Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Estimating agreement coefficients from sample survey data

by Hung-Mo Lin, Hae-Young Kim, John M. Williamson
and Virginia M. Lesser

June 2012

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 63-72
Statistics Canada, Catalogue No. 12-001-X

Estimating agreement coefficients from sample survey data

Hung-Mo Lin, Hae-Young Kim, John M. Williamson and Virginia M. Lesser [1]

Abstract

We present a generalized estimating equations approach for estimating the concordance correlation coefficient and the kappa coefficient from sample survey data. The estimates and their accompanying standard error need to correctly account for the sampling design. Weighted measures of the concordance correlation coefficient and the kappa coefficient, along with the variance of these measures accounting for the sampling design, are presented. We use the Taylor series linearization method and the jackknife procedure for estimating the standard errors of the resulting parameter estimates. Body measurement and oral health data from the Third National Health and Nutrition Examination Survey are used to illustrate this methodology.

Key Words: Clustering; Concordance correlation coefficient; Generalized estimating equations; Jackknife estimator; Kappa coefficient; Sample weighting; Stratification; Taylor series linearization.

1. Introduction

Surveys often collect multiple measures of latent conditions such as quality of life and aspiration for a college education, as well as multiple measures of difficult- to-classify conditions such as having chronic fatigue syndrome. When multiple measures are collected, interest naturally focuses on the agreement between the multiple measures and in obtaining confidence intervals on those agreement measures. Also, there may be interest in contrasting agreement across population subgroups and across alternate pairings of measurements. In this context, one might be interested in testing equality of agreement measures. This paper focuses on two measures of agreement between such multiple measures, the concordance correlation coefficient (CCC, ρ_c) and the kappa (κ) coefficient. The former is useful for continuous measurements with natural scales. If a measure of a latent concept has no natural scale, then it can be arbitrarily rescaled to have mean zero and unit variance. When this is possible, it is meaningless to talk about differences in marginal moments. However, if there is a natural scale, then rescaling is not desirable and a good measure of agreement will take into account both correlation and agreement of marginal moments. The kappa coefficient is most useful for binary classifications.

The CCC has been shown to be more appropriate for measuring agreement or reproducibility (Lin 1989; Lin 1992) than the Pearson correlation coefficient (ρ). It evaluates the accuracy between two readings by measuring the variation of the fitted linear relationship from the 45^0 line through the origin (the concordance line) and precision by measuring how far each observation deviates from the fitted line. Let Y_{i1} and Y_{i2} denote a pair of continuous random variables measured on the same subject i using two methods. The CCC for measuring the agreement of Y_{i1} and Y_{i2} is defined as follows:

$$\rho_c = 1 - \frac{E[(Y_{i1} - Y_{i2})^2]}{E_{\text{indep}}[(Y_{i1} - Y_{i2})^2]} = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \quad (1)$$

where $\sigma_1^2 = \text{var}(Y_{i1})$, $\sigma_2^2 = \text{var}(Y_{i2})$, and $\sigma_{12} = \text{cov}(Y_{i1}, Y_{i2})$ (Lin 1989). As noted by Lin (1989), $\rho_c = 0$ if and only if $\rho = 0$. It can also be shown algebraically that ρ_c is proportional to ρ and that $-1 \le -|\rho| \le \rho_c \le |\rho| \le 1$ (Lin 1989). Hence imprecision can be reflected by a smaller ρ and systematic bias can be reflected by a smaller ratio of ρ_c relative to ρ. Together, information on ρ and ρ_c provide a set of tools to identify which corrective actions, either to improve accuracy and/or to improve precision, is most beneficial (Lin and Chinchilli 1997).

The intraclass correlation coefficient (ICC) is also a popular measure of agreement for variables measured on a continuous scale (Fleiss 1986). Suppose Y_{i1} and Y_{i2} can be described in a linear model as follows: $y_{ij} = \mu_j + \theta_i + e_{ij}$ where μ_j is the mean of the measurement from the j^{th} method, $\theta_i \sim (0, \sigma_\theta^2)$ is the latent variable for the i^{th} subject, and the $e_{ij} \sim (0, \sigma_e^2)$ are independent errors terms. Carrasco and Jover (2003, page 850) used a model with variance components to demonstrate that the CCC is the intraclass correlation coefficient (ICC) when one takes into account the difference in averages of the methods:

$$\rho_{\text{ICC}} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_e^2 + \sigma_\mu^2} = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}.$$

1. Hung-Mo Lin, Department of Anesthesiology, Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1010, New York, NY 10029, U.S.A. E-mail: hung-mo.lin@mountsinai.org; Hae-Young Kim, Center for Statistical Analysis and Research, New England Research Institutes, 9 Galen Street, Watertown, MA 02472, U.S.A.; John M. Williamson, Center for Global Health Research, Centers for Disease Control and Prevention/Kenya Medical Research Institute, 1578 Kisumu-Busia Road, Kisian Kisumu, Kenya; Virginia M. Lesser, Department of Statistics and Survey Research Center, Oregon State University, 44 Kidder Hall, Corvallis, OR 97731-4606, U.S.A.

Therefore, one can estimate the CCC using the variance components of a mixed effects model or the common method of moments. Because of its superiority to the Pearson correlation coefficient and its link to the ICC, application of the CCC has gained popularity in recent years (Chinchilli, Martel, Kumanyika and Lloyd 1996; Zar 1996). In 2009 and the 2010, the CCC was used as a measure of agreement in more than 60 medical publications in areas such as respiratory illness (Dixon, Sugar, Zinreich, Slavin, Corren, Naclerio, Ishii, Cohen, Brown, Wise and Irvin 2009; Kocks, Kerstjens, Snijders, de Vos, Biermann, van Hengel, Strijbos, Bosveld and van der Molen 2010), sleep (Khawaja, Olson, van der Walt, Bukartyk, Somers, Dierkhising and Morgenthaler 2010), pediatrics (Liottol, Radaelli, Orsi1, Taricco, Roggerol, Giann, Consonni, Mosca1 and Cetin 2010), neurology (MacDougall, Weber, McGarvie, Halmagyi and Curthoys 2009), and radiology (Mazaheri, Hricak, Fine, Akin, Shukla-Dave, Ishill, Moskowitz, Grater, Reuter, Zakian, Touijer and Koutcher 2009).

The kappa coefficient (κ) (Cohen 1960) and the weighted kappa coefficient (Cohen 1968) are the most popular indices for measuring agreement for discrete and ordinal outcomes, respectively (Fleiss 1981). Let Y_{i1} and Y_{i2} denote two binary random variables taking values 0 and 1 with probabilities denoted by $\pi_1 = \Pr(Y_{i1} = 1)$ and $\pi_2 = \Pr(Y_{i2} = 1)$. Kappa corrects the percentage of agreement between raters by taking into account the proportion of agreement expected by chance (calculated under independence), and is defined as follows:

$$\kappa = \frac{P_o - P_e}{1.0 - P_e}, \qquad (2)$$

where P_e is the probability that the pair of binary responses are equal assuming independence $(\pi_1\pi_2 + (1-\pi_1)(1-\pi_2))$ and P_o is the probability that the pair are equal (Cohen 1960). The difference $P_o - P_e$ is the excess of agreement over chance agreement. A value of 0 for κ indicates no agreement beyond chance and a value of 1 indicates perfect agreement (Fleiss 1981). Disadvantages of kappa are that is a function of the marginal distribution of the raters (Fleiss, Nee and Landis 1979; Tanner and Young 1985) and its range depends on the number of ratings per subject (Fleiss *et al.* 1979). Robieson (1999) noted that the CCC computed from ordinal scaled data is equivalent to the weighted kappa when integer scores are used. Kappa has been used to measure the validity and reproducibility of the similarity between twins (Klar, Lipsitz and Ibrahim 2000), different epidemiologic tools (Maclure and Willett 1987), and control-informant agreement from case-control studies (Korten, Jorm, Henderson, McCusker and Creasey 1992).

The value of sample surveys have been well recognized and estimation for data collected from sample surveys has been widely documented (Hansen, Hurwitz and Madow 1953; Cochran 1963; Kish 1965). For example, a number of federal studies conducted in the U.S. to obtain estimates of the health of the population are based on national surveys, such as the National Health Interview Survey (NHIS), the Behavioral Risk Factor Surveillance System (BRFSS), and the National Health and Nutrition Examination Surveys (NHANES). Each of these studies incorporates complex survey design structure, namely oversampling of subpopulations, stratification and clustering. These designs are often used to improve precision, provide estimates for subpopulations, or reduce costs associated with frame development. In order to draw design-based inference to the targeted population for complex survey designs, estimators and their variances include sampling weights and account for the design structure to obtain unbiased estimates. In addition, by including the sampling weights and incorporating the sample design in analyses, any potential correlation from the clusters in a multistage design is taken into account so that the standard errors of the estimators are not underestimated.

Often researchers are not interested in testing whether their estimation of agreement using either the CCC or kappa is significantly different from zero. Their interest is to report the confidence intervals along with their estimates (*e.g.*, Dixon *et al.* 2009; Mazaheri *et al.* 2009). Similar to the Pearson correlation coefficient, there is no target value that can be used to judge if agreement is strong. Therefore, it is essential that judgment of agreement between any test and reference methods should be made with an established degree of certainty. In some situations, studies are conducted that require hypothesis testing or comparisons of agreement indexes for more than one new methods against a reference method. For examples, Khawaja *et al.* (2010) tested the equality of two CCCs that compared the apnea hypopnea index (AHI) from the first 2 and 3 hours of sleep with the gold standard AHI from FN-PSG (FN-AHI). In radiology research, associations between volume measurements of prostate tumor from imaging and also from pathologic examination were assessed by comparing CCCs. The two imaging methods were tested for equality of agreement with the pathologic results (Mazaheri *et al.* 2009). Tests of equal kappa have been used to compare visual assessment and computerized planimetry in assessing cervical ectopy (Gilmour, Ellerbrock, Koulos, Chiasson, Williamson, Kuhn and Wright 1997; Williamson, Manatunga and Lipsitz 2000), and in comparing monozygotic and dizygotic twins in terms of cholesterol levels (Feinleib, Garrison, Fabsitz, Christian, Hrubec, Borhani, Kannel, Roseman, Schwartz and Wagner 1977).

As illustrated in the two NHANES III examples in Section 3, large differences can exist between the weighted and unweighted estimates of parameter estimate standard errors in survey studies. Failure to include sampling weights and take into account the sample design in analyses will result in underestimation of standard errors and incorrect inference. This is especially important for surveys repeated every few years, and researchers often have a special interest in comparing changes among domains or subpopulations. For instance, in the first NHANES III application, we compare the agreement between self reported and measured body weights at examination in adolescents. Computing accurate standard errors (confidence intervals) are necessary if interest is to compare the CCC across domains, such as normal weight and obese subgroups.

We provide weighted measures of the CCC and kappa coefficient, along with the variance estimators of these measures accounting for the sampling design. In Section 2, we present a generalized estimating equations approach for estimating these two agreement coefficients from sample survey data. In Section 3, we illustrate our method with data collected from the NHANES III study. We use body measurement data to estimate ρ_c for assessing the agreement between self-reported and actual weight. We also use oral health data to estimate κ for assessing the agreement between two definitions of periodontal disease. We account for stratification and clustering, and incorporate weights of the survey design in both examples. We conclude with a short discussion.

2. Methods

We propose a general approach for estimating the CCC and kappa from sample survey data using two GEE approaches. For the CCC, three sets of estimating equations are required. A first set of estimating equations models the distribution of the continuous responses. Following Barnhart and Williamson (2001), a second set of estimating equations is used to estimate the variances of the continuous responses. A third set of estimating equations estimates the CCC by modeling the covariance between the paired continuous responses and the estimates of the means and variances from the first two sets of estimating equations. For κ, only two sets of estimating equations are required. A first set of estimating equations models the marginal distribution of the binary responses. Following Lipsitz, Laird and Brennan (1994), a second set of estimating equations is introduced to estimate κ by modeling a binary random variable depicting agreement between two responses on a subject.

In order to account for variable selection probabilities, weight matrices are incorporated into each set of estimating equations. Standard error estimation of the proposed $\hat{\rho}_c$ and $\hat{\kappa}$ from sample survey data are conducted with the Taylor series linearization method. We also show how standard error estimation of the proposed estimators can be accomplished by using the jackknife approach.

Assume a sample survey is conducted with stratification, clustering, and unequal probabilities of selection. Let Y_{hij} denote the response variable for the j^{th} member ($j = 1, ..., m_{hi}$) of the i^{th} cluster ($i = 1, ..., n_h$) of the h^{th} stratum ($h = 1, ..., H$). Averaging over all possible samples, the corresponding expected value is $E[Y_{hij}] = \mu_{hij}$ if Y_{hij} is a continuous response, and the corresponding probability $E[Y_{hij}] = \Pr[Y_{hij} = 1] = \pi_{hij}$ if Y_{hij} is a binary response. The sampling weight w_{hij} is the inverse of the probability of selection for the j^{th} member of the i^{th} cluster of the h^{th} stratum.

2.1 The concordance correlation coefficient

Liang and Zeger (1986) developed moment-based methods for analyzing correlated observations from the same cluster (*e.g.*, repeated measurements over time on the same individual or observations on multiple members of the same family). The GEE approach results in consistent marginal parameter estimation, even with misspecification of the correlation structure by using a robust "sandwich" estimator of variance. We use the GEE approach to analyze sample survey data by additionally incorporating a sampling weight matrix as follows:

$$\sum_{h=1}^{H}\sum_{i=1}^{n_h}\mathbf{D}'_{hi}\,\mathbf{W}_{hi}\,\mathbf{V}_{hi}^{-1}\,(\mathbf{Y}_{hi}-\boldsymbol{\mu}_{hi}(\hat{\boldsymbol{\mu}}))=\mathbf{0},$$

where $\mathbf{D}'_{hi}$ is the $(q \times m_{hi})$ derivative matrix $d[\boldsymbol{\mu}_{hi}]'/d\boldsymbol{\mu}$, $\mathbf{W}_{hi}$ is a $(m_{hi} \times m_{hi})$ main diagonal matrix consisting of the person-specific sampling weights w_{hij}, $\mathbf{V}_{hi}$ is a $(m_{hi} \times m_{hi})$ working variance-covariance matrix for the within-cluster responses, $\mathbf{Y}_{hi}$ is a $(m_{hi} \times 1)$ response vector consisting of the responses Y_{hij}, and $\boldsymbol{\mu}_{hi} = E[\mathbf{Y}_{hi}]$ is possibly a function of the $(q \times 1)$ parameter vector $\boldsymbol{\beta}$. The GEE can then be solved non-iteratively, resulting in the usual estimate

$$\hat{\mu}=\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}\,Y_{hij}\right)\Bigg/\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}\right)$$

if we are estimating a common mean $\mu = \boldsymbol{\beta}$ ($q = 1$) and are using an independence working covariance matrix.

Assume a pair of continuous responses are observed for the j^{th} member of the i^{th} cluster of the h^{th} stratum, Y_{hij1} and Y_{hij2}, and their expected values are μ_{hij1} and μ_{hij2}. Again, assume we are estimating common means μ_1 and μ_2 without covariates for the pair of within-subject continuous responses, which can be estimated by using the above generalized estimating equation.

Barnhart and Williamson (2001) demonstrated how three sets of generalized estimating equations can be used to model the CCC defined in (1) using correlated data. We extend Barnhart and Williamson's (2001) second set of GEE equations to estimate the variances of the continuous responses by again incorporating a weight matrix as follows:

$$\nu_2(\hat{\sigma}_1^2, \hat{\sigma}_1^2, \hat{\mu}_1, \hat{\mu}_2) = \sum_{h=1}^{H}\sum_{i=1}^{n_h} \mathbf{F}'_{hi}\, \mathbf{W}_{hi}\, \mathbf{H}_{hi}^{-1}(\mathbf{Y}_{hi}^2 - \boldsymbol{\delta}_{hi}^2(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\mu}_1, \hat{\mu}_2)) = \mathbf{0},$$

where $\mathbf{F}'_{hi}$ is the $(2 \times 2m_{hi})$ derivative matrix $d[\boldsymbol{\delta}_{hi}^2]'/d\boldsymbol{\sigma}^2$ with $\boldsymbol{\sigma}^2 = [\sigma_1^2, \sigma_2^2]'$, $\mathbf{W}_{hi}$ is a $(2m_{hi} \times 2m_{hi})$ main diagonal matrix consisting of the person-specific sampling weights w_{hij}, $\mathbf{H}_{hi}$ is a $(2m_{hi} \times 2m_{hi})$ working variance-covariance matrix for the within-cluster squared responses, $\mathbf{Y}_{hi}^2 = [Y_{hi11}^2, Y_{hi12}^2, Y_{hi21}^2, Y_{hi22}^2, ..., Y_{him_{hi}1}^2, Y_{him_{hi}2}^2]'$ is a $(2m_{hi} \times 1)$ response vector of the continuous variables, and $\boldsymbol{\delta}_{hi}^2 = E[\mathbf{Y}_{hi}^2]$. Although $\boldsymbol{\delta}_{hi}^2$ is a function of both the variance terms σ_1^2 and σ_2^2 and the means μ_1 and μ_2, it is assumed that the means are fixed in $\boldsymbol{\delta}_{hi}^2$ and one only takes derivatives of $\boldsymbol{\delta}_{hi}^2$ with respect to the variances. Again we choose the $(2m_{hi} \times 2m_{hi})$ matrix $\mathbf{H}_{hi}$ to be the "independence" working variance-covariance matrix and the $(2m_{hi} \times 1)$ column vector $\boldsymbol{\delta}_{hi}^2 = [\sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2, ..., \sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2]'$ because we are assuming common variances and means across all strata and clusters. The above GEE can thus be solved non-iteratively:

$$\hat{\sigma}_p^2 = \left(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} W_{hijp}\, Y_{hijp}^2\right) \Bigg/ \left(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} W_{hijp}\right) - \hat{\mu}_p^2,$$

for the p^{th} measurement in the pair, $p = 1, 2$.

The CCC can be estimated in a third set of estimating equations by using the pairwise products of the responses to model σ_{12}, once the means and variances are estimated. Let $\mathbf{U}_{hi} = [Y_{hi11}Y_{hi12}, Y_{hi21}Y_{hi22}, ..., Y_{him_{hi}1}Y_{him_{hi}2}]'$ be a $(m_{hi} \times 1)$ vector of pairwise products of the responses and denote $\boldsymbol{\theta}_{hi} = E[\mathbf{U}_{hi}]$, which is a function of the means, variances, and CCC. We solve for $\hat{\rho}_c$ in a third set of estimating equations:

$$\nu_3(\hat{\rho}_c, \hat{\sigma}_1^2, \hat{\sigma}_1^2, \hat{\mu}_1, \hat{\mu}_2) = \sum_{h=1}^{H}\sum_{i=1}^{n_h} \mathbf{C}'_{hi}\, \mathbf{W}_{hi}\, \mathbf{K}_{hi}^{-1}(\mathbf{U}_{hi} - \boldsymbol{\theta}_{hi}(\hat{\rho}_c, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\mu}_1, \hat{\mu}_2)) = 0,$$

where $\mathbf{C}'_{hi}$ is a $(1 \times m_{hi})$ derivative vector $= \partial\boldsymbol{\theta}_{hi}/\partial\rho_c$, $\mathbf{W}_{hi}$ is a $(m_{hi} \times m_{hi})$ main diagonal matrix consisting of the person-specific sampling weights w_{hij}, and $\mathbf{K}_{hi}$ is a $(m_{hi} \times m_{hi})$ working covariance matrix that we choose to be the "independence" covariance matrix. The above GEE can be solved non-iteratively:

$$\hat{\rho}_c = \frac{2\hat{\sigma}_{12}}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + (\hat{\mu}_1 - \hat{\mu}_2)^2},$$

where

$$\hat{\sigma}_{12} = \frac{\left(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} W_{hij12}\, Y_{hij1}\, Y_{hij2}\right)}{\left(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} W_{hij12}\right)} - \hat{\mu}_1\hat{\mu}_2.$$

2.2 Linearization estimator of variance

The usual robust estimators of variance for the means and CCC from the GEE approach are invalid here because they do not take into account the sampling structure, only the correlation of observations made on the same individual. We propose standard error estimation using the Taylor series linearization method (Binder 1983; Binder 1996). The first derivatives of ρ_c (equation 1) with respect to μ_1, μ_2, σ_1^2, σ_2^2, and σ_{12} are:

$$\frac{\partial\rho_c}{\partial\mu_1} = \frac{-4\sigma_{12}(\mu_1 - \mu_2)}{D^2},$$
$$\frac{\partial\rho_c}{\partial\mu_2} = \frac{-4\sigma_{12}(\mu_2 - \mu_1)}{D^2},$$
$$\frac{\partial\rho_c}{\partial\sigma_1^2} = \frac{-2\sigma_{12}}{D^2},$$
$$\frac{\partial\rho_c}{\partial\sigma_2^2} = \frac{-2\sigma_{12}}{D^2},$$
$$\frac{\partial\rho_c}{\partial\sigma_{12}} = \frac{2}{D},$$

where $D = \sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2$. Thus

$$\begin{aligned}\hat{\rho}_c - \rho_c \approx{}& \left(\frac{\partial\rho_c}{\partial\mu_1}\right)(\hat{\mu}_1 - \mu_1) + \left(\frac{\partial\rho_c}{\partial\mu_2}\right)(\hat{\mu}_2 - \mu_2) \\ &+ \left(\frac{\partial\rho_c}{\partial\sigma_1^2}\right)(\hat{\sigma}_1^2 - \sigma_1^2) \\ &+ \left(\frac{\partial\rho_c}{\partial\sigma_2^2}\right)(\hat{\sigma}_2^2 - \sigma_2^2) + \left(\frac{\partial\rho_c}{\partial\sigma_{12}}\right)(\hat{\sigma}_{12} - \sigma_{12}) \\ ={}& \frac{-4\sigma_{12}(\mu_1 - \mu_2)}{D^2}(\hat{\mu}_1 - \mu_1) \\ &+ \frac{-4\sigma_{12}(\mu_2 - \mu_1)}{D^2}(\hat{\mu}_2 - \mu_2) \\ &+ \frac{-2\sigma_{12}}{D^2}(\hat{\sigma}_1^2 - \sigma_1^2) + \frac{-2\sigma_{12}}{D^2}(\hat{\sigma}_2^2 - \sigma_2^2) \\ &+ \frac{2}{D}(\hat{\sigma}_{12} - \sigma_{12}).\end{aligned}$$

The above equation can be rearranged into two parts, one involving the parameter estimates $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2$, and $\hat{\sigma}_{12}$ and the other involving only parameters which does not

contribute to the variance estimation of $\hat{\rho}_c$. Thus the first part becomes

$$
\begin{aligned}
&-\frac{4\sigma_{12}(\mu_1-\mu_2)}{D^2}\hat{\mu}_1-\frac{4\sigma_{12}(\mu_2-\mu_1)}{D^2}\hat{\mu}_2\\
&\quad-\frac{2\sigma_{12}}{D^2}\hat{\sigma}_1^2-\frac{2\sigma_{12}}{D^2}\hat{\sigma}_2^2+\frac{2}{D}\hat{\sigma}_{12}\\
&\quad=-\frac{2\sigma_{12}}{D^2}(2(\mu_1-\mu_2)(\hat{\mu}_1-\hat{\mu}_2)+\hat{\sigma}_1^2+\hat{\sigma}_2^2)+\frac{2}{D}\hat{\sigma}_{12}\\
&\quad=-\frac{2\sigma_{12}}{D^2}\Bigg(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}2(\mu_1-\mu_2)(w_{hij}^{*}Y_{hij1}-w_{hij}^{*}Y_{hij2})\\
&\qquad\qquad+w_{hij}^{*}(Y_{hij1}-\mu_1)^2+w_{hij}^{*}(Y_{hij2}-\mu_2)^2\Bigg)\\
&\quad+\frac{2}{D}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}^{*}(Y_{hij1}-\mu_1)(Y_{hij2}-\mu_2) \qquad (3)
\end{aligned}
$$

where $w_{hij}^{*}=w_{hij}/(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij})$. Equation (3) becomes a linear function of the data after the summation is moved to the front, which we can then express as $\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}^{*}z_{hij}$, where

$$
\begin{aligned}
z_{hij}=&-\frac{2\sigma_{12}}{D^2}(2(\mu_1-\mu_2)(Y_{hij1}-Y_{hij2})\\
&+(Y_{hij1}-\mu_1)^2+(Y_{hij2}-\mu_2)^2)\\
&+\frac{2}{D}(Y_{hij1}-\mu_1)(Y_{hij2}-\mu_2). \qquad (4)
\end{aligned}
$$

One then creates a random variable $\hat{z}_{hij}$ based on equation (4) that replaces the parameters with their respective estimates. The variance of this new estimator $\hat{z}_{hij}$ is an approximation for the variance of $\hat{\rho}_c$, which can be estimated using standard survey software (see Appendix).

2.3 Jackknife estimator of variance

We also use the jackknife technique for standard error estimation of the parameters following Rust and Rao (1996, Section 2.1) for comparison with the linearization estimates. The jackknife technique is implemented by calculating a set of replicate estimates and estimating the variance using them. A replicate data set is created for each cluster by deleting all observations from the given cluster from the sample. The weights of all other observations in the stratum containing the cluster are inflated by a factor $n_h/(n_h-1)$. Weights in the other strata remain unchanged. Thus, the new weights for the replicated data set created by removing cluster i from stratum h are:

$\omega_{klj}^{(hi)}=w_{klj}$ if $k\neq h$ (different strata)

$\omega_{hlj}^{(hi)}=w_{hlj}n_h/(n_h-1)$ if $l\neq i$ (same strata but different clusters)

$\omega_{hij}^{(hi)}=0$ (for the cluster being removed).

The resulting jackknife variance estimator for $\hat{\rho}_c$ is

$$v_J(\hat{\rho}_c)=\sum_{h=1}^{H}\left(\frac{n_h-1}{n_h}\right)\sum_{i=1}^{n_h}(\hat{\rho}_{c(hi)}-\hat{\rho}_c)^2$$

where $\hat{\rho}_{c(hi)}$ is estimated in the same way as $\hat{\rho}_c$, but using the recalculated weights $\omega^{(hi)}$ instead of the weights ω. The jackknife estimators for the means are similarly calculated.

2.4 The kappa coefficient

Assume a pair of binary responses are observed for the j^{th} member of the i^{th} cluster of the h^{th} stratum, Y_{hij1} and Y_{hij2}, and their expected values are the probabilities π_{hij1} and π_{hij2}. Again assume we are estimating common probabilities π_1 and π_2 without covariates for the pair of within-subject binary responses. Lipsitz *et al.* (1994) demonstrated how two sets of generalized estimating equations can be used to develop simple non-iterative estimates of the κ-coefficient that can be used for unbalanced data as previous estimates of kappa and its variance were only proposed for balanced data. They defined the binary random variable $U_{hij}=Y_{hij1}Y_{hij2}+(1-Y_{hij1})(1-Y_{hij2})=1$ if both responses in the pair agree and 0 otherwise. Accordingly, $E[U_{hij}]=P_o$, which denotes the probability of observed agreement and is assumed here to be constant over all strata, clusters, and pairs of observations. Now let $E[Y_{hij1}Y_{hij2}]=\Pr[Y_{hij1}=Y_{hij2}=1]=\omega$. The probability of observed agreement can be expressed as $P_o=1-\pi_1-\pi_2+2\omega$. The probability of expected agreement by chance is defined as $P_e=\pi_1\pi_2+(1-\pi_1)(1-\pi_2)$ and is estimated by $\hat{P}_e=\hat{\pi}_1\hat{\pi}_2+(1-\hat{\pi}_1)(1-\hat{\pi}_2)$, where $\hat{\pi}_1$ and $\hat{\pi}_2$ are calculated in the first set of estimating equations.

We can derive estimates of κ from sample survey data following the approach for the CCC in Section 2.1. We can incorporate the survey weight matrices into Lipsitz *et al.*'s (1994) two sets of GEE equations for estimating kappa. Then, by choosing "independence" working covariance matrices for the two sets of equations as in Lipsitz *et al.*'s (1994) approach, we arrive at the following non-iterative estimate of kappa for sample survey data:

$$\hat{\kappa}=\frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}U_{hij}-\hat{P}_e\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}-\hat{P}_e\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}}. \qquad (5)$$

This estimator is identical to Lumley's (2010), which can be computed using the R software survey package and svykappa function.

Standard error estimation of $\hat{\kappa}$ can be conducted similarly to that of $\hat{\rho}_c$ using the Taylor series linearization method. The first derivatives of kappa with respect to P_o, π_1, and π_2 are:

$$\frac{\partial \kappa}{\partial P_o} = \frac{1}{1-P_e},$$
$$\frac{\partial \kappa}{\partial \pi_1} = \frac{(1-P_o)(1-2\pi_2)}{(1-P_e)^2},$$
$$\frac{\partial \kappa}{\partial \pi_2} = \frac{(1-P_o)(1-2\pi_1)}{(1-P_e)^2}.$$

Thus

$$\hat{\kappa} - \kappa \approx \left(\frac{\partial \kappa}{\partial P_o}\right)(\hat{P}_o - P_o) + \left(\frac{\partial \kappa}{\partial \pi_1}\right)(\hat{\pi}_1 - \pi_1) + \left(\frac{\partial \kappa}{\partial \pi_2}\right)(\hat{\pi}_2 - \pi_2) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w^*_{hij}\, z_{hij},$$

where $w^*_{hij} = w_{hij} / (\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij})$ and

$$z_{hij} = \left(\frac{\partial \kappa}{\partial P_o}\right)U_{hij} + \left(\frac{\partial \kappa}{\partial \pi_1}\right)Y_{hij1} + \left(\frac{\partial \kappa}{\partial \pi_2}\right)Y_{hij2} = \frac{U_{hij}}{1-P_e} + \frac{(1-P_o)[Y_{hij1}(1-2\pi_2) + Y_{hij2}(1-2\pi_1)]}{(1-P_e)^2}. \quad (6)$$

Replacing the parameters in (6) with their respective estimates, one then treats $\hat{z}_{hij}$ as a random variable and estimates its variance using standard survey software that accounts for the sampling design. The variance of this new estimator $\hat{z}_{hij}$ is an approximation for the variance of $\hat{\kappa}$. The jackknife method can also be used to estimate the variance of $\hat{\kappa}$.

3. NHANES III survey

We used data from the Third National Health and Nutrition Examination Survey to illustrate our method. NHANES III was conducted by the National Center for Health Statistics of the Centers for Disease Control and Prevention and was designed as a six-year survey divided into two phases (1988-1991 and 1991-1994). The data were collected using a complex, multistage, probability sampling design to select participants representative of the civilian, non-institutionalized US population. Details of the survey design and analytic and reporting guidelines were published in the NHANES III reference manuals and reports (National Center for Health Statistics 1996).

3.1 The adolescent weight study

Obesity is a rapidly increasing public health problem with surveillance most often based on self-reported values of height and weight. A series of recent studies and systemic reviews have attempted to assess the agreement between self-reported and measured weight, especially in the adolescent population. The general findings suggest that self-reported weight was slightly lower than measured weight, and that a significant number of normal weight adolescents misperceive themselves as overweight and are engaging in unhealthy weight control behaviors (Field, Aneja and Rosner 2007; Gorber, Tremblay, Moher and Gorber 2007; Sherry, Jefferds and Grummer-Strawn 2007). Therefore, researchers have suggested that obesity prevention programs should address weight misperceptions and the harmful effects of unhealthy weight control methods even among normal weight adolescents (Talamayan, Springer, Kelder, Gorospe and Joye 2006). A similar Canadian study from the 2005 Canadian Community Health Survey that focused on adult individuals also showed that associations between obesity and health conditions may be overestimated if self-reported weight is used (Shield, Gorber and Tremblay 2008). We use data obtained from the Body Measurements (Anthropometry) component of the NHANES III study to estimate the CCC that measures agreement between self-reported and measured weight (pounds) obtained from adolescents (aged 12 through 16 years).

The self-reported weight was obtained just prior to the actual measurement of weight. We use data from the entire six-year survey period (both 1988-1991 and 1991-1994). For simplicity, we excluded one stratum which only had one PSU. Hence, there were 48 strata and each stratum had two PSUs. The sample weight labeled wtpfex6 accounting for the differential selection probability was used in our analyses. There were 1,651 subjects with complete data for both weight measurements. The estimates of the self-reported and actual weights (in pounds) were 135.5 (s.e. = 1.8) and 136.3 (s.e. = 1.8), respectively, calculated using PROC SURVEYMEANS in SAS. The estimates of the standard errors based on the jackknife approach are the same as above.

The CCC is a natural choice for assessing the agreement between the two weight measurements because they are measured on the same scale and their ranges are similar (self-reported weight: 78 lbs ~ 350 lbs and actual weight: 73 lbs ~ 372 lbs) (Lin and Chinchilli 1997). The estimate of the CCC for measuring the agreement between the two definitions of weight using the proposed method is 0.93. The standard error of the estimate is 0.021 using the Taylor series linearization method. The jackknife standard error of 0.021 agrees closely with the linearization standard error. These statistics are summarized in Table 1 along with their values computed when the sampling structure is ignored. The standard errors for the estimates incorporating the sampling structure are much larger than the unweighted estimates.

Table 1
Unweighted and weighted average, CCC, and respective standard errors for adolescent self-reported and actual weight in pounds

	Self-reported	Actual	CCC
Unweighted Estimate	135.31	136.96	0.890
SE	0.76	0.80	0.0005
Weighted Estimate	135.47	136.30	0.926
SE	1.75	1.82	0.0205

Similar to the CCC, the usual Pearson correlation coefficient between the self-reported and the actual weight measures is also 0.93. In this case, the mean difference between the two weight measurements is just less than one pound. When subpopulations are examined, differences are noted in the CCC and the Pearson correlation coefficient. Consider a subpopulation of those individuals that had a measured weight $>$ 200 lbs at examination. Summarizing the data for this subpopulation, the self-reported weight is on average 8 pounds less than the measured weight (223.2 lbs *vs* 231.4 lbs). There is a slight departure of the CCC (0.72) from the Pearson correlation coefficient (0.76). The discrepancy between the two measures increases in the more obese subgroup. In the subpopulation where measured weight is $>$ 220 lbs, the means of self-reported and measured weights are 231.9 lbs and 248.8 lbs, respectively. The CCC is 0.67, whereas the Pearson correlation coefficient is 0.85. In this situation, the CCC reflects both the reproducibility and differences between the self-reported and measured means. Therefore, the CCC is informative and advantageous when considering these comparisons, particularly in domain analysis within a complex survey.

3.2 The oral health study

Slade and Beck (1999) used extent of pocket depth and loss of attachment as indices of periodontal conditions. Prevalence of periodontal disease using previously reported thresholds of pocket depth ≥ 4 mm and attachment loss ≥ 3 mm were estimated by Slade and Beck (1999, Table 1). Pocket depth may be reflective of inflammation rather than chronic periodontal disease and, thus, attachment level may be a more meaningful measure of periodontal destruction. However, pocket depth remains the recommended measurement in clinical practice (Winn, Johnson and Kingman 1999). Therefore, we compare the agreement of these two definitions of periodontal disease using the kappa coefficient.

We use the sample that was analyzed by Slade and Beck (1999). The data include 14,415 persons aged 13 or older who had complete pocket depth and attachment loss assessment by six designated dentists. We again use data from the entire six-year survey period (both 1988-1991 and 1991-1994). There were a total of 49 strata and each stratum had two PSUs. The variable labeled sample weight, wtpfex6, accounting for differential selection probability, was used in our analyses.

The first definition of periodontal disease is pocket depth ≥ 4 mm and the second is maximum attachment loss ≥ 3 mm. For both variables we are using the maximum values among all teeth in an individual's mouth. The probability estimates of the attachment loss and pocket depth variables are 0.358 (jackknife s.e. = 0.0088) and 0.212 (jackknife s.e. = 0.016), respectively, using the proposed method. The asymptotic standard errors based on the usual Taylor series expansion (Woodruff 1971, produced by PROC SURVEYFREQ in SAS, version 9.1) are 0.0088 and 0.015, respectively.

Kappa is a natural choice for assessing the agreement between two binary ratings as it corrects for chance agreement (Fleiss 1981). The estimate of kappa for measuring the agreement between the two definitions of periodontal disease (pocket depth of ≥ 4 mm and attachment loss of ≥ 3 mm) using the proposed method is 0.307. The standard error of 0.0158 was obtained by both the Taylor series linearization and jackknife methods. Table 2 compares these results to the measures when the complex sampling structure is ignored. The standard error of the kappa coefficient is larger when accounting for the survey structure.

Table 2
Unweighted and weighted average, kappa, and respective standard errors for attachment loss and pocket depth

	Attachment Loss	Pocket Depth	Kappa
Unweighted Estimate	0.393	0.283	0.334
SE	0.004	0.004	0.008
Weighted Estimate	0.358	0.212	0.307
SE	0.009	0.016	0.0158

4. Discussion

The CCC and kappa evaluate the agreement between two measurements for continuous and categorical responses, respectively. In this paper, we have proposed a generalized estimating equation approach for estimating the CCC for a pair of continuous variables, and kappa for a pair of binary variables, from sample survey data where the data have been collected using complex survey features such as stratification or clustering. The usual sandwich estimator of the variance only accounts for repeated measurements made on the same individual, and does not account for the sampling framework (*e.g.*, clustering, stratification, and weighting). In the GEE approach, standard error estimation of the estimators is conducted with the Taylor series linearization and jackknife approaches. If the data are not collected using complex survey features, the proposed estimators will be identical to the usual estimators. As is

evident in the two examples from the NHANES III study, we have shown the need to incorporate sampling weights and the sampling design features so that the standard errors are not underestimated when data are collected from a complex sampling design. Tables 1 and 2 show that there were large differences in the standard errors between weighted and unweighted estimates of the standard errors for both CCC and kappa. Confidence intervals that incorporate weights and the design features will allow correct inference.

In the appendix, we show steps for calculating the weighted measures of the CCC and kappa, along with their standard errors using standard survey software that incorporates the sampling weights, clustering and stratification. The GEE approach is advantageous because it is a convenient framework for developing estimators of the agreement coefficients and is easily extended to multiple raters, multiple methods, covariate adjustment and unbalanced cluster sizes. This design-based approach results in correct standard error estimation without assuming an underlying model and accounting for the sampling structure. If one is interested in estimating the agreement between two ordinal variables with kappa then Williamson *et al.*'s (2000) generalized estimating equation approach can be extended similarly to the proposed method.

Acknowledgements

We thank the anonymous reviewers and editor for their helpful comments. In particular, the editor gave extremely valuable suggestions for the introduction section.

Appendix

Steps for calculating the CCC and its standard error using standard survey software

Step 1: Calculate the means of the continuous variables Y_{hij1} and Y_{hij2} using software for survey data that incorporates stratification, clustering, and sample weighting (*e.g.*, PROC SURVEYMEANS in SAS).

Step 2: Square the centered Y_{hij1} and Y_{hij2} values around their respective means.

Step 3: Calculate the means of the squared centered Y_{hij1} and Y_{hij2} values using standard software for survey data. These means are the variance estimates of Y_{hij1} and Y_{hij2}. Calculate the mean of the product of the centered Y_{hij1} and Y_{hij2} values using standard software for survey data. This mean is the estimated covariance of Y_{hij1} and Y_{hij2}.

Step 4: Calculate the CCC by substituting the estimated means and variances into equation (1). Create the new variable Z_{hij} based on equation (4).

Step 5: Calculate the standard error of Z_{hij} using standard software for survey data. The standard error of Z_{hij} estimates the standard error of $\hat{\rho}_c$.

SAS CODE:

Let $y1$ and $y2$ denote the variables for the pair of continuous responses, and *s*, *c* and *w* denote the variables for strata, cluster and weight:

```
PROC SURVEYMEANS DATA=dataset MEAN; /* Step 1 above */;
      STRATA s;
      CLUSTER c;
      WEIGHT w;
      VAR y1 y2;
      ODS OUTPUT STATISTICS=stat;
data _null_;
      set stat (where=(varname='y1' ));
      call symputx('muy1', mean);
data _null_;
      set stat (where=(varname='y2' ));
      call symputx('muy2', mean);
data dataset; set dataset;                    /* Step 2 above */;
      cy1 = y1 – &muy1;
      cy2 = y2 – &muy2;
      vary1 = cy1 * *2;
      vary2 = cy2 * *2;
      covy12 = cy1 * cy2;
PROC SURVEYMEANS MEAN;                        /* Step 3 above */;
      STRATA s;
      CLUSTER c;
      WEIGHT w;
      VAR vary1 vary2 covy12;
      ODS OUTPUT STATISTICS=stat;
run;
data _null_;
      set stat (where=(varname='vary1' ));
      call symputx('vary1', mean);
data _null_;
      set stat (where=(varname='vary2' ));
      call symputx('vary2', mean);
data _null_;
      set stat (where=(varname='covy12' ));
      call symputx('covy12', mean);
data dataset; set dataset;                    /* Step 4 above */;
      d = &vary1 + &vary2 + (&muy1 – &muy2) ** 2;
      CCC = 2 * &covy12 / d;
      z = (2 / d) * (cy1 * cy2) – (2 * &covy12 / d / d) * ((cy1 * *2) +
      (cy2 * *2) + 2 * (&muy1 – &muy2) * (y1 – y2));
PROC SURVEYMEANS MEAN;                        /* Step 5 above */ ;
      STRATA s;
      CLUSTER c;
      WEIGHT w;
      VAR CCC z;
run;
```

Steps for calculating kappa and its standard error using standard survey software

Step 1: Estimate the probabilities of the binary variables Y_{hij1} and Y_{hij2} using software for survey data that incorporates stratification, clustering, and sample weighting (*e.g.*, PROC SURVEYFREQ in SAS).

Step 2: Estimate $P_e(= \hat{\pi}_1\hat{\pi}_2 + (1-\hat{\pi}_1)(1-\hat{\pi}_2))$.

Step 3: Create the new agreement variable $U_{hij}(= Y_{hij1}$ $Y_{hij2} + (1-Y_{hij1})(1-Y_{hij2}))$.

Step 4: Calculate the sum of the sample survey weights and the sum of the weighted U_{hij} (*e.g.*, using PROC SURVEYMEANS in SAS). Estimate kappa using equation (2).

Step 5: Create a new variable z_{hij} using equation (6).

Step 6: Calculate the standard error of z_{hij} using standard software for survey data. The standard error of z_{hij} estimates the standard error of $\hat{\kappa}$.

References

Barnhart, H.X., and Williamson, J.M. (2001). Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics*, 57, 931-940.

Behavioral Risk Factor Surveillance System (BRFSS). http://www.cdc.gov/BRFSS.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 20, 37-46.

Binder, D.A. (1996). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology*, 22, 17-22.

Carrasco, J.L., and Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*, 59, 849-858.

Chinchilli, V.M., Martel, J.K., Kumanyika, S. and Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics*, 52, 341-353.

Cochran, W.G. (1963). *Sampling Techniques*, 2nd Ed. New York: John Wiley & Sons, Inc.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.

Dixon, A.E., Sugar, E.A., Zinreich, S.J., Slavin, R.G., Corren, J., Naclerio, R.M., Ishii, M., Cohen, R.I., Brown, E.D., Wise, R.A. and Irvin, C.G. (2009). Criteria to screen for chronic sinonasal disease. *Chest*, 136 (5), 1324-1332.

Feinleib, M., Garrison, R.J., Fabsitz, R.R., Christian, J.C., Hrubec, Z., Borhani, N.O., Kannel, W.B., Roseman, R., Schwartz, J.T. and Wagner, J.O. (1977). The NHLBI Twin Study of cardiovascular disease risk factors: Methodology and summary of results. *American Journal of Epidemiology*, 106, 284-295.

Field, A.E., Aneja, P. and Rosner, B. (2007). The validity of self-reported weight change among adolescents and young adults. *Obesity*, 15, 2357-2364.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Edition. New York: John Wiley & Sons, Inc.

Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, Inc.

Fleiss, J.L., Nee, J.C.M. and Landis, J.R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86, 974-977.

Gilmour, E., Ellerbrock, T.V., Koulos, J.P., Chiasson, M.A., Williamson, J.M., Kuhn, L. and Wright, T.C. (1997). Measuring cervical ectopy: Direct visual assessment versus computerized planimetry. *American Journal of Obstetrics and Gynecology*, 176, 108-111.

Gorber, S.C., Tremblay, M., Moher, D. and Gorber, B. (2007). A comparison of direct *vs.* self-report measures for assessing height, weight and body mass index: A systematic review. *Obesity Review*, 8, 373-374.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons, Inc. Vols I and II.

Khawaja, I.S., Olson, E.J., van der Walt, C., Bukartyk, J., Somers, V., Dierkhising, R. and Morgenthaler, T.I. (2010). Diagnostic accuracy of split-night polysomnograms. *Journal of Clinical Sleep Medicine*, 6 (4), 357-362.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Klar, N., Lipsitz, S.R. and Ibrahim, J.G. (2000). An estimating equations approach for modeling kappa. *Biometrical Journal*, 42, 45-58.

Kocks, J.W., Kerstjens, H.A., Snijders, S.L., de Vos, B., Biermann, J.J., van Hengel, P., Strijbos, J.H., Bosveld, H.E. and van der Molen, T. (2010). Health status in routine clinical practice: validity of the clinical COPD questionnaire at the individual patient level. Health and Quality of Life Outcomes. 8, 135-141.

Korten, A.E., Jorm, A.F., Henderson, A.S., McCusker, E. and Creasey, H. (1992). Control-informant agreement on exposure history in case-control studies of Alzheimer's disease. *International Journal of Epidemiology*, 21, 1121-1131.

Liang, K.Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255-268.

Lin, L. (1992). Assay validation using the concordance correlation coefficient. *Biometrics*, 48, 599-604.

Lin, L., and Chinchilli, V. (1997). Rejoinder to the letter to the editor from Atkinson and Nevill. *Biometrics*, 53, 777-778.

Liottol, N., Radaelli, T., Orsi1, A., Taricco, E., Roggerol, P., Giann, M.L., Consonni, D., Mosca1, F. and Cetin, I. (2010). Relationship between in utero sonographic evaluation and subcutaneous plicometry after birth in infants with intrauterine growth restriction: An exploratory study. *Italian Journal of Pediatrics*, 36, 70-77.

Lipsitz, S.R., Laird, N.M. and Brennan, T.A. (1994). Simple moment estimates of the κ -coefficient and its variance. *Applied Statistics*, 43, 309-323.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis using R. Hoboken*. New Jersey: John Wiley & Sons, Inc.

MacDougall, H.G., Weber, K.P., McGarvie, L.A., Halmagyi, G.M. and Curthoys, I.S. (2009). The video head impulse test. Diagnostic accuracy in peripheral vestibulopathy. *Neurology*, 73, 1134-1141.

Maclure, M., and Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, 126, 161-169.

Mazaheri, Y., Hricak, H., Fine, S.W., Akin, O., Shukla-Dave, A., Ishill, N.M., Moskowitz, C.S., Grater, J.E., Reuter, V.E., Zakian, K.L., Touijer, K.A. and Koutcher, J.A. (2009). Prostate tumor volume measurement with combined T2-weighted imaging and diffusion-weighted MR: Correlation with pathologic tumor volume. *Radiology*, 252 (2), 449-457.

National Center for Health Statistics (2011). Third National Health and Nutrition Examination Survey, 1988-1994, NHANES III Examination data file (CD-ROM). http://www.cdc.gov/ nchs/nhanes.htm.

National Health Interview Survey (NHIS) (2011). http://www. cdc.gov/nchs/ nhis.htm.

Robieson, W. (1999). On weighted kappa and concordance correlation coefficient. Ph.D. thesis, University of Illinois in Chicago/Graduate College/Mathematics.

Rust, K.F., and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Sherry, B., Jefferds, M.E. and Grummer-Strawn, L.M. (2007). Accuracy of adolescent self-report of height and weight in assessing overweight status: A literature review. *Archive of Pediatrics Adolescent Medicine*, 161, 1154-1161.

Shield, M., Gorber, S.C. and Tremblay, M.S. (2008). Effects of measurement on obesity and morbidity. *Health Reports*, 19, 77-84.

Slade, G.D., and Beck, J.D. (1999). Plausibility of periodontal disease estimates from NHANES III. *Journal of Public Health Dentistry*, 59, 67-72.

Talamayan, K.S., Springer, A.E., Kelder, S.H., Gorospe, E.C. and Joye, K.A. (2006). Prevalence of overweight misperception and weight control behaviors among normal weight adolescents in the United States. *The Scientific World Journal*, 6, 365-373.

Tanner, M.A., and Young, M.A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association*, 80, 175-180.

Williamson, J.M., Manatunga, A.K. and Lipsitz, S.R. (2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics*, 1, 191-202.

Winn, D.M., Johnson, C.L. and Kingman, A. (1999). Periodontal disease estimates in NHANES III: Clinical measurement and complex sample design issues. *Journal of Public Health Dentistry*, 59, 73-78.

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.

Zar, J.H. (1996). *Biostatistical Analysis. Upper Saddle River*. New Jersey: Prentice Hall International.