HHS Public Access

Author manuscript

Int J Biostat. Author manuscript; available in PMC 2019 April 30.

Published in final edited form as:

Int J Biostat. ; 9(1): . doi:10.1515/ijb-2012-0001.

Assessing Agreement of Repeated Binary Measurements with an Application to the CDC's Anthrax Vaccine Clinical Trial

Yi Pan,
Immunization Safety Office, Division of Healthcare Quality Promotion, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA; Logistics Health Inc, La Crosse, WI, USA, jun5@cdc.gov

Charles E. Rose,
Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA, cvr7@cdc.gov

Michael Haber,
Department of Biostatistics and Bioinformatics, Emory University, Rollins School of Public Health, Atlanta, GA, USA, mhaber@emory.edu

Yan Ma,
Biostatistics, Public Health Department, Hospital for Special Surgery, Weill Medical College of Cornell University, New York City, NY, USA, yam2007@med.cornell.edu

Josep L. Carrasco,
University of Barcelona, Barcelona, Spain, jlcarrasco@ub.edu

Brock Stewart,
Immunization Safety Office, Division of Healthcare Quality Promotion, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA, jnn6@cdc.gov

Wendy A. Keitel,
Baylor College of Medicine, Houston, TX, USA, wkeitel@bcm.edu

Harry Keyserling,
Emory University School of Medicine, Atlanta, GA, USA, hkeyser@emory.edu

Robert M. Jacobson,
Department of Pediatric and Adolescent Medicine, Mayo Clinic, Rochester, MN, USA, jacobson.robert@mayo.edu

[*]**Corresponding author:**.

Conflict of interest

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention (CDC). Mention of a product or company name does not constitute endorsement by the CDC. The protocol for this study was approved by an Institutional Review Board of the CDC. Dr. Gregory Poland is the chair of a Safety Evaluation Committee for novel investigational vaccine trials being conducted by Merck Research Laboratories. Dr. Poland offers consultative advice on vaccine development to Merck & Co. Inc., CSL Biotherapies, Avianax, Sanofi Pasteur, Dynavax, Novartis Vaccines and Therapeutics, PAXVAX Inc, and Emergent Biosolutions. These activities have been reviewed by the Mayo Clinic Conflict of Interest Review Board and are conducted in compliance with Mayo Clinic Conflict of Interest policies. This research has been reviewed by the Mayo Clinic Conflict of Interest Review Board and was conducted in compliance with Mayo Clinic Conflict of Interest policies. Rest of the authors have conflicts of interest to declare.

Gregory Poland, and
Mayo Clinic, Rochester, MN, USA, poland.gregory@mayo.edu

Michael M. McNeil*
Immunization Safety Office, Division of Healthcare Quality Promotion, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention; Centers for Disease Control and Prevention, 1600 Clifton Road NE, Atlanta, GA 30333, mmm2@cdc.gov

Abstract

Cohen's kappa coefficient, which was introduced in 1960, serves as the most widely employed coefficient to assess inter-observer agreement for categorical outcomes. However, the original kappa can only be applied to cross-sectional binary measurements and, therefore, cannot be applied in the practical situation when the observers evaluate the same subjects at repeated time intervals. This study summarizes six methods of assessing agreement of repeated binary outcomes under different assumptions and discusses under which condition we should use the most appropriate method in practice. These approaches are illustrated using data from the CDC anthrax vaccine adsorbed (AVA) human clinical trial comparing the agreement for two solicited adverse events after AVA between the 1–3 day in-clinic medical record and the patient's diary on the same day. We hope this article can inspire researchers to choose the most appropriate method to assess agreement for their own study with longitudinal binary data.

Keywords

agreement; binary; kappa; longitudinal; adverse event; anthrax vaccine

1 Introduction

In epidemiologic and medical science, many statistical approaches have been proposed for assessing agreement among different observers or measurement methods. For categorical measurements, Cohen's kappa [1] and weighted kappa [2] are the most popular indices of agreement. For quantitative data, a very popular unscaled agreement index is the limits of agreement proposed by Bland and Altman [3]. More recently, the commonly used scaled agreement coefficients are the intraclass correlation coefficient (ICC) [4–7] and the concordance correlation coefficient (CCC) [8].

In many modern-day applications, data are often clustered, making inference difficult to perform. In addition, longitudinal studies where repeated observations are recorded for one subject by each observer at different time points have become increasingly more common. Our objective is to review and summarize the available methods that can be used to assess agreement of repeated binary outcomes. This study can serve a very useful tool whenever agreement needs to be assessed from repeated binary measurements under different scenarios. To illustrate the methods for obtaining agreement coefficients for repeated binary outcomes, we use data from the CDC anthrax vaccine adsorbed (AVA) human clinical trial [9], and we provide an overview of this in the following section. Details of each statistical method are presented in Section 3. Results of applying each of the statistical methods to data

for redness and tenderness at the site of vaccination from the CDC AVA human clinical trial are shown in Section 4. Summary and discussion follow in Section 5.

2 Motivating example

The CDC AVA human clinical trial was conducted during 2002–2005 with participants enrolled and followed at five major U.S. vaccine research centers [9]. This was a Phase 4, multicenter, randomized, placebo-controlled clinical trial to evaluate route change [subcutaneous (SQ) to intramuscular (IM)] and dose reduction (priming schedule of 0, 2, 4 weeks and 6, 12, 18 months and an annual booster vs reduced priming schedule of 0, 4 weeks and 6 months and a triennial booster). At each site, participants were randomly assigned to one of seven arms (TRT-8SQ, TRT-8IM, TRT-7IM, TRT-5IM, TRT-4IM, CNT-8IM, and CNT-8SQ) based on treatment (AVA vs saline placebo), route (SQ vs IM), and full/reduced AVA schedule (full = 0.5 mL doses at 0, 2, and 4 weeks, and 6, 12, 18, 30, and 42 months and reduced = substituting one/more placebo doses). Details of the clinical trial and the results of the interim analysis on data collected through the first four AVA doses for the first 1,005 participants were previously published [9].

Participants (n = 1,563) received a total of eight injections of AVA or saline placebo during 43 months (Figure 1). Following each injection, participants were routinely monitored (a) using a self-reported adverse event diary for 14 days after each of the first two doses and for 28 days after all subsequent doses and (b) by a study nurse during an in-clinic interview and exam at 15–60 min (not shown in Figure 1) and 1–3 days after all injections, 14 days after the first two injections, and 28 days after injections 3–8. Eight solicited injection-site adverse events (warmth, tenderness, itching, pain, arm motion limitation, redness, swelling, and bruise) were recorded both in the clinic record and separately by the individual participant in their diary. For our agreement methods' comparison, we restricted the dataset to only participants from study site A (n = 299) and compared the agreement for two solicited adverse events, redness and tenderness, at the injection site, in the in-clinic record, and in the participant's diary. We further limited our dataset to include only the record obtained at the participant's in-clinic visit scheduled for 1–3 days following each vaccination, and we compared redness and tenderness indicated in the clinical record and in the participant's diary on the exact same date. This time point was chosen to maximize the agreement of having redness and tenderness at the injection site recorded by both observers. We did not evaluate data from the early (15–60 min) in-clinic assessment, since the participants were instructed to use the diary to only record adverse events occurring later in the day after their in-clinic visit or the much later (14/28 days) in-clinic evaluations.

Table 1 presents an example of the data layout of redness. The redness and tenderness measurements of one subject were from both diary and in-clinic measurements after each of the 8 injections, consisting of 16 observations. In addition, important covariates such as treatment arm, age, and gender are also included.

3 Methods

The following agreement approaches were considered: (1) modeling the observed agreement with generalized estimating equation (GEE) by Coughlin et al. [10]; (2) extended kappa statistic with two-stage logistic regression by Lipsitz et al. [11]; (3) extended kappa statistic based on U-statistics by Ma et al. [12]; (4) ICC on repeated measurements by Carrasco et al. [13]; (5) CCC on repeated measurements with U-statistics by King et al. [14]; and (6) weighted CCC on repeated measurements with variance components by Carrasco et al. [13].

In the following discussion of methods, the common notation is applied. Let y_{ijt} denote the binary readings from the ith subject, jth observer at time point t, where $i = 1,\ldots, n$, denoting the subject index; $j = 1, 2$, standing for the two observers and $t = 1,\ldots, T$. For example, if $y_{i1t} = y_{i2t}$, then the new defined dependent variable $z_{it} = 1$ at time point t. Otherwise $z_{it} = 0$. Define $P(z_{it} = 1) = p_{it}$. In addition, some baseline covariates $X_1, X_2, \ldots X_M$ were also included in some methods.

3.1 Logistic regression modeling of the agreement proportion

Coughlin et al. [10] introduced logistic modeling of inter-observer agreement. The dependent variable is defined to be 1, if the two raters agree and 0 otherwise. Based on the notations, the logistic regression to estimate the percent agreement, adjusting for covariates in order to obtain adjusted or subgroup-specific estimates of percent agreement can be expressed as follows:

$$\log it(p_{it}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_M X_M,$$

where β_0 is the intercept, β_1 is the coefficient of covariate X_1 and so on. GEE with unstructured correlation was applied to estimate model-based percent agreement.

If each of N subjects is assigned independently by two raters to one of the I categories, then the cell frequencies (n_{iit}) along the main diagonal of the two-way contingency table represent the agreement between the raters, at time point t. The crude agreement is estimated as follows:

$$p_{0t} = \frac{1}{N} \sum_{i=1}^{I} n_{iit}.$$

By applying this logistic regression approach, the proportion of agreement for particular subgroups can be estimated. Suppose there are M explanatory variables ($X_1, X_2, \ldots, X_M$), then the model-based agreement is

$$E(p_{0t} \mid X_1, X_2, \ldots, X_M) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{m=1}^{M} \beta_m X_m\right)\right]}.$$

The variance of the logit of the proportion agreement can be estimated using the sandwich estimator [15–17], and inference is carried out using standard normal approaches.

3.2 Extended Kappa statistic based on two-stage logistic regression

Although the estimation of the agreement proportion and its interpretation [10] is straightforward, two observers can simply agree by chance [11]. Lipsitz et al. [11] proposed two-stage logistic regression to estimate kappa, and this can also be applied to repeated binary measurements. Kappa was introduced by Cohen [1] to assess the agreement of two methods/observers having binary readings, and it is defined as $\kappa = \frac{\pi_0 - \pi_e}{1 - \pi_e}$, estimated by $\hat{\kappa} = \frac{p_0 - p_e}{1 - p_e}$ where π_0 denotes the observed agreement and π_e denotes the agreement expected by chance, which is also the agreement under independence.

Let $P(Y_{i1t} = 1) = p_{i1t}$ which denotes the probability of the first rater having the measurements as "1". Similarly, $P(Y_{i2t} = 1) = p_{i2t}$ and $\hat{p}_{i1t}$, $\hat{p}_{i2t}$ are the corresponding estimates based on the following logistic regressions.

$\log it(p_{i1t}) = \beta_{01} + \beta_{11}X_1 + \beta_{12}X_2 + \cdots + \beta_{1M}X_M$, where β_{01} is the intercept and β_{11} is the coefficient of variable time, which is denoted by X_1 and similarly, $\log it(p_{i2t}) = \beta_{02} + \beta_{21}X_1 + \beta_{22}X_2 + \cdots + \beta_{2M}X_M$. In summary, there are two steps that are considered to assess the agreement between two observers:

1. Use the standard logistic regression to obtain $\hat{p}_{i1t}$ and $\hat{p}_{i2t}$.
2. Form the estimated offset, $\hat{\eta}_{it} = \log it[(\hat{p}_{i1t}\hat{p}_{i2t} + (1 - \hat{p}_{i1t})(1 - \hat{p}_{i2t})]$.

Finally, the model is $\log it(p_{it}) = \eta_{it} + \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_MX_M$, where η_{it} is the pre-specified offset, β_0 is the intercept, β_1 is the coefficient of X_1, β_2 is the coefficient of X_2 and so on. A summary measure of how agreement differs from chance agreement, for any given covariate pattern is included by the estimated linear predictor, $\pi_{eit} = \log it(p_{it}) - \eta_{it}$. Lipsitz et al. [11] showed that the estimate of the kappa coefficient is

$$\hat{k}_{it} = \frac{\hat{p}_{it} - e^{\hat{\eta}_{it}}/[1 + e^{\hat{\eta}_{it}}]}{1 - e^{\hat{\eta}_{it}}/[1 + e^{\hat{\eta}_{it}}]} = \frac{\hat{p}_{it} - [\hat{p}_{i,\text{rater}=1,t}\hat{p}_{i,\text{rater}=2,t} + (1 - \hat{p}_{i,\text{rater}=1,t})(1 - \hat{p}_{i,\text{rater}=2,t})]}{1 - [\hat{p}_{i,\text{rater}=1,t}\hat{p}_{i,\text{rater}=2,t} + (1 - \hat{p}_{i,\text{rater}=2,t})(1 - \hat{p}_{i,\text{rater}=2,t})]}.$$

Although the jackknife estimator was originally proposed for this method, in our study we applied the bootstrap standard error approach. GEE with unstructured correlation was also used. To ensure the results in the two stages were consistent, any data point with only one observation in the in-clinic record or in the diary was not included. And we only adjusted for time in all the models fitted under this approach.

3.3 Extended Kappa statistic based on U-statistics

Ma et al. [12] introduced a new class of kappa coefficients based on U-statistics to tackle the complexities involved in addressing missing data and other related issues arising from a multirater scenario and repeated categorical measurements. For illustration purposes, we only consider a longitudinal study with n subjects, j raters, t assessments, and a binary outcome g. The addition of index g is to create a dummy variable of each Y_{ijt}. For example,

if $y_{ijt} = 1$, then $y_{ijt1} = 1$ ($g = 1$) and $y_{ijt0} = 0$ ($g = 0$). On the other hand, if $y_{ijt} = 0$, then $y_{ijt1} = 0$ ($g = 1$) and $y_{ijt0} = 1$ ($g = 0$). The motivation of such a notation is to accommodate the missing data structure. The estimate of kappa at time t is given by the ratio of two U-statistics:

$$\kappa_t = \frac{\sum_{(i,u)\in C_2^n} \frac{1}{2}\sum_{k=0}^{1}\sum_{g=0}^{1}(y_{i1tk} - y_{u1tk})(y_{i2tg} - y_{u2tg})}{\sum_{(i,u)\in C_2^n}\{1 - \frac{1}{2}\sum_{k=0}^{1}\sum_{g=0}^{1}(y_{i1tk}y_{u2tg} + y_{u1tk}y_{i2tg})\}}, \quad t = 1, 2, \ldots, T, \quad (1)$$

where i and u are two subjects belonging to $C_2^n = \{(i,u); 1 \le i < u \le N\}$.

By the theory of U-statistics and the delta method, the proposed estimate of kappa is proved to be consistent and asymptotically normal. Further, the U-statistics based estimate in eq. [1] can be modified to account for missing data:

$$\kappa_t = \frac{\sum_{(i,u)\in C_2^n} \frac{r_{i1t}r_{u1t}r_{i2t}r_{u2t}}{2\Delta_{it}\Delta_{ut}}\sum_{k=0}^{1}\sum_{g=0}^{1}(y_{i1tk} - y_{u1tk})(y_{i2tg} - y_{u2tg})}{\sum_{(i,u)\in C_2^n}\frac{r_{i1t}r_{u1t}r_{i2t}r_{u2t}}{\Delta_{it}\Delta_{ut}}\{1 - \frac{1}{2}\sum_{k=0}^{1}\sum_{g=0}^{1}(y_{i1tk}y_{u2tg} + y_{u1tk}y_{i2tg})\}}, \quad t = 1, 2, \ldots,$$

T,

(2)

where $r_{ijt} = 1$ if the jth rater's rating on the ith subject is observed at the tth assessment and $r_{ijt} = 0$ if the rating is missing, and Δ_{it} represents the probability of missing data. This estimate is designed for modeling kappa under two missing data patterns – missing completely at random (MCAR) and missing at random (MAR). MCAR means that the missing data are independent of both the observed and the unobserved variables. On the other hand, MAR means that given the observed data, missingness does not depend on the unobserved data.

By the theory of multivariate U-statistics, the joint distribution of the kappas assessed at multiple time points is readily derived. Inferences for longitudinal kappas can be further developed based on their joint distribution. For example, one of the research interests in practice is to identify the trend in agreement over time. In particular, a test of equal agreements (i.e. $\kappa_1 = \kappa_2 = \cdots = \kappa_T$) is proposed in this method.

When applying this method to our illustrative dataset, we made the following assumptions as the major missing data pattern in our study: (1) if the measurement was missing in the in-

clinic record, then it was also missing for the participant's diary (missing simultaneously); (2) if the record of injection-site redness was missing at one vaccine dose, then it was also missing for all the participant's later doses, representing a monotone missing data pattern (MMDP). In fact in our study, all missingness occurred in both in-clinic and diary records simultaneously and 90% of the missing observations followed MMDP. Missing in-clinic record and diary observations were due to a participant's missed in-clinic visit and participant's negligence, respectively. Furthermore, the occurrence of missing data in our study did not relate to either prior or future observations. Therefore, we observed a MCAR missing pattern in our study, by which Δ_{it} in eq. [2] is constant and can be estimated by the sample proportion $\frac{\sum_{i=1}^{n} r_{i1t} r_{i2t}}{n}$.

3.4 ICC for binary repeated measurements

Historically, agreement between quantitative measurements has been evaluated via the ICC. Numerous versions of ICCs [4–7, 18–20] have been proposed in many areas of research by assuming different underlying analysis of variance (ANOVA) models for the situation where none of the observers is treated as the reference. The simplest ICC is defined in the following ANOVA model:

$Y_{ijt} = \mu + \alpha_i + \varepsilon_{ij}$ with assumptions: $\alpha_i \sim N(0, \delta_\alpha^2)$; $\varepsilon_{ij} \sim N(0, \delta_\varepsilon^2)$ and ε_{ij} is independent of α_i, where $i = 1, \ldots, n$, denoting the subject index; $j = 1, \ldots, k_i$, standing for the observer index. Then, $\mathrm{ICC} = \frac{\delta_\alpha^2}{\delta_\alpha^2 + \delta_\varepsilon^2}$ and its estimate is $\widehat{\mathrm{ICC}} = \frac{\mathrm{MS}_\alpha - \mathrm{MS}_\varepsilon}{\mathrm{MS}_\alpha + (K-1)\mathrm{MS}_\varepsilon}$, where MS_α and MS_ε are the mean sums of squares from the one-way ANOVA model for between and within subjects, respectively. This method has been extended to binary data just coding Y_{ijt} as 0 or 1 [21]. For binary measurement,

$$\widehat{\mathrm{ICC}}_{\mathrm{binary}} = \frac{\mathrm{MS}_\alpha - \mathrm{MS}_\varepsilon}{\mathrm{MS}_\alpha + (n_0 - 1)\mathrm{MS}_\varepsilon}, \text{ where } n_0 = \frac{1}{(n-1)}\left[K - \sum_{i=1}^{n} \frac{k_i^2}{K}\right] \text{with } K = \sum_{i=1}^{n} k_i.$$

Besides this ANOVA estimator, Ridout et al. [21] also introduced other estimation methods of ICC for binary observations such as the moment estimators, the quasi-likelihood and pseudo-likelihood estimators and the maximum likelihood estimator for beta-binomial data. Furthermore, Carrasco et al. [13] extended ICCs to repeated measurements. Consider the following linear mixed model:

$$Y_{ijt} = \mu + \alpha_i + \beta_j + \gamma_t + \alpha\beta_{ij} + \alpha\gamma_{it} + \beta\gamma_{jt} + e_{ijt}, \quad (3)$$

where μ is the overall mean, α_i is the random subject effect ($i = 1, \ldots, n$) assumed to be distributed as $\alpha_i \sim N(0, \sigma_\alpha^2)$, β_j is the fixed observer effect ($j = 1, \ldots, k$), γ_t is the fixed time effect ($t = 1, \ldots, p$), $\alpha\beta_{ij}$ is the random subject–observer interaction effect assumed to be

distributed as $\alpha\beta_{ij} \sim N(0, \sigma^2_{\alpha\beta})$, $\alpha\gamma_{it}$ is the random subject–time interaction effect assumed to be distributed as $\alpha\gamma_{ij} \sim N(0, \sigma^2_{\alpha\gamma})$, $\beta\gamma_{jt}$ is the fixed observer–time interaction effect, and e_{ijt} is the random error effect assumed to be distributed as $e_i \sim MVN(0, \sigma^2_e R)$ when e_i is the vector of residuals of each subject. All the effects of the model are assumed to be independent. This ANOVA model is a special case of linear mixed models. Thus, the appropriate expression of the ICC for measuring agreement between observers is

$$\mathrm{ICC}_2 = \frac{\sigma^2_\alpha + \sigma^2_{\alpha\gamma}}{\sigma^2_\alpha + \sigma^2_{\alpha\gamma} + \sigma^2_{\alpha\beta} + \sigma^2_{\beta\gamma} + \sigma^2_e}. \quad (4)$$

ICC_2 is estimated by replacing the variance components in eq. [4] by their corresponding estimates obtained using restricted maximum likelihood (REML) estimation.

3.5 CCC based on U-statistics

The CCC is commonly used for assessing agreement for continuous outcomes. It was first published by Lin [8] for the simplest case where each of two raters makes one reading per subject. Lin's CCC is defined as follows: assume that the observations are from a bivariate distribution with mean vector (μ_1, μ_2) and variance covariance matrix $\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, the Lin's CCC between two observers Y_1 and Y_2 is proposed as

$$\mathrm{CCC} = 1 - \frac{E(Y_2 - Y_1)^2}{E[(Y_2 - Y_1)^2 \mid \rho = 0]} = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}, \quad (5)$$

where ρ is the Pearson correlation coefficient between two observers. King et al. [14] introduced a generalized CCC for both continuous and categorical data. Furthermore, King et al. [22, 23] reported the extension of CCC to repeated measurements, mainly with repeated continuous data.

Let the elements of the vector $\mathbf{Y_1}$ (y_{i1t}) represent the tth repeated measure on the ith subject for the repeated measurements on the first observer. Let the elements of the vector $\mathbf{Y_2}$ (y_{i2t}) represent the tth repeated measure on the ith subject for measurements on the second observer. Assume that the elements of the vector $\mathbf{[Y_1, Y_2]}$ are selected from a multivariate normal population with $2T \times 1$ mean vector $[\mu_{Y_1}, \mu_{Y_2}]$, and $2T \times 2T$ covariance matrix $\sum$, which consists of the following four $T \times T$ matrices: $\sum_{Y_1 Y_1}$, $\sum_{Y_2 Y_2}$, $\sum_{Y_1 Y_2}$, and $\sum_{Y_2 Y_1}$.

Extending from the derivation of Lin's CCC shown in eq. [5], we can then construct a repeated measures CCC as

$$\mathrm{CCC}_{rm} = 1 - \frac{E[(Y_1 - Y_2)'D(Y_1 - Y_2)]}{E_I[(Y_1 - Y_2)'D(Y_1 - Y_2)]} = \frac{\mathrm{trace}(D\sum_{Y_1Y_2} + D\sum_{Y_2Y_1})}{\mathrm{trace}(D\sum_{Y_1Y_1} + D\sum_{Y_2Y_2} + (\mu_{Y_1} - \mu_{Y_2})'D(\mu_{Y_1} - \mu_{Y_2})}$$

$$= \frac{\sum_{t=1}^{T}\sum_{s=1}^{T} d_{ts}(\sigma_{Y_{1t}Y_{2s}} + \sigma_{Y_{2t}Y_{1s}})}{\sum_{t=1}^{T}\sum_{s=1}^{T} d_{ts}(\sigma_{Y_{1t}Y_{1s}} + \sigma_{Y_{2t}Y_{2s}}) + \sum_{t=1}^{T}\sum_{s=1}^{T} d_{ts}(\mu_{Y_{1t}} - \mu_{Y_{2s}})(\mu_{Y_{1t}} - \mu_{Y_{2s}})},$$

(6)

where **D** is a $T \times T$ non-negative definite matrix of weights between the different repeated measurements; both t and s index two sample moments between and within the measures being evaluated for agreement, i.e. when both t and s index the T repeated visits in a longitudinal study, $t = s$ represents the within-visit agreement, and $t \neq s$ represents the between-visit agreement between the two measures of interest. When $T = 1$, the repeated measures CCC reduces to Lin's simple concordance coefficient in eq. [5].

Following the extension in King et al. [22], eq. [6] has been extended to repeated categorical outcomes as:

$$\mathrm{CCC}_{rm} = 1 - \frac{\sum_{t=1}^{T}\sum_{s=1}^{T} d_{ts}\Pr(Y_{1t} \neq Y_{2t}, Y_{1s} \neq Y_{2s})}{\sum_{t=1}^{T}\sum_{s=1}^{T} d_{ts}\Pr I(Y_{1t} \neq Y_{2t}, Y_{1s} \neq Y_{2s})}.$$

Furthermore, in King et al. [22, 23], four options of **D** were proposed. In our study, we considered the following two options for D:

1. Weight 1: **D** = (d_{ts}), where d_{ts} is the non-missing proportions when $t = s$ and $d_{ts} = 0$ when $t \neq s$.
2. Weight 2: **D** = (d_{ts}), where $d_{ts} = T - t + 1$ when t = s and $d_{ts} = 0$ when $t \neq s$.

In the first set of weights, we used the actual non-missing proportion to give more weights to the time points with lower missing proportions. It is reasonable to assign more weight to the time point with more information. While the second set of weights gave us more weight to the earlier time points. The data from the first measurements are possibly more reliable, since there are more data. Therefore, more weight was assigned to the earlier time points in the second weight. A basic consideration for statistical inference concerning CCC_{rm} is to recognize that the estimator $\widehat{\mathrm{CCC}}_{rm}$ can be expressed as a ratio of functions of U-statistics.

CCC based on U-statistics is not applicable when missing data occur. In our study, if any observation for a participant was missing, then we excluded all that participant's data.

3.6 Weighted CCC based on variance components

Carrasco et al. [13] proposed a CCC for repeated binary measurements through the appropriate specification of the ICC from a variance components linear mixed model. Combining the notations in Sections 3.4 and 3.5, $\mathrm{CCC}_{rm} = 1 - \frac{E[(Y_1 - Y_2)'D(Y_1 - Y_2)]}{E_1[(Y_1 - Y_2)'D(Y_1 - Y_2)]}$, where **D** is an identity matrix, the expression of CCC under model (1) can be reduced to $\frac{(\sigma_\alpha^2 + \sigma_{\alpha\gamma}^2)\sum_{j=1}^{T} d_{tt}}{(\sigma_\alpha^2 + \sigma_{\alpha\gamma}^2 + \sigma_{\alpha\beta}^2 + \sigma_e^2)\sum_{j=1}^{T} d_{tt} + \frac{1}{2}\sum_{t=1}^{T} d_{tt}(\mu_{1t} - \mu_{2t})^2}$. In particular, the case of **D** as an identity matrix where $d_{tt} = 1$ if $t =$ s and 0 otherwise, then $\sum d_{ts} = T$ and the CCC becomes $\mathrm{CCC}_{rm} = \frac{\sigma_\alpha^2 + \sigma_{\alpha\gamma}^2}{\sigma_\alpha^2 + \sigma_{\alpha\gamma}^2 + \sigma_{\alpha\beta}^2 + \sigma_{\beta\gamma}^2 + \sigma_e^2} = \mathrm{ICC}_2$ as shown in eq. [4]. Besides the two weights, compound symmetry (CS) and first-order auto correlation (AR(1)) structures were considered with the CCC variance components method.

4 Results

Of the total trial's enrollment, 299 participants were enrolled at study site A. The mean age of these participants was 42.2 years with standard deviation 10.1 years. Fifty-one percent (155/299) of participants were female. Participants by treatment arms were 48 (16.1%) in 8 IM; 50 (16.7%) in 7 IM; 51 (17.1%) in 5 IM; 52 (17.4%) in 4 IM; 24 (8.0%) in the 8 IM placebo arm, and the remaining 24 (8.0%) in the 8 SQ placebo arm.

Tables 2 and 3 present the result of modeling agreement proportions of redness and tenderness on the 299 participants enrolled at study center A. A univariate model with only time effect and a multivariate model with age, gender, treatment arm, and time effect were considered. In both models, all data at each time point were fitted using a single model. In the univariate model, only time effect was included with GEE and unstructured correlation. The significance of the time effect indicated that agreement proportions changed significantly across visits ($p = 0.003$ for redness and $p = 0.005$ for tenderness). In the multivariable model for redness, age ($p = 0.02$), visit time ($p = 0.002$), and treatment arm ($p < 0.0001$) were found to be significantly associated with agreement and for tenderness, only visit time ($p = 0.005$) and arm ($p < 0.0001$) were significant. Both univariate and multivariable models showed high agreement proportions. After adjusting for age, gender, and treatment arm, the overall agreement proportions were almost all above 90% for redness and above 85% for tenderness. However, two observers can simply agree with each other by chance. The observed agreement proportion does not tell us the complete story. Therefore, we used the modified kappa statistic which takes chance agreement into account. Extended kappa coefficients based on two-stage logistic regression models and U-statistics are presented in Tables 4 and 5, respectively. For redness, kappa statistic based on two-stage logistic regression was 0.719 with 95% CI (0.597, 0.833) at AVA dose 1, while it was 0.630 with 95% CI (0.522, 0.738) at the AVA dose 8. Kappa coefficient based on U-statistics gave us similar results. Starting with a value of 0.720 with 95% CI (0.591, 0.849) at AVA dose 1 kappa decreased to 0.646 with 95% CI (0.521, 0.771) at AVA dose 8. Kappa statistics were not significantly different across time ($p = 0.8$) for redness. A kappa value ranging from 0.2

to 0.4 indicates fair agreement, 0.4 to 0.6 means moderate agreement, 0.6 to 0.8 means good agreement, and greater than 0.8 means excellent agreement [24]. Therefore, good agreement was achieved when comparing the records of injectionsite redness from both patients' diaries and their in-clinic records. On the other hand, the kappa coefficients of tenderness ranged from 0.523 to 0.667 based on the two-stage logistic regression (Table 5). Kappa based on U-statistic gave us comparable results except for the seventh time point where U-statistic gave us 0.627, and kappa based on two-stage logistic regression was 0.584. Kappa statistics were not significantly different across time ($p = 0.7$) for tenderness.

Table 6 combines the results of ICC with CS and AR(1) correlation structures, CCC based on U-statistics with two weights and CCC based on variance components with CS and AR(1) correlations and two weights specified in Section 3.5. Here, weight 1 based on non-missing proportions is (0.98, 0.95, 0.95, 0.91, 0.90, 0.89, 0.82, 0.77) and weight 2 is simply (8, 7, 6, 5, 4, 3, 2, 1). Overall agreement across eight doses is presented for each method, and CCC based on U-statistics with weight 2 gave us the highest estimate 0.7090 with 95% CI (0.6614, 0.7510) for redness. Furthermore, more discrepancy was observed for tenderness applying different methods and weighting schemes. ICC and CCC based on variance components with weight 2 gave us almost identical results. However, CCC based on variance components applying weight 1 gave us much higher agreement of tenderness between in-clinic interview and patients' diaries. For example, CCC_VC with AR(1) was 0.6861 with weight 1 and 0.5943 with weight 2. CCC based on U-statistic gave us comparable results no matter which weight we choose from (Table 7).

5 Discussion

In this study, six approaches for measuring agreement on repeated binary outcomes were reviewed and applied to our illustrative dataset. In general, it is apparent that those methods should be applied under different situations. In Table 8, we list the details, main characteristics of each method and when we recommend using each of them. Generally speaking, those six approaches can be classified into the following categories by their model assumptions: (1) full parametric model which is based on linear mixed models and is estimated via maximum likelihood paradigm, such as ICC and Weighted CCC based on variance components; (2) semi-parametric model which is based on GEE, such as logistic regression modeling of the agreement proportion and extended kappa statistic based on two-stage logistic regression; and (3) nonparametric model, which is based on U-statistic, such as extended kappa and CCC based on U-statistic. The sample programs can be found under http://web1.sph.emory.edu/observeragreement/ and requested from the authors.

In practice, estimating the agreement proportion has been widely used in medical research [25–27]. Besides the unstructured correlation, other correlation structures such as AR(1) and CS can also be considered. Furthermore, important covariates that may affect agreement proportions can be included in the logistic modeling. However, this method does not overcome potential limitations of general observations of agreement such as the tendency for percent agreement to be high whenever the frequency of a particular diagnostic category is very low or very high. Therefore, estimates of kappa, which take chance agreement into account, may be preferable to the agreement proportion method.

The two extended kappa-based methods serve as good alternatives to modeling only the agreement proportion. In Lipsitz's logistic regression model for chance-corrected agreement, the offset term ensures that agreement due to chance is properly accounted for when attempting to identify covariates that are predictive of agreement. In this method, any data point with only an observation in the in-clinic record or in the diary was not included. Although results adjusting for covariates are not presented, covariates may also potentially be included in the two-stage logistics model. Besides the GEE method, which accommodates for the correlations among different time points for the same subjects, random effects models can be applied. On the other hand, Ma et al. [12] developed an approach to address missing data when modeling multiobserver kappa within a longitudinal study setting. In particular, they extended MMDP assumption for longitudinal data analysis involving a single response to a bivariate setting and integrated the inverse probability weighting approach with the theory of U-statistics. The missingness in "redness" follows MMDP and MCAR. Comparing those two extended kappa coefficients, model-based estimation requires more assumptions to correctly specify the model while the U-statistics approach is nonparametric.

CCC based on U-statistics can be applied as an extension of Lin's CCC to responses measured repeatedly over time, or clustered by some other design. This method can handle any number of repeated measurements, and the variance can be estimated in a straightforward manner by U-statistics methodology. Furthermore, the CCC based on U-statistics is not applicable, when any missing data appear and when the design is unbalanced. It indicated that if there was one missing data point in any method at any time point, the whole subject will be removed from the dataset. On the other hand, ICC and CCC based on variance components approaches were built up with the random effects model described in eq. [3]. By eq. [4], ICC can be considered as a special case (un-weighted version) of the weighted CCC when the **D** is an identity matrix. In addition, we found that the standard error is substantially higher for both CCC_U methods than for ICC and CCC_VC approaches. It may be possible that ICC and CCC_VC underestimate the SE, so that a smaller SE means worse performance in this case. This finding is consistent with Carrasco et al. [28]. So, perhaps the VC approach, or more properly the maximum likelihood approach, gives smaller SE, but this does not always mean "better performance". Furthermore, the ICC and CCC_VC can accommodate multiple raters and possible covariates in the model.

As explained in Section 3.4, one of the approaches Ridout et al. [21] evaluated ICC based on the linear mixed model (ANOVA) with the binary data coded as 0 and 1. In another more recent article on agreement, Zou and Donner [29] estimated ICC for binary data in the same way. Actually, it is not uncommon to use a linear mixed model to estimate variance components for binary data. In estimating those variance components, the categorical or binary nature of the response is usually ignored, and the analysis is carried out using ANOVA or mixed models. The rule-of-thumb generally applied is that the ANOVA is reasonably accurate, as long as the proportions in each of the categories of the response are not extreme. The variance components approach in Carrasco et al. [13] is similar to that of these articles, because the fixed effects and variance components are estimated from the

linear mixed model by restricted maximum likelihood (REML) which gives the same estimates than ANOVA in case of balanced design.

In summary, all six statistical methods give us comparable estimates, indicating good agreement for assessing the record of redness and moderate to good agreement on tenderness between participants' diaries and their in-clinic record. However, each approach has its own pros and cons under different situations. This article provides several alternatives for assessing the agreement with longitudinal binary data. We hope this article can inspire researchers to choose the most appropriate method to assess agreement for their own study.

Acknowledgments:

The authors thank Dr. David G. Kleinbaum and other colleagues at CDC for their helpful review of the manuscript. The authors thank Dr. Lawrence Lin for his very insightful suggestions for reviewing this manuscript. The authors also thank Kamesha Smith from ISO, CDC for drawing Figure 1.

References

1. Cohen J A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
2. Cohen J Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213–20. [PubMed: 19673146]
3. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–10.
4. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966;19:3C11. [PubMed: 5942109]
5. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420C428. [PubMed: 18839484]
6. Ebel RL. Estimation of the reliability of ratings. Psychometrika 1951;16:407–24.
7. Haggard EA. Intraclass correlation and the analysis of variance. New York: Dryden Press, 1958.
8. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255–68. [PubMed: 2720055]
9. Marano N, Plikaytis BD, Martin SW, et al. and Anthrax Vaccine Research Program Working Group. Effects of a reduced dose schedule and intramuscular administration of anthrax vaccine adsorbed on immunogenicity and safety at 7 months: a randomized trial. JAMA 2008;300:1532–43. [PubMed: 18827210]
10. Coughlin SS, Pickle LW, Goodman MT, Wilkens LR. The logistic modeling of interobserver agreement. J Clin Epidemiol 1992;45:1237–41. [PubMed: 1432004]
11. Lipsitz SR, Parzen M, Fitzmaurice GM, Klar N. A two-stage logistic regression model for analyzing inter-rater agreement. Psychometrika 2003;68:289–98.
12. Ma Y, Tang W, Feng C, Tu XM. Inference for kappas for longitudinal study data: applications to sexual health research. Biometrics 2008;64:781–9. [PubMed: 18047535]
13. Carrasco JL, King T, Chinchilli VM. The concordance correlation coefficient for repeated observations estimated by variance components. J Biopharm Stat 2009;19:90–105. [PubMed: 19127469]
14. King TS, Chinchilli VM. A generalized concordance correlation coefficient for continuous and categorical data. Stat Med 2001;20:2131–47. [PubMed: 11439426]
15. Diggle PJ, Liang KY, Zeger SL. Analysis of longitudinal data. Oxford: Clarendon Press, 1994.
16. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
17. Liang KY, Zeger SL, Qaqish B. Multivariate regression analysis for categorical data. J R Stat Soc Ser B 1992;54:3–40.

18. Muller R, Buttner P. A critical discussion of intraclass correlation coefficient. Stat Med 1984;13:2465–76.
19. Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: Using goniometric measurements as an example. Phys Ther 1994;74:777–88. [PubMed: 8047565]
20. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30–46.
21. Ridout MS, Demetrio CGB, Firth D. Estimating intraclass correlation for binary data. Biometrics 1999;55:137–48. [PubMed: 11318148]
22. King TS, Chinchilli VM, Wang KL, Carrasco JL. A class of repeated measures concordance correlation coefficients. J Biopharm Stat 2007;17:653–72. [PubMed: 17613646]
23. King TS, Chinchilli VM, Carrasco JL. A repeated observations concordance correlation coefficient. Stat Med 2007;26:3095–113. [PubMed: 17216594]
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74. [PubMed: 843571]
25. Phillips KA, Milne RL, Buys S, Friedlander ML, Ward JH, McCredie MRE, et al. Agreement between self-reported breast cancer treatment and medical records in a population-based breast cancer family registry. J Clin Oncol 2005;23:4679–86. [PubMed: 15851764]
26. Phillip D, Lyngberg AC, Jensen R. Assessment of headache diagnosis: a comparative population study of a clinical interview with a diagnostic headache diary. Cephalalgia 2007;27:1. [PubMed: 17212676]
27. Tannenbaum C, Brouillette J, Corcos J. Rating improvements in urinary incontinence: do patients and their physicians agree? Age Aging 2008;37:379–83.
28. Carrasco JL, Jover L, King T, Chinchilli VM. Comparison of concordance correlation coefficient estimating approaches with skewed data. J Biopharm Stat 2007;17:673–84. [PubMed: 17613647]
29. Zou G, Donner A. Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. Biometrics 2004;60:807–11. [PubMed: 15339305]

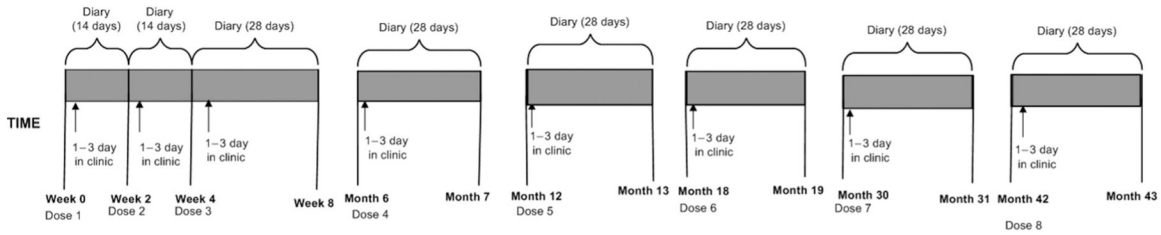

Figure 1.
Timeline of data collection of participants' diary and in-clinic evaluation.

Table 1

Data layout of repeated measures of redness in diary and clinic visits.

Study ID	Redness	Source	Injection	Treatment arm	Age	Gender
1	0	Diary	1	TRT-8IM	45	F
1	1	Diary	2	TRT-8IM	45	F
1	0	Diary	–	TRT-8IM	45	F
1	0	Diary	8	TRT-8IM	45	F
1	0	Clinic	1	TRT-8IM	45	F
1	1	Clinic	2	TRT-8IM	45	F
1	0	Clinic	–	TRT-8IM	45	F
1	0	Clinic	8	TRT-8IM	45	F
2	1	Diary	1	CNT-8SQ	37	M
2	1	Diary	2	CNT-8SQ	37	M
2	1	Diary	–	CNT-8SQ	37	M
2	0	Diary	8	CNT-8SQ	37	M
2	0	Clinic	1	CNT-8SQ	37	M
2	1	Clinic	2	CNT-8SQ	37	M
2	1	Clinic	–	CNT-8SQ	37	M
2	0	Clinic	8	CNT-8SQ	37	M

Table 2

Estimating agreement proportions of redness

Injection	Univariate model			Multivariable model		
	Estimate (%)	Lower (%)	Upper (%)	Estimate (%)	Lower (%)	Upper (%)
1	94.5	91.3	96.6	96.1	93.4	97.7
2	94.4	91.0	96.5	96.0	93.0	97.7
3	93.0	89.3	95.4	95.0	91.7	97.0
4	88.4	84.0	91.7	91.3	87.5	94.0
5	91.1	87.1	94.0	93.4	90.1	95.6
6	86.2	81.5	89.8	89.5	85.2	92.6
7	89.8	85.4	93.0	92.4	88.3	95.1
8	87.2	82.3	90.9	90.2	85.6	93.4

Notes: In the univariate model, time effect was significant, $p = 0.003$; In the multivariate model, age, gender, and arm were adjusted besides time. p values for time = 0.002, age = 0.02, gender = 0.3, and arm < 0.0001.

Table 3

Estimating agreement proportions of tenderness.

Injection	Univariate model			Multivariable model		
	Estimate (%)	Lower (%)	Upper (%)	Estimate (%)	Lower (%)	Upper (%)
1	79.8	74.9	84.0	86.4	81.7	90.0
2	88.3	84.0	91.5	92.6	89.1	95.0
3	78.1	72.9	82.5	85.1	80.3	88.9
4	79.9	74.7	84.2	86.5	82.0	90.1
5	86.9	82.4	90.4	91.6	88.1	94.2
6	83.6	78.6	87.6	89.3	85.1	92.4
7	84.9	79.9	88.8	90.3	85.9	93.4
8	78.0	72.3	82.8	85.1	80.0	89.1

Notes: In the univariate model, time effect was significant, $p = 0.005$; In the multivariate model, age, gender, and arm were adjusted besides time. p values for time = 0.005 and arm < 0.0001.

Table 4

Extended kappa statistics assessing agreement for redness.

Injection	Two-stage logistic				U-statistic*			
	Kappa	SE	95% CI		Kappa	SE	95% CI	
1	0.719	0.065	0.597	0.833	0.720	0.066	0.591	0.849
2	0.752	0.055	0.632	0.834	0.752	0.059	0.636	0.868
3	0.734	0.059	0.590	0.849	0.731	0.057	0.619	0.843
4	0.692	0.052	0.581	0.761	0.689	0.052	0.587	0.791
5	0.656	0.066	0.511	0.777	0.647	0.066	0.518	0.776
6	0.657	0.044	0.571	0.753	0.647	0.053	0.543	0.751
7	0.682	0.061	0.560	0.789	0.685	0.060	0.567	0.803
8	0.630	0.055	0.522	0.738	0.646	0.064	0.521	0.771

Notes:

* Test of equal kappa values, $p = 0.8$; A kappa value ranging from 0.2 to 0.4 indicates fair agreement, 0.4 to 0.6 means moderate agreement, 0.6 to 0.8 means good agreement, and greater than 0.8 means excellent agreement [24].

Table 5

Extended kappa statistics assessing agreement for tenderness.

Injection	Two-stage logistic				U-statistic*			
	Kappa	SE	95% CI		Kappa	SE	95% CI	
1	0.58	0.048	0.49	0.662	0.582	0.048	0.488	0.676
2	0.667	0.043	0.591	0.757	0.671	0.052	0.569	0.773
3	0.523	0.05	0.425	0.61	0.532	0.052	0.43	0.634
4	0.579	0.046	0.49	0.667	0.591	0.049	0.495	0.687
5	0.645	0.055	0.517	0.734	0.649	0.054	0.543	0.755
6	0.615	0.051	0.521	0.7	0.619	0.053	0.515	0.723
7	0.584	0.053	0.462	0.668	0.627	0.059	0.511	0.743
8	0.549	0.056	0.459	0.659	0.578	0.056	0.654	0.872

Note:

*Test of equal kappa values, $p = 0.7$.

Table 6

Assessing the overall agreement of redness between patients' diaries and their in-clinic visits with ICC, CCC based on U-statistics and variance components.

Methods	Correlation structure	Estimate	SE	95% CI	
ICC	CS	0.6861	0.0140	0.6577	0.7126
ICC	Autoregressive (1) (AR(1))	0.6862	0.0140	0.6578	0.7127
CCC_U, weight 1 *	NA	0.6942	0.0213	0.6497	0.7339
CCC_U, weight 2 **	NA	0.7090	0.0227	0.6614	0.7510
CCC_VC, weight 1	CS	0.6863	0.0140	0.6579	0.7127
CCC_VC, weight 1	AR(1)	0.6861	0.0140	0.6577	0.7126
CCC_VC, weight 2	CS	0.6870	0.0140	0.6587	0.7134
CCC_VC, weight 2	AR(1)	0.6869	0.0140	0.6585	0.7133

Notes:

* Weight 1 = (0.98, 0.95, 0.95, 0.91, 0.90, 0.89, 0.82, 0.77);

** Weight 2 = (8, 7, 6, 5, 4, 3, 2, 1).

Table 7

Assessing the overall agreement of tenderness between patients' diaries and their in-clinic visits with ICC, CCC based on U-statistics and variance components.

Methods	Correlation structure	Estimate	SE	95% CI	
ICC	CS	0.5943	0.0162	0.5616	0.6251
ICC	Autoregressive (1) (AR(1))	0.5937	0.0163	0.5609	0.6247
CCC_U, weight 1 *	NA	0.6285	0.0138	0.6006	0.6549
CCC_U, weight 2 **	NA	0.6360	0.0143	0.6069	0.6634
CCC_VC, weight 1	CS	0.6863	0.0140	0.6579	0.7127
CCC_VC, weight 1	AR(1)	0.6861	0.0140	0.6577	0.7126
CCC_VC, weight 2	CS	0.5937	0.0163	0.5609	0.6247
CCC_VC, weight 2	AR(1)	0.5943	0.0162	0.5615	0.6252

Notes:

* Weight 1 = (0.98, 0.95, 0.95, 0.91, 0.90, 0.89, 0.82, 0.77);

** Weight 2 = (8, 7, 6, 5, 4, 3, 2, 1).

Table 8

Comparison of six methods on assessing agreement of longitudinal binary data.

Methods	Summary	Main characteristics	Recommendations for use
[1] Agreement proportion	Logistic regression with GEE approach is used to estimate the crude agreement proportion.	(1) Covariates are allowed. (2) MCAR missing scheme is assumed. (3) Probability of agreement is estimated in a semi-parametric model.	Because the agreement proportion does not adjust for chance agreement as in kappa, it is only recommended for a crude agreement assessment.
[2] Kappa by two-stage logistic regression	It is the extension of estimating the agreement proportion. In the first step, chance agreement is estimated, and it is included as an offset in the second step.	(1) Kappa is estimated at each time point. (2) Covariates are allowed. (3) MCAR is assumed. (4) Observer specific and marginal agreement proportions are estimated in semi-parametric models.	To be used in order to calculate kappa with covariates for repeated measurements.
[3] Kappa U-statistic	It estimates kappa for longitudinal binary data with U-statistic, which is a nonparametric approach.	(1) Kappa is estimated at each time point. (2) The equality of kappa at different time points can be tested. (3) This approach allows MAR missing scheme and monotone missing patterns. (4) No covariates are allowed. (5) Nonparametric model assumed.	Preferred method for estimating kappa from repeated measurements for its robustness compared to method [2], if we are not interested in including any covariates.
[4] ICC	It estimates ICC through a linear mixed model which estimates the variance components ignoring the binary nature of the data.	(1) Aggregate agreement measure over all time points is generated. (2) Identity link is used for binary outcome. (3) All time points are equally weighted. (4) MAR missing scheme is allowed. (5) Covariates are allowed. (6) Full parametric model estimated via maximum likelihood paradigm.	It works well, when we do not have extreme proportions for binary outcomes, and all time points are considered to be equally important. Furthermore, it is easy to extend to the scenario with multiple raters.
[5] CCC U	It estimates CCC through U-statistics.	(1) Aggregate agreement measurements over all time points. (2) Different weights can be assigned to different time points according to importance. (3) No missing data are allowed. (4) No covariates are allowed. (5) Nonparametric model assumed.	It is preferred over ICC, because the inference is based on the robust nonparametric approach. But similar to ICC, only aggregate agreement estimate is given, and no information on each time point is available. Furthermore, the approach cannot handle any missing data or unbalanced design.
[6] CCC variance components	It estimates CCC through a linear mixed model which estimates the variance components ignoring the binary nature of the data.	(1) Aggregate agreement measurements over all time points. (2) Different weights can be assigned to different time points based on importance. (3) MAR missing scheme is allowed. (4) Covariates are allowed. (5) Full parametric model estimated via maximum likelihood paradigm.	It works well when we do not have extreme proportions for binary outcomes, and it is easy to extend the scenario to multiple raters.