Li}))\Phi(-\omega_{xy}(y_{Li})) + \sigma_x\sqrt{1-\rho^2}\,\phi(\omega_{xy}(y_{Li}))$, where $\omega_{xy}(Y) = \tau_x/\sqrt{1-\rho^2} - \rho v_y(Y)/\sqrt{1-\rho^2}$ with $\tau_x = (L_x - \mu_x)/\sigma_x$ (see Appendix for details). For simplicity and ease of computation, we use the same expression for $\gamma_i(\rho, \boldsymbol{\theta})$ with $y_{Li} = y_0$ to approximate $E(X_{Li}|Y_{Li} = y_0) = E(X_L|Y < L_y)$ (see Appendix for the exact expression of $E(X_L|Y < L_y)$). We show that this approximation is reasonable in the simulation study and the example if x_0 and y_0 are chosen to satisfy $E(X) = E(X_L)$ and $E(Y) = E(Y_L)$. Our approach will yield asymptotically unbiased parameter estimates for any choices of x_0 and y_0 if we do not use the above approximation. However, specifying x_0 and y_0 based on the conditions $E(X) = E(X_L)$ and $E(Y) = E(Y_L)$ generally tends to provide better efficiency and 95 per cent coverage for ρ [7].

We solve for ρ by using the following second set of estimating equations:

$$\sum_{i=1}^{N} C_i' W_i^{-1}(x_{Li} - \gamma_i(\rho, \boldsymbol{\theta})) = 0 \tag{3}$$

where

$$C_i = \frac{\partial\gamma_i}{\partial\rho} = \sigma_x v_y(y_{Li})\Phi(-\omega_{xy}(y_{Li})) - \frac{\partial\omega_{xy}(y_{Li})}{\partial\rho}(L_x - x_0)\phi(\omega_{xy}(y_{Li})) - \frac{\rho}{\sqrt{1-\rho^2}}\sigma_x\phi(\omega_{xy}(y_{Li}))$$

with

$$\frac{\partial\omega_{xy}(y_{Li})}{\partial\rho} = \frac{\rho}{(1-\rho^2)^{3/2}}\tau_x - \frac{1}{(1-\rho^2)^{3/2}}v_y(y_{Li})$$

Here, W_i is the working variance of $x_{Li}|y_{Li}$. Again, because the GEE method is robust to misspecification of the working variance, we use $W_i = \mathrm{var}(X|Y) = (1-\rho^2)\sigma_x^2$, as obtained under normality for simplicity.

To obtain the point estimates of $\boldsymbol{\theta}$ and ρ, a modified Fisher-scoring iterative procedure is used. Specifically, we obtain the estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, by the iteration process,

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} + \left(\sum_{i=1}^{N} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1} \sum_{i=1}^{N} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} (\mathbf{Z}_i - \mathbf{U}(\hat{\boldsymbol{\theta}}^{(t)}))$$

By replacing $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}$ in the second set of estimating equations (3), the estimate $\hat{\rho}$ of ρ can be obtained by the iteration process,

$$\hat{\rho}^{(t+1)} = \hat{\rho}^{(t)} + \left(\sum_{i=1}^{N} \hat{C}_i' \hat{W}_i^{-1} \hat{C}_i\right)^{-1} \sum_{i=1}^{N} \hat{C}_i' \hat{W}_i^{-1} (x_{Li} - \gamma_i(\hat{\rho}^{(t)}, \hat{\boldsymbol{\theta}}))$$

Following similar arguments used for the generalized estimating equations [8], we can show that the parameter estimates are consistent provided that $\mathbf{U}(\boldsymbol{\theta})$ is correctly specified and $\gamma_i(\rho, \boldsymbol{\theta})$ is a good approximation to $E(X_{Li}|Y_{Li} = y_0)$. This is true whether or not the working covariance matrices in the two sets of equations are correctly specified.

Following Prentice [9] and Barnhart and Williamson [10], we can show that the joint asymptotic distribution of $N^{1/2}((\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), (\hat{\rho} - \rho))'$ is multivariate normally distributed with mean $\mathbf{0}$ and variance matrix N times $\mathbf{B}$, where

$$\begin{aligned} \mathbf{B} &= \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Psi}'^{-1} = \boldsymbol{\Psi}^{-1} \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \Lambda_{22} \end{pmatrix} \boldsymbol{\Psi}'^{-1} \\ \boldsymbol{\Psi} &= \begin{pmatrix} \sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i & \mathbf{0} \\ \sum_{i=1}^{N} C_i' W_i^{-1} \mathbf{G}_i & \sum_{i=1}^{N} C_i' W_i^{-1} C_i \end{pmatrix} \end{aligned} \tag{4}$$

and

$$\mathbf{G}_i = \frac{\partial \gamma_i}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \gamma_i}{\partial \mu_x}, \quad \frac{\partial \gamma_i}{\partial \mu_y}, \quad \frac{\partial \gamma_i}{\partial \sigma_x}, \quad \frac{\partial \gamma_i}{\partial \sigma_y}\right)' \quad \text{with}$$

$$\frac{\partial \gamma_i}{\partial \mu_x} = \Phi(-\omega_{xy}(y_{Li})) - \frac{(x_0 - L_x)}{(\sigma_x \sqrt{1-\rho^2})} \phi(\omega_{xy}(y_{Li}))$$

$$\frac{\partial \gamma_i}{\partial \mu_y} = \rho \frac{\sigma_x}{\sigma_y} \Phi(-\omega_{xy}(y_{Li})) + \frac{\rho(x_0 - L_x)}{(\sigma_y \sqrt{1-\rho^2})} \phi(\omega_{xy}(y_{Li}))$$

$$\frac{\partial \gamma_i}{\partial \sigma_x} = \rho v_y(y_{Li}) \Phi(-\omega_{xy}(y_{Li})) + \sqrt{1-\rho^2} \phi(\omega_{xy}(y_{Li})) - \frac{\tau_x(x_0 - L_x)}{(\sigma_x \sqrt{1-\rho^2})} \phi(\omega_{xy}(y_{Li}))$$

and

$$\frac{\partial \gamma_i}{\partial \sigma_y} = -\frac{\sigma_x}{\sigma_y} \rho v_y(y_{Li}) \Phi(-\omega_{xy}(y_{Li})) + \frac{\rho(x_0 - L_x)}{(\sigma_y \sqrt{1-\rho^2})} v_y(y_{Li}) \phi(\omega_{xy}(y_{Li}))$$

The estimates of the elements in matrix $\boldsymbol{\Lambda}$ are as follows:

$$\hat{\boldsymbol{\Lambda}}_{11} = \sum_{i=1}^{N} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} (\mathbf{Z}_i - \mathbf{U}(\hat{\boldsymbol{\theta}}))(\mathbf{Z}_i - \mathbf{U}(\hat{\boldsymbol{\theta}}))' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i$$

$$\hat{\boldsymbol{\Lambda}}_{12} = \sum_{i=1}^{N} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} (\mathbf{Z}_i - \mathbf{U}(\hat{\boldsymbol{\theta}}))(x_{Li} - \gamma_i(\hat{\rho}, \hat{\boldsymbol{\theta}}))' \hat{W}_i^{-1} \hat{C}_i$$

$$\hat{\Lambda}_{22} = \sum_{i=1}^{N} \hat{C}_i' \hat{W}_i^{-1} (x_{Li} - \gamma_i(\hat{\rho}, \hat{\boldsymbol{\theta}}))(x_{Li} - \gamma_i(\hat{\rho}, \hat{\boldsymbol{\theta}}))' \hat{W}_i^{-1} \hat{C}_i$$

$$\hat{\boldsymbol{\Lambda}}_{12} = \hat{\boldsymbol{\Lambda}}_{21}'$$

We refer to $\hat{\mathbf{B}}$ as the empirically corrected estimate of the variance–covariance matrix of $\hat{\boldsymbol{\theta}}$ and $\hat{\rho}$.

To estimate the agreement index ρ_c, we insert the estimates of $\boldsymbol{\theta}$ and ρ in the definition for ρ_c (equation (1)). The delta method is used to estimate the standard error of $\hat{\rho}_c$ based on the empirically corrected estimate of the covariance matrix for $\hat{\boldsymbol{\theta}}$ and $\hat{\rho}$.

3. SIMULATION

To compare the *ad hoc*, ML and GEE approaches, we performed a simulation study. Bivariate normal data with sample size of 100 were generated using one of the following six combinations of parameter settings: $\mu_x = 0, \mu_y = 0.2, \sigma_x = 0.8, \sigma_y = 1, \rho = 0.25, 0.50, 0.75$, and left censoring rate of (25 per cent, 25 per cent) or (40 per cent, 25 per cent). True values of L_x and L_y are determined by the censoring rates. Results based on 1000 simulated data sets are reported in Table I. We used the formula in Reference [2] to compute the SE for the *ad hoc* method. To check on the approximation of $\gamma_i(\rho, \boldsymbol{\theta})$ to $E(X_L | Y < L_Y)$ when $y_{Li} = y_0$, we find that the absolute differences between these two quantities ranged from 0.00314 to 0.04494 under these six parameter settings. The simulation results from the GEE approach in Table I also support that the approximation is reasonable.

The *ad hoc* estimates are obviously biased. In general, both the ML and GEE methods perform well but tend to slightly underestimate the true value. The GEE method performs slightly better than the ML method when one compares the mean standard error (SE) to the empirical standard deviation (SD) based on the 1000 estimates. The mean SE from the GEE agrees well with the empirical SD while the mean SE from ML is slightly larger than the empirical SD. This resulted in a slightly better 95 per cent coverage for the ML method than the GEE method due to the slight parameter underestimation in both methods. As suggested in Barnhart *et al.* [11], the 95 per cent coverage for the GEE method is improved when the SE is multiplied by a factor of $\sqrt{N/(N-2)}$.

Table I. Simulation results based on 1000 data sets with sample size of 100.

Per cent censoring	True ρ	True ρ_c	Method	Mean $\hat{\rho}_c$	Empirical SD	Mean SE	95 per cent coverage (*)
(25 per cent, 25 per cent)	0.25	0.238	ML	0.233	0.092	0.094	0.943
			GEE	0.233	0.093	0.093	0.935 (0.940)
			Ad Hoc	0.195	0.092	0.087	0.901
	0.50	0.476	ML	0.468	0.077	0.079	0.951
			GEE	0.464	0.078	0.078	0.941 (0.941)
			Ad Hoc	0.406	0.083	0.076	0.852
	0.75	0.714	ML	0.706	0.050	0.053	0.963
			GEE	0.692	0.052	0.052	0.946 (0.949)
			Ad Hoc	0.640	0.059	0.053	0.741
(40 per cent, 25 per cent)	0.25	0.238	ML	0.232	0.095	0.098	0.939
			GEE	0.232	0.095	0.096	0.939 (0.940)
			Ad Hoc	0.189	0.093	0.086	0.893
	0.50	0.476	ML	0.467	0.079	0.083	0.952
			GEE	0.463	0.080	0.081	0.945 (0.949)
			Ad Hoc	0.396	0.085	0.075	0.815
	0.75	0.714	ML	0.705	0.052	0.056	0.970
			GEE	0.696	0.055	0.055	0.953 (0.956)
			Ad Hoc	0.628	0.062	0.052	0.648

*The number in parentheses is the adjusted 95 per cent coverage where the standard error is multiplied by a factor of $\sqrt{N/(N-2)}$.

4. EXAMPLE

We use an HIV example to illustrate the use of the ML and GEE methods to assess agreement for two assays with lower limits of detection. The data set came from the Multicenter AIDS Cohort Study (MACS). It was originally analysed by Mellors *et al.* [12] for a quality control analysis of the bDNA assay methodology. The HIV-1 RNA concentration in plasma was measured using the bDNA signal-amplification assay. The LOD of this assay is 500 copies/mL. To perform quality control of the assay technology, a random sample of size 300 was obtained from the original data and the plasma HIV-1 RNA concentration was also measured by the RT-PCR assay (LOD = 400 copies/mL). The purpose of studying this sub-data is to examine the agreement between readings produced by the two assay technologies. The bDNA data contained 17 undetectable readings, implying a 5.6 per cent censoring rate. The RT-PCR contained 4 undetectable readings, for a 1.33 per cent censoring rate. A natural log transformation was applied to both the bDNA and RT-PCR readings. We treat log(RT-PCR) as the X variable and log(bDNA) as the Y variable.

Figure 1 shows histograms and Q–Q plots of the log-transformed measurements for the observed log(RT-PCR) and log(bDNA) data where subjects with censored data are excluded. In the Q–Q plots, expected quantiles were obtained based on the appropriate truncated normal distributions. These plots suggest that both log(RT-PCR) and log(bDNA) are approximately normal. Figure 2 displays the scatter plot for log(bDNA) *versus* log(RT-PCR). The cloud of points is shifted to the right, indicating that readings from RT-PCR are systematically higher

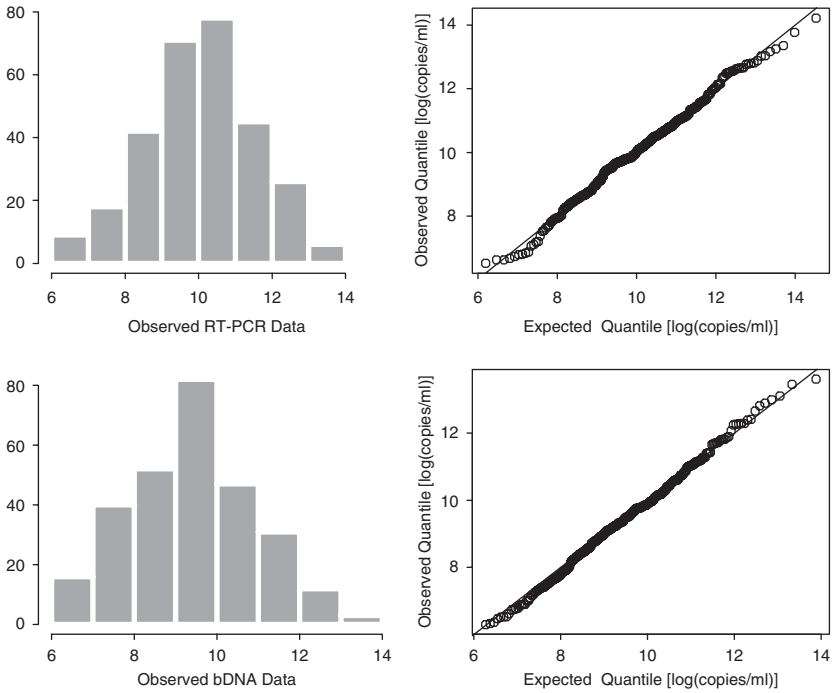

Figure 1. Histograms and Q–Q plots for log transformed RT-PCR (top row) and bDNA (bottom row). Expected quantiles based on a truncated normal distribution with mean and variance equal to their MLEs from the truncated normal model.

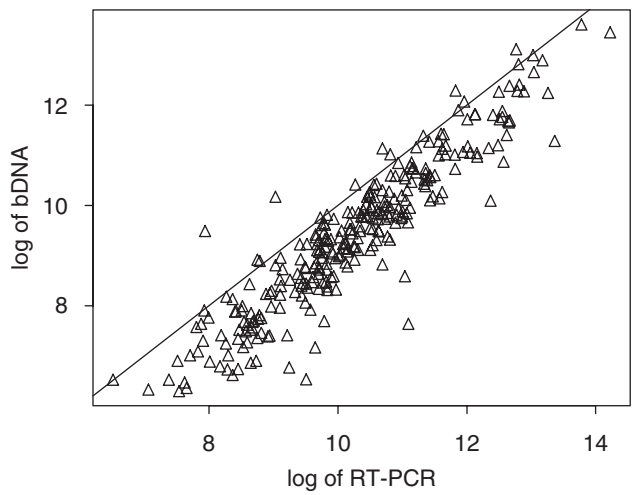

Figure 2. Scatter plot of observed log RT-PCR *versus* log bDNA data.

Table II. Agreement results for comparing log-transformed HIV RNA readings based on bDNA and RT-PCR assays.

	LOD = (400, 500) copies/mL				LOD = (2000, 2000) copies/mL			
Method	$\hat{\rho}^c$	95 per cent *CI*	$\hat{\rho}$	$\hat{\chi}^a$	$\hat{\rho}^c$	95 per cent *CI*	$\hat{\rho}$	$\hat{\chi}^a$
ML	0.831	(0.764, 0.881)	0.942	0.882	0.830	(0.759, 0.881)	0.943	0.880
GEE	0.850	(0.808, 0.884)	0.962	0.884	0.851	(0.804, 0.888)	0.962	0.885
*Ad hoc**	0.795	(0.761, 0.829)	0.910	0.874	0.740	(0.697, 0.784)	0.846	0.875

*Assigned $\frac{1}{2}$ of the log(LOD) to censored data.

than readings from bDNA. Data points are clustered, indicating a strong correlation between the two readings.

Table II summarizes results based on the ML method, the GEE method and the *ad hoc* method, for estimating ρ_c, ρ and χ^a when comparing log(bDNA) to log(RT-PCR). In the *ad hoc* approach, we replaced non-detects by $x_0 = \frac{1}{2}\log(X_L)$ and $y_0 = \frac{1}{2}\log(Y_L)$. Due to the low censoring rate of both assays, the GEE, ML and *ad hoc* estimates of CCC are not too far apart (0.850, 0.831 and 0.795, respectively). However, if the censoring rate were high, the *ad hoc* method would be likely to produce a much less reasonable estimate. To illustrate this point, we reconducted the analyses assuming that the LOD for both assays was 2000 copies/mL. In this case, the censoring rate for RT-PCR is 6.7 per cent and the censoring rate for bDNA is 17.7 per cent. We note that the GEE and ML estimates are little changed when the LODs are altered from (400, 500) copies/mL to (2000, 2000) copies/mL. However, the *ad hoc* estimate changed from 0.791 to 0.740.

As this example suggests, the *ad hoc* approach can lead to severe bias when the amount of left censoring is non-negligible. Standard error estimates are also invalid when applying the *ad hoc* method.

The GEE method produced a slightly higher estimate of ρ_c than the ML method (0.850 *versus* 0.831). The precision (accuracy) components were estimated as 0.962 (0.884) and 0.942 (0.882) for the GEE and ML methods, respectively. This implies that the bDNA reading is highly correlated with the RT-PCR reading, while overall agreement is slightly lower due to the moderate to high values on accuracy. The original analysis of the relationship of bDNA and RT-PCR [12] was based upon a sample size of 400 instead of 300, where the additional 100 subjects came from a random sample from readings below 3000 copies/mL. By design, this was therefore not a random sample from the MACS data. Mellors *et al.* [12] reported the sample correlation coefficient to be 0.93 between log(bDNA) and log(RT-PCR) when 400 copies/mL was assigned to subjects with censored bDNA readings (below 500 copies/mL) and 300 copies/mL was assigned to subjects with censored RT-PCR readings (below 400 copies/mL). Based on the random sample of 300 subjects, the sample correlation coefficient is 0.94. Note that this is larger than the *ad hoc* result in Table II because larger values of $x_0 = \log(300)$ and $y_0 = \log(400)$ were used in the *ad hoc* approach as employed by Mellors *et al.* [12]. Due to the high correlation, Mellors *et al.* [12] suggested a linear relationship between log(RT-PCR) and log(bDNA) and expressed the estimated relationship on the original scale as: RT-PCR (copies/mL) $= 5.13 \times$ (bDNA) (copies/mL)$^{0.9}$. If we use the random sample of 300 subjects, this estimated relationship (with non-detects replaced by $x_0 = \log(300)$ and $y_0 = \log(400)$) becomes

RT-PCR (copies/mL) $= 7.04 \times (\text{bDNA})\ (\text{copies/mL})^{0.88}$. For comparison, we assume that there is a linear relationship between log(RT-PCR) and log(bDNA). Our GEE results with $\hat{\boldsymbol{\theta}} = (\hat{\mu}_x, \hat{\mu}_y, \hat{\sigma}_x, \hat{\sigma}_y) = (9.207, 10.039, 1.735, 1.574)$ and $\hat{\rho} = 0.962$ imply that RT-PCR (copies/mL) $= 7.42 \times (\text{bDNAcopies/mL})^{0.87}$ where $7.42 = \exp(\hat{\mu}_y - \hat{\mu}_x \hat{\rho} \hat{\sigma}_y / \hat{\sigma}_x)$ and $0.87 = \hat{\rho} \hat{\sigma}_y / \hat{\sigma}_x$. We conclude that the bDNA and RT-PCR measurements agree up to location and scale shifts, where the bDNA assay has consistently lower average readings than the ones from the RT-PCR assay.

To check on the approximation of $\gamma_i(\rho, \boldsymbol{\theta})$ to $E(X_L | Y < L_y)$ when $y_{Li} = y_0$, we found that the differences between these two quantities are 0.00571 and 0.00697, respectively, for the two LOD settings, when these two quantities are evaluated at the GEE estimates.

5. DISCUSSION

We have proposed a GEE approach to estimate the concordance correlation coefficient for left censored data that often occur in assay validation. Our example showed that the GEE approach works well and was comparable to the maximum likelihood approach based on the bivariate normality assumption.

In the proposed GEE approach, we used the conditional mean in the second set of estimating equations. A straightforward GEE would be using $\mathbf{Z}_i = (x_{L_i}, y_{L_i}, x_{L_i}^2, y_{L_i}^2, x_{L_i} y_{L_i})'$. This approach, termed as GEE(p), was investigated in Dr Song's dissertation research [7] where the expectation of a product term instead of a conditional mean is considered. We found that the GEE(p) approach is more computationally intensive than the conditional GEE approach because GEE(p) requires evaluation of $E(x_{L_i} y_{L_i})$ involving double integration. In terms of performance, both the GEE(p) and conditional GEE methods performed similarly except that the GEE(p) approach may be more stable when censoring rates in both variables are higher than 40 per cent. Thus, the conditional GEE approach is recommended because the censoring rate is unlikely to be higher than 40 per cent in both variables in practice.

We have used two stage GEE in parameter estimation. This is equivalent to one set of estimating equations by combining the proposed two sets of equations together with a block diagonal working covariance matrix. If general working covariance matrix is used, then misspecification from one set of estimating equations may affect consistent estimation of parameters from the other set of equations.

In using the proposed GEE method, one would need to decide which of the two variables should be used as the X variable. This is not a choice when the data were collected, but a choice at the time of parameter estimation when the GEE method is used. It should not affect consistency if any one of the two variables is used as the X variable. Intuitively we felt that it is more efficient to use the variable with lower censoring rate as the X variable.

For stability, a Fisher's Z transformation may be used on ρ in parameter estimation. In addition, the Fisher's Z transformation can also be applied to ρ_c to improve 95 per cent coverage [2]. This is especially useful when the true value of ρ or ρ_c is close to the boundary value (-1 or 1),

The computing programs from this paper are available from Jingli Song upon request.

APPENDIX

Let $\mathbf{Z}_i = (x_{L_i}, y_{L_i}, x_{L_i}^2, y_{L_i}^2)'$. Note that with $\tau_x = (L_x - \mu_x)/\sigma_x$ and $\tau_y = (L_y - \mu_y)/\sigma_y$

$$\begin{aligned}
E(x_{L_i}) &= \int_{-\infty}^{L_x} x_0 \frac{1}{\sqrt{2\pi}\sigma_x} \mathrm{e}^{-(x-\mu_x)^2/2\sigma_x^2}\,\mathrm{d}x + \int_{L_x}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma_x} \mathrm{e}^{-(x-\mu_x)^2/2\sigma_x^2}\,\mathrm{d}x \\
&= x_0 \int_{-\infty}^{\tau_x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}\,\mathrm{d}t + \int_{\tau_x}^{\infty} (\mu_x + t\sigma_x) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}\,\mathrm{d}t \\
&= x_0 \Phi(\tau_x) + \mu_x \Phi(-\tau_x) + \sigma_x \phi(\tau_x)
\end{aligned}$$

and

$$\begin{aligned}
E(x_{L_i}^2) &= \int_{-\infty}^{L_x} x_0^2 \frac{1}{\sqrt{2\pi}\sigma_x} \mathrm{e}^{-(x-\mu_x)^2/2\sigma_x^2}\,\mathrm{d}x + \int_{L_x}^{\infty} x^2 \frac{1}{\sqrt{2\pi}\sigma_x} \mathrm{e}^{-(x-\mu_x)^2/2\sigma_x^2}\,\mathrm{d}x \\
&= x_0^2 \int_{-\infty}^{\tau_x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}\,\mathrm{d}t + \int_{\tau_x}^{\infty} (\mu_x + t\sigma_x)^2 \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}\,\mathrm{d}t \\
&= x_0^2 \Phi(\tau_x) + \mu_x^2 \Phi(-\tau_x) + \int_{\tau_x}^{\infty} (2\mu_x\sigma_x t + t^2\sigma_x^2) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}\,\mathrm{d}t \\
&= x_0^2 \Phi(\tau_x) + \mu_x^2 \Phi(-\tau_x) + 2\mu_x\sigma_x \phi(\tau_x) + \sigma_x^2(\tau_x \phi(\tau_x) + \Phi(-\tau_x)) \\
&= x_0^2 \Phi(\tau_x) + (\mu_x^2 + \sigma_x^2)\Phi(-\tau_x) + (\sigma_x^2 \tau_x + 2\sigma_x\mu_x)\phi(\tau_x) \\
&= x_0^2 \Phi(\tau_x) + (\mu_x^2 + \sigma_x^2)\Phi(-\tau_x) + (L_x + \mu_x)\sigma_x \phi(\tau_x)
\end{aligned}$$

Thus,

$$\mathbf{U} = E(\mathbf{Z}_i) = \begin{pmatrix}
x_0 \Phi(\tau_x) + \mu_x \Phi(-\tau_x) + \sigma_x \phi(\tau_x) \\
y_0 \Phi(\tau_y) + \mu_y \Phi(-\tau_y) + \sigma_y \phi(\tau_y) \\
x_0^2 \Phi(\tau_x) + (\mu_x^2 + \sigma_x^2)\Phi(-\tau_x) + (L_x + \mu_x)\sigma_x \phi(\tau_x) \\
y_0^2 \Phi(\tau_y) + (\mu_y^2 + \sigma_y^2)\Phi(-\tau_y) + (L_y + \mu_y)\sigma_y \phi(\tau_y)
\end{pmatrix}$$

and

$$\mathbf{D}_i = \frac{\partial \mathbf{U}}{\partial \boldsymbol{\theta}} = \begin{pmatrix}
d_{11} & 0 & d_{13} & 0 \\
0 & d_{22} & 0 & d_{24} \\
d_{31} & 0 & d_{33} & 0 \\
0 & d_{42} & 0 & d_{44}
\end{pmatrix}$$

where

$$d_{11} = \Phi(-\tau_x) + (L_x - x_0)\phi(\tau_x)/\sigma_x$$

$$d_{13} = \phi(\tau_x)[\tau_x(L_x - x_0)/\sigma_x + 1]$$

$$d_{22} = \Phi(-\tau_y) + (L_y - y_0)\phi(\tau_y)/\sigma_y$$

$$d_{24} = \phi(\tau_y)[\tau_y(L_y - y_0)/\sigma_y + 1]$$

$$d_{31} = 2\mu_x\Phi(-\tau_x) + \phi(\tau_x)(L_x^2 + 2\sigma_x^2 - x_0^2)/\sigma_x$$

$$d_{33} = 2\sigma_x\Phi(-\tau_x) + \phi(\tau_x)[(L_x + \mu_x)(1 + \tau_x^2) + \tau_x(\mu_x^2 + \sigma_x^2 - x_0^2)/\sigma_x]$$

$$d_{42} = 2\mu_y\Phi(-\tau_y) + \phi(\tau_y)(L_y^2 + 2\sigma_y^2 - y_0^2)/\sigma_y$$

$$d_{44} = 2\sigma_y\Phi(-\tau_y) + \phi(\tau_y)[(L_y + \mu_y)(1 + \tau_y^2) + \tau_y(\mu_y^2 + \sigma_y^2 - y_0^2)/\sigma_y]$$

Derivation of $E(X_{Li}|Y_{Li} = y_{Li})$:
If $y_{Li} \geqslant L_y$ then

$$\gamma_i(\rho, \boldsymbol{\theta}) = E(X_{Li}|Y_{Li} = y_{Li}) = \int_{L_x}^{\infty} x f(x|y_{Li})\,\mathrm{d}x + x_0 \int_{-\infty}^{L_x} f(x|y_{Li})\,\mathrm{d}x$$

$$= \int_{L_x}^{\infty} \frac{x}{\sqrt{2\pi}\sigma_x\sqrt{1-\rho^2}} \mathrm{e}^{-1/2\left(\frac{1}{\sqrt{1-\rho^2}}\frac{x-\mu_x}{\sigma_x} - \frac{\rho}{\sqrt{1-\rho^2}}\frac{y_{L_i}-\mu_y}{\sigma_y}\right)^2} \mathrm{d}x$$

$$+x_0 \int_{-\infty}^{L_x} \frac{1}{\sqrt{2\pi}\sigma_x\sqrt{1-\rho^2}} \mathrm{e}^{-1/2\left(\frac{1}{\sqrt{1-\rho^2}}\frac{x-\mu_x}{\sigma_x} - \frac{\rho}{\sqrt{1-\rho^2}}\frac{y_{L_i}-\mu_y}{\sigma_y}\right)^2} \mathrm{d}x$$

$$= \int_{\omega_{xy}(y_{L_i})}^{\infty} (\mu_x + \rho\sigma_x v_y(y_{L_i}) + s\sigma_x\sqrt{1-\rho^2}) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(1/2)s^2}\,\mathrm{d}s + x_0 \int_{-\infty}^{\omega_{xy}(y_{L_i})} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(1/2)s^2}\,\mathrm{d}s$$

$$= x_0\Phi(\omega_{xy}(y_{Li})) + (\mu_x + \rho\sigma_x v_y(y_{Li}))\Phi(-\omega_{xy}(y_{Li})) + \sigma_x\sqrt{1-\rho^2}\phi(\omega_{xy}(y_{Li}))$$

where $\omega_{xy}(Y) = \tau_x/\sqrt{1-\rho^2} - \rho v_y(Y)/\sqrt{1-\rho^2}$ with $v_y(Y) = (Y - \mu_y)/\sigma_y$ and $\tau_x = (L_x - \mu_x)/\sigma_x$.
If $y_{Li} = y_0$, i.e. $y_i < L_y$ then

$$\gamma_i(\rho, \boldsymbol{\theta}) = E(X_{Li}|Y_{Li} = y_0) = E(X_{Li}|y_i < L_y)$$

$$= \int_{L_x}^{\infty} x f(x|y_i < L_y)\,\mathrm{d}x + x_0 \int_{-\infty}^{L_x} f(x|y_i < L_y)\,\mathrm{d}x$$

$$= \int_{L_x}^{\infty} x \frac{f(x, y_i < L_y)}{Pr(y_i < L_y)}\,\mathrm{d}x + x_0 \int_{-\infty}^{L_x} \frac{f(x, y_i < L_y)}{Pr(y_i < L_y)}\,\mathrm{d}x$$

$$= \frac{1}{\Phi(\tau_y)} \left[\int_{L_x}^{\infty} xPr(y_i < L_y|x) f(x)\,\mathrm{d}x + x_0 \int_{-\infty}^{L_x} Pr(y_i < L_y|x) f(x)\,\mathrm{d}x \right]$$

$$= \frac{1}{\Phi(\tau_y)} \left[\int_{L_x}^{\infty} x \int_{-\infty}^{L_y} f(y|x)\,\mathrm{d}y f(x)\,\mathrm{d}x + x_0 \int_{-\infty}^{L_x} \int_{-\infty}^{L_y} f(y|x)\,\mathrm{d}y f(x)\,\mathrm{d}x \right]$$

$$= \frac{1}{\Phi(\tau_y)} \left[\int_{-\infty}^{L_y} \int_{L_x}^{\infty} x f(y|x) f(x)\,\mathrm{d}x\,\mathrm{d}y + x_0 \int_{-\infty}^{L_y} \int_{-\infty}^{L_x} f(y|x) f(x)\,\mathrm{d}x\,\mathrm{d}y \right]$$

$$= \frac{1}{\Phi(\tau_y)} \int_{-\infty}^{L_y} \left[\int_{L_x}^{\infty} x f(x, y)\,\mathrm{d}x + x_0 \int_{-\infty}^{L_x} f(x, y)\,\mathrm{d}x \right] \mathrm{d}y$$

$$= \frac{1}{\Phi(\tau_y)} \int_{-\infty}^{L_y} f(y) \left[\int_{L_x}^{\infty} x f(x|y)\,\mathrm{d}x + x_0 \int_{-\infty}^{L_x} f(x|y)\,\mathrm{d}x \right] \mathrm{d}y$$

$$= \frac{1}{\Phi(\tau_y)} \int_{-\infty}^{L_y} [x_0 \Phi(\omega_{xy}(y)) + (\mu_x + \rho\sigma_x v_y(y)) \Phi(-\omega_{xy}(y))$$

$$+ \sigma_x \sqrt{1-\rho^2}\, \phi(\omega_{xy}(y))] f(y)\,\mathrm{d}y$$

$$\approx x_0 \Phi(\omega_{xy}(y_0)) + (\mu_x + \rho\sigma_x v_y(y_0)) \Phi(-\omega_{xy}(y_0)) + \sigma_x \sqrt{1-\rho^2}\, \phi(\omega_{xy}(y_0))$$

where x_0 and y_0 are chosen to satisfy $E(X_L) = E(X)$ and $E(Y_L) = E(Y)$.

ACKNOWLEDGEMENTS

We would like to thank Drs Alvaro Muñoz and Stephen Gange from Johns Hopkins University, as well as the other investigators and participants of the MACS study, for providing the HIV-1 RNA concentration data based on bDNA and RT-PCR assays. The MACS study was funded by the National Institute of Allergy and Infectious Disease with additional supplemental funding from the National Cancer Institute. This research was supported in part by Emory University Quadrangle Fund and by NIH Grant R01 MH070028-01A1. We appreciate two anonymous reviewers' comments that lead to an improved manuscript.

REFERENCES

1. Food and Drug Administration. *Guidance for Industry*: *Bioanalytical Method Validation*. http://www.fda.gov/cder/guidance/index.htm, 2001.
2. Lin L. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**:255–268.
3. Lin L. Assay validation using the concordance correlation coefficient. *Biometrics* 1992; **48**:599–604.
4. Lyles RH, Williams JK, Chuachoowong R. Correlating two viral load assay with known detection limits. *Biometrics* 2001; **57**:1238–1244.
5. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
6. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; **42**: 121–130.
7. Song J. Assessing agreement/association for continuous measurement scales. *Ph.D. Dissertation*, Department of Biostatistics, Emory University, Atlanta, GA, 2003.
8. Liang KY, Zeger SL, Qaqish B. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society Series B-Methodological* 1992; **54**:3–24.

9. Prentice RL. Correlated binary regression with covariates specific to each binary observation. *Biometrics* 1988; **44**:1033–1048.
10. Barnhart HX, Williamson JM. Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics* 2001; **57**:931–940.
11. Barnhart HX, Haber M, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* 2002; **58**:1020–1027.
12. Mellors JW, Muñoz A, Giorgi JV, Margolick JB, Tassoni CJ, Gupta P, Kingsley LA, Todd JA, Saah AJ, Detels R, Phair JP, Rinaldo Jr CR. Plasma viral load and CD4(+ lymphocytes as prognostic markers of HIV-1 infection. *Annals of Internal Medicine* 1997; **126**:946–954.

