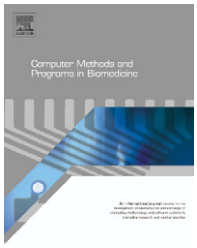
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 88 (2007) 62–74

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Computer programs for the concordance correlation coefficient

Sara B. Crawford[a,*], *Andrzej S. Kosinski*[b], *Hung-Mo Lin*[c], *John M. Williamson*[a], *Huiman X. Barnhart*[b]

[a] *Division of Parasitic Diseases, Centers for Disease Control and Prevention, 4770 Buford Highway NE (MS-F22), Atlanta, GA 30341, United States*
[b] *Department of Biostatistics and Bioinformatics and Duke Clinical Research Institute, Duke University, PO Box 17969, Durham, NC 27715, United States*
[c] *Department of Health Evaluation Sciences, Penn State College of Medicine, A210 600 Centerview Dr., Hershey, PA 17033, United States*

ARTICLE INFO

Article history:
Received 18 December 2006
Received in revised form 5 July 2007
Accepted 5 July 2007

Keywords:
Agreement
Bootstrap
Concordance correlation coefficient
Dependence
Reproducibility

ABSTRACT

The CCC macro is presented for computation of the concordance correlation coefficient (CCC), a common measure of reproducibility. The macro has been produced in both SAS and R, and a detailed presentation of the macro input and output for the SAS program is included. The macro provides estimation of three versions of the CCC, as presented by Lin [L.I.-K. Lin, A concordance correlation coefficient to evaluate reproducibility, Biometrics 45 (1989) 255–268], Barnhart et al. [H.X. Barnhart, J.L. Haber, J.L. Song, Overall concordance correlation coefficient for evaluating agreement among multiple observers, Biometrics 58 (2002) 1020–1027], and Williamson et al. [J.M. Williamson, S.B. Crawford, H.M. Lin, Resampling dependent concordance correlation coefficients, J. Biopharm. Stat. 17 (2007) 685–696]. It also provides bootstrap confidence intervals for the CCC, as well as for the difference in CCCs for both independent and dependent samples. The macro is designed for balanced data only. Detailed explanation of the involved computations and macro variable definitions are provided in the text. Two biomedical examples are included to illustrate that the macro can be easily implemented.

1. Introduction

In the health sciences, it is often necessary to study the reproducibility of continuous measurements made using a certain diagnostic tool or method. As technology brings forth new tools and methods, we are interested in evaluating the consistency of evaluations made using the new method as well as comparing this measure to the current gold standard if one exists. The concordance correlation coefficient (CCC) provides a means for examining the reproducibility of continuous measurements made by multiple raters using a single method or by two or more raters using two methods. Several other reproducibility measures are available, such as the Pearson correlation coefficient, the intraclass correlation coefficient [4,5], and the within-subject coefficient of variation [6]. In general, these measures do not address both precision and accuracy as does the concordance correlation coefficient; however, the equivalency and similarities of the intraclass correlation coefficient to the concordance correlation coefficient under certain scenarios has been discussed by Nickerson [7],

* *Corresponding author*. Tel.: +1 770 488 4204.
E-mail address: sgv0@cdc.gov (S.B. Crawford).
0169-2607/$ – see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2007.07.003

Carrasco and Jover [8], and Barnhart et al. [9]. The CCC measures how far the fitted linear relationship of two variables X and Y deviates from the concordance line (accuracy) and how far each observation deviates from the fitted line (precision).

There are several forms of the concordance correlation coefficient (CCC). The CCC for two raters evaluating a single method was presented by Lin [1,10]. Barnhart et al. [2] considered an overall CCC for multiple raters evaluating a single method, which is equivalent to the functions presented by Lin [1,11] and King and Chinchilli [12]. Williamson et al. [3] examined the agreement between two methods for multiple raters. The estimation of these forms of the CCC can be conducted by estimating the means, variances, and covariances for the ratings. As an alternative to estimation by the method of moments, Carrasco and Jover [8] propose estimating the CCC using variance components from a mixed model. Several methods have been proposed for inference regarding the CCC. For two raters evaluating a single method, Lin [1] proposed an asymptotic approach for computing variance estimates. For the overall CCC for multiple raters evaluating a single method, King and Chinchilli [12] conducted inference using a U-statistics approach, while Barnhart et al. [2] explored both a GEE and bootstrap approach. Williamson et al. [3] explored permutation testing and the bootstrap for agreement between two methods for multiple raters.

Here we describe the CCC macro written in SAS [13] which is designed to estimate all three forms of the CCC. The macro also provides confidence intervals for these estimates as well as for the difference in two CCCs. Where applicable, an asymptotic confidence interval is computed for both the estimation of the CCC as well as for the estimation of the difference in CCCs [1]. Otherwise, bootstrap confidence intervals are computed with the CCC macro [2,3,14]. The CCC macro was also written in R with the same input and output [15]. General examples for the macro call in R are included in Section 3.3 detailing the required parameters and output for each analysis, but all of the practical examples are presented in SAS.

2. Methods

2.1. Estimation of concordance correlation coefficients

A major component of the CCC macro is to provide an estimate of the concordance correlation coefficient (CCC). The formula used to compute the CCC is dependent upon the number of raters and the number of methods specified by the user. When a single method is evaluated by two raters, the CCC proposed by Lin [1] is used:

$$\rho^c = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}, \tag{1}$$

where μ_1 and σ_1^2 represent the mean and variance for the first rater, μ_2 and σ_2^2 represent the mean and variance for the second rater, and σ_{12} is the covariance for the first and second rater. An estimate for the CCC can be computed by substituting the corresponding sample estimates for the parameters. When a single method is evaluated by multiple raters, we apply the overall CCC presented by Barnhart et al. [2] for R raters:

$$\rho_o^c = \frac{2\sum_{r=1}^{R-1}\sum_{s=r+1}^{R}\sigma_{rs}}{(R-1)\sum_{r=1}^{R}\sigma_r^2 + \sum_{r=1}^{R-1}\sum_{s=r+1}^{R}(\mu_r - \mu_s)^2}. \tag{2}$$

Here, μ_r and σ_r^2 denote the mean and variance for the rth rater, and σ_{rs} is the covariance for raters r and s. The overall CCC in Eq. (2) is the same as that presented by Lin in the section on future studies [1,11] and the generalized CCC presented by King and Chinchilli [12] when the squared distance function is used. Williamson et al. [3] proposed another version of an overall CCC for measuring the agreement between two methods by the same R raters, such as an experimental method and the current gold standard:

$$\rho_{12}^c = \frac{2\sum_{r=1}^{R}\sigma_{12r}}{\sum_{r=1}^{R}(\sigma_{1r}^2 + \sigma_{2r}^2) + \sum_{r=1}^{R}(\mu_{1r} - \mu_{2r})^2}, \tag{3}$$

Here μ_{mr} and σ_{mr}^2 represent the mean and variance for the rth rater evaluating the mth method, and σ_{12r} represents the covariance of the first and second methods for the rth rater. Again in Eqs. (2) and (3) the corresponding sample estimates can be substituted for the parameters to estimate the CCC. These three equations for the concordance correlation coefficient are included in the CCC macro. The application of each equation depends on the number of methods and raters entered by the macro user.

2.2. Confidence intervals for the CCC

In the case of two raters evaluating a single assessment method, we can calculate an asymptotic confidence interval for the CCC. Lin showed that the estimate for the CCC, $\hat{\rho}^c$, has an asymptotic normal distribution with mean ρ^c and variance:

$$\sigma_{\hat{\rho}^c}^2 = \frac{1}{n-2}\left[\frac{(1-\rho^2)(\rho^c)^2(1-(\rho^c)^2)}{\rho^2} + \frac{2(\rho^c)^3(1-\rho^c)u^2}{\rho} - \frac{(\rho^c)^4u^4}{2\rho^2}\right]. \tag{4}$$

where $u = (\mu_1 - \mu_2)/\sqrt{\sigma_1\sigma_2}$ and ρ represents the Pearson correlation coefficient. Note that Eq. (4) is undefined when $\rho = 0$. Lin also suggests the use of a Z-transformation in order to achieve asymptotic normality of the estimator:

$$\hat{Z} = \tanh^{-1}(\hat{\rho}^c) = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}^c}{1-\hat{\rho}^c}\right). \tag{5}$$

The Z-transformation is asymptotically normally distributed with mean

$$Z = \tanh^{-1}(\hat{\rho}^c) = \frac{1}{2}\ln\left(\frac{1+\rho^c}{1-\rho^c}\right) \tag{6}$$

and variance

$$\sigma_{\hat{Z}}^2 = \frac{1}{n-2}\left[\frac{(1-\rho^2)(\rho^c)^2}{(1-(\rho^c)^2)\rho^2} + \frac{2(\rho^c)^3(1-\rho^c)u^2}{\rho(1-(\rho^c)^2)^2} - \frac{(\rho^c)^4u^4}{2\rho^2(1-(\rho^c)^2)^2}\right]. \tag{7}$$

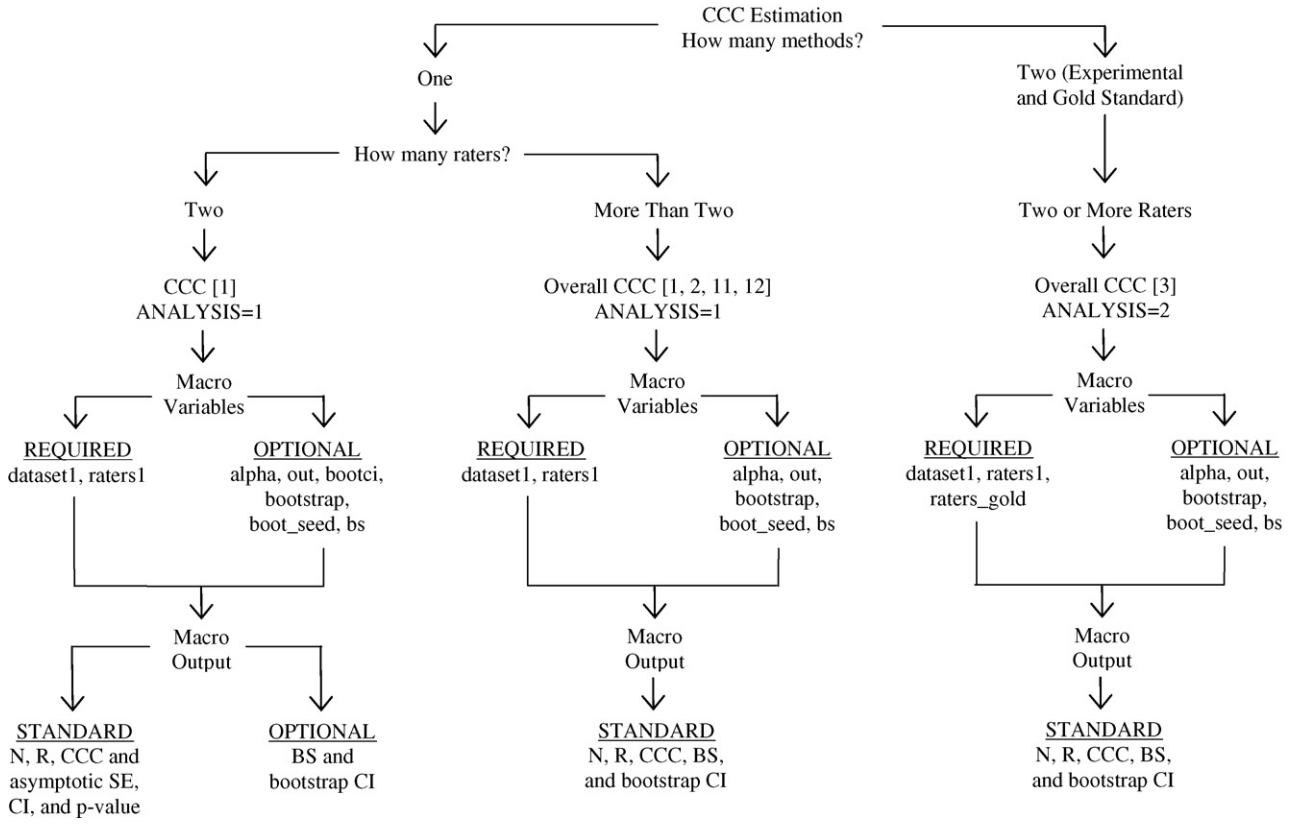


Fig. 1 – Macro options, variables, and output for CCC estimation.

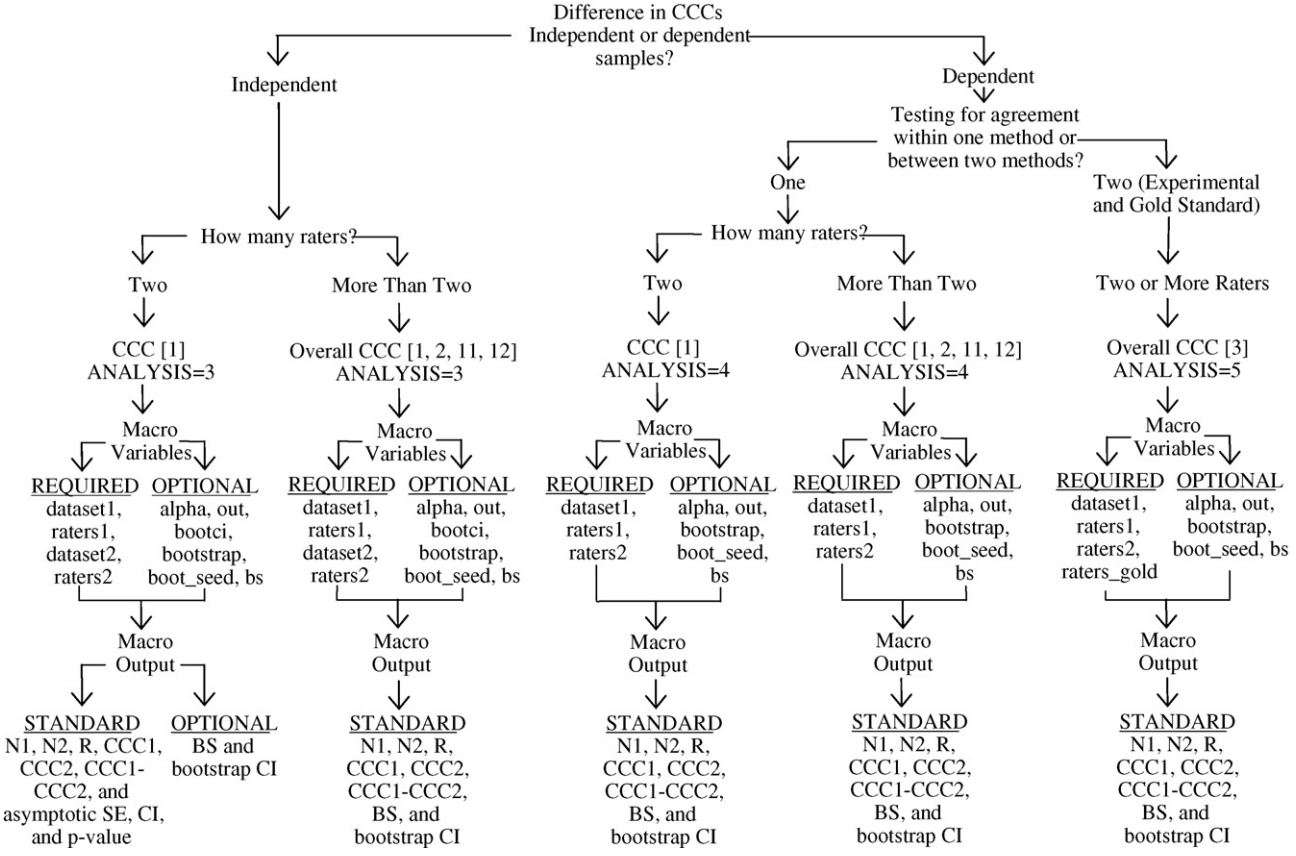


Fig. 2 – Macro options, variables, and output for the estimation and testing of the difference in CCCs.

This Z-transformation can be used to calculate an asymptotic, non-symmetric confidence interval for the CCC. The macro provides the option of performing a hypothesis test for the equality of two independent CCCs. When only two raters are evaluating a single method for each group of independent subjects, the variance for individual CCC presented in Eq. (4) can be used to construct an asymptotic confidence interval for the difference in CCCs.

The bootstrap is a resampling approach used frequently in applied statistics [14]. It involves creating a large number of resampled datasets sampled with replacement from the original sample. The test statistic is calculated for each resampled dataset in order to create a probability distribution for the test statistic of interest. When subjects are evaluated by multiple raters and/or two methods, the macro generates bootstrap confidence intervals for the CCC. Both the percentile and the bias accelerated and corrected (BCa) bootstrap confidence intervals can be computed. For the calculation of the $100 \times (1-\alpha)\%$ bootstrap confidence intervals, a large number of bootstrap samples are selected from the original sample with replacement. The estimate of the CCC, $\hat{\rho}^{c*}$, is calculated for each resampled dataset. The resulting percentile bootstrap confidence interval is $\left(\hat{\rho}^{c*(\alpha/2)}, \hat{\rho}^{c*((1-\alpha)/2)}\right)$, where $\hat{\rho}^{c*(\alpha/2)}$ is the $(\alpha/2)$th empirical percentile of the distribution of resampled CCCs. The weighted average definition of percentile is used. The BCa confidence interval is also a percentile interval, but the percentiles are not based solely on $(\alpha/2)$ and $((1-\alpha)/2)$. The percentiles are adjusted by the median bias of the bootstrap samples as well as the acceleration of the standard error [14,16].

2.3. *Estimation of the difference in CCCs*

The CCC macro also can estimate the difference in CCCs and calculate a bootstrap confidence interval for the difference. This confidence interval can be used to test for equal concordance correlation coefficients by determining whether the null value lies in the confidence interval. The difference in CCCs can be assessed under three scenarios: a set of raters evaluating a method in each of two independent groups of subjects; a set of raters evaluating two methods over the same group of subjects; and a set of raters evaluating three methods (such as for two experimental methods and a gold standard) over the same group of subjects. Note that the third scenario involves the evaluation of three methods, but does not require the third method to be a gold standard. In the writing of the macro and all of the macro description, the reference method is referred to as the gold standard in order to clar-

Table 1 – CCC estimation for one method with two or more raters

Summary	Performs estimation of the CCC for one method with two or more raters (Lin [1, 10, 11], Barnhart et al. [2], King and Chinchilli [12]) as shown in eq. (1) for two raters and in eq. (2) for more than two raters.
Example Input Dataset	DATASET: INPUTDATA ID observer1 observer2 observer3 1 x_{11} x_{12} x_{13} 2 x_{21} x_{22} x_{23} 3 x_{31} x_{32} x_{33}
Example SAS Code	%include 'cccmacro.sas'; %ccc(analysis=1, dataset1=*inputdata*, raters1=*observer1 observer2 observer3*, alpha=0.05, out=*outputdata*, bootci='Y', bootstrap='B', boot_seed=123, bs=2000);
Example R Code	out <- f.analysis(analysis=1, dataset1=*inputdata*, raters1=c("*observer1*","*observer2*","*observer3*"), alpha=0.05, bootci="Y", bootstrap="B", boot.seed=123, bs=2000) out
Hypothetical Output	N R CCC BS BOOTSTRAP_LCL BOOTSTRAP_UCL 100 3 0.50 2000 0.25 0.75

ify which method is serving as the referent method. Under the first scenario, we are interested in testing the hypothesis $H_0 : \rho_1^c = \rho_2^c$ versus $H_1 : \rho_1^c \neq \rho_2^c$, under the second scenario, $H_0 : \rho_{o1}^c = \rho_{o2}^c$ versus $H_1 : \rho_{o1}^c \neq \rho_{o2}^c$, and under the third scenario, $H_0 : \rho_{1,\text{Gold}}^c = \rho_{2,\text{Gold}}^c$ versus $H_1 : \rho_{1,\text{Gold}}^c \neq \rho_{2,\text{Gold}}^c$. The test statistics are $\hat{\rho}_{\text{diff}}^c = \hat{\rho}_1^c - \hat{\rho}_2^c$, $\hat{\rho}_{\text{o,diff}}^c = \hat{\rho}_{o1}^c - \hat{\rho}_{o2}^c$, and $\hat{\rho}_{\text{Gold,diff}}^c = \hat{\rho}_{1,\text{Gold}}^c - \hat{\rho}_{2,\text{Gold}}^c$, respectively.

In estimating and testing the difference of independent CCCs where each group of subjects is evaluated by two raters (scenario 1), an asymptotic hypothesis test can be conducted based on Lin's CCC, shown in Eq. (1), and asymptotic standard error, shown in Eq. (4). A bootstrap confidence interval also can be used for testing the hypothesis $H_0 : \rho_{c1} = \rho_{c2}$. With this approach, a 95% bootstrap confidence interval is calculated for the difference in CCCs and the null hypothesis is rejected at an alpha level of 0.05 if the confidence interval does not contain zero. The construction of a bootstrap confidence interval requires resampling the original sample with replacement and then calculating the test statistic of interest a large number of times (e.g. 2000) in order to create a probability distribution for the test statistic. Because we are making an assumption of the equality of the CCCs but not necessarily the equality of the underlying distributions under the null hypothesis, each bootstrap resample for independent groups will be conducted within each of the two groups and then the difference in the CCCs will be calculated. In the case of dependent CCCs, either for two competing methods (scenario 2) or two experimental methods and a gold standard (scenario 3), we have one set of subjects over which multiple raters can evaluate multiple methods. Because a single subject is being evaluated in the case of dependent CCCs, the bootstrap resampling will be done at the subject level [3,14,16].

3. Computer program

3.1. Macro overview

The CCC macro is designed to perform a variety of analyses pertaining to the concordance correlation coefficient. The macro can provide estimates of the overall CCC along with confidence intervals, either asymptotic or bootstrap, for one method with multiple raters, or for two methods with multiple raters (i.e. an experimental method and a gold standard). It will also perform estimation of the difference in two CCCs under the assumptions of both independence and dependence, and calculate confidence intervals using either asymptotic theory or bootstrap methodology. See Fig. 1 for a flowchart describing the macro capabilities, the required and optional macro variables for input, and the standard macro output for estimation of a single CCC and see Fig. 2 for a flowchart describing

Table 2 – CCC estimation for agreement between two methods with two or more raters

Summary	Performs estimation of the CCC for assessing the agreement between two methods with two or more raters (Williamson [3]) as shown in Eq. (3).
Example Input Dataset	DATASET: INPUTDATA ID method1rater1 method1rater2 goldrater1 goldrater2 1 x_{111} x_{112} x_{1G1} x_{1G2} 2 x_{211} x_{212} x_{2G1} x_{2G2} 3 x_{311} x_{312} x_{3G1} x_{3G2}
Example SAS Code	%include 'cccmacro.sas'; %ccc(analysis=2, dataset1=*inputdata*, raters1=*method1rater1 method1rater2*, raters_gold=*goldrater1 goldrater2*, alpha=0.05, out=*outputdata*, bootstrap='B', boot_seed=123, bs=2000);
Example R Code	out <- f.analysis(analysis=2, dataset1=*inputdata*, raters1=c("*method1rater1*", "*method1rater2*"), raters.gold=c("*goldrater1*", "*goldrater2*"), alpha=0.05, bootstrap="B", boot.seed=123, bs=2000) out
Hypothetical Output	N R CCC BS BOOTSTRAP_LCL BOOTSTRAP_UCL 100 2 0.50 2000 0.25 0.75

the same macro characteristics for estimation and testing of the difference in CCCs. The macro is written for balanced data, and will delete any incomplete lines of data from the analysis.

3.2. *Input parameters*

If the program is used to compute an estimate of the overall CCC or to perform a hypothesis test for dependent data, one dataset is required for input, with one line of data per subject. The value assessed by each rater for each method must be recorded in a separate variable. If the program is used to perform a hypothesis test for independent data, two datasets are required. The first will contain one line of data for each subject assessed in the first group with one variable for each rater, while the second will contain one line of data for each subject assessed in the second group. The following is a list of all macro variables available for input.

ANALYSIS: Possible values are 1, 2, 3, 4, or 5. An entry of 1 or 2 requests estimation of a single CCC. An entry of 3, 4, or 5 will request testing for the difference in CCCs, and will provide both estimation of the difference in CCCs as well as an asymptotic or bootstrap confidence interval for the difference. Specifically
1: indicates estimation of the overall CCC for one method with two or more raters (Table 1)
2: indicates estimation of the CCC for assessing the agreement between two methods with two or more raters (Table 2)
3: requests hypothesis testing of the difference in independent CCCs for a set of raters evaluating a single method in each of two independent groups of subjects (Table 3)
4: requests hypothesis testing of the difference in dependent CCCs for a set of raters evaluating two methods over the same group of subjects (Table 4)
5: requests hypothesis testing of the difference in two dependent CCCs, where each CCC is measuring the agreement between an experimental method and a gold standard over the same group of subjects (Table 5)

DATASET1: Main input dataset for all estimation and dependent hypothesis tests

DATASET2: Second dataset for the independent hypothesis test

RATERS1: String of names for the variables representing the raters evaluating the first or only method, separated by a space. For example, RATERS1 = rater1 rater2 rater3

RATERS2: String of names for the variables representing the raters for the second method when applicable, separated by a space

RATERS_GOLD: String of names for the variables representing the raters evaluating the gold standard when applicable, separated by a space

ALPHA: Type-I error rate. Generates $(1-\alpha) \times 100\%$ confidence intervals (default = 0.05)

OUT: Name of output dataset where the analysis results are stored (default = work.outdata)

BOOTCI: For analyses where the bootstrap confidence interval is optional, indicates whether or not a bootstrap confidence interval will be produced. Possible values are 'Y' for yes and 'N' for no (default = 'N')

BOOTSTRAP: Indicates the type of bootstrap confidence interval to be generated. Possible values are 'P' for percentile and 'B' for BCa (default = 'B')

BS: Number of bootstrap samples (default = 2000)

BOOT_SEED: Seed for generating the bootstrap samples (default = clock)

3.3. *Required parameters and output for each analysis*

When invoking the CCC macro, the macro variable ANALYSIS must always be specified. The other variables that are required as well as the output that will be produced are dependent on the value of this macro variable ANALYSIS. The required and optional macro variables for each type of analysis, as well as the output that will be produced, is summarized in Figs. 1 and 2. A summary of the macro functioning, an example of the required form of the dataset, the macro invocation for each type of analysis specifying all required and optional macro variables, and an example of the macro output is presented in Tables 1–5. Table 1 provides an example for the overall estimation of the CCC for one method with multiple raters; Table 2, the estimation of the CCC for agreement between two methods with multiple raters; Table 3, the estimation of the difference between two independent CCCs for one method with multiple raters, which is an extension of analysis = 1; Table 4, the estimation of the difference between two dependent CCCs for one method with multiple raters, which is also an extension of analysis = 1; Table 5, the estimation of the difference between two dependent CCCs for two methods with multiple raters, which is an extension of analysis = 2. The words in italics represent generic dataset and variable names that would be altered according to the users data.

4. Examples

4.1. *Biochemical in vitro assays*

In a study of biochemical in vitro assays, researchers were interested in the reproducibility of toxicity measurements made by two different assays: cellular adenosine triphosphate activity using cell line 76 (ATP-76) and cellular adhesion using cell line 74 (CLA-74) [1]. The percent cell function measured

Table 3 – Testing for the difference between two independent CCCs with two or more raters

Summary	Performs hypothesis testing of the difference in independent CCCs for a set of raters evaluating a single method in each of two independent groups of subjects, where the CCC for each independent group is calculated using eq. (1) for two raters and eq. (2) for more than two raters.
Example Input Dataset	DATASET1: INPUTDATA1 DATASET2: INPUTDATA2 ID observer1 observer2 ID observer1 observer2 1 x_{11} x_{12} 101 $x_{101,1}$ $x_{101,2}$ 2 x_{21} x_{22} 102 $x_{102,1}$ $x_{102,2}$ 3 x_{31} x_{32} 103 $x_{103,1}$ $x_{103,2}$
Example SAS Code	%include 'cccmacro.sas'; %ccc(analysis=3, dataset1=*inputdata1*, raters1=*observer1 observer2*, dataset2=*inputdata2*, raters2=*observer1 observer2*, alpha=0.05, out=*outputdata*, bootci='Y', bootstrap='B', boot_seed=123, bs=2000);
Example R Code	out <- f.analysis(analysis=3, dataset1=*inputdata1*, raters1=c("*observer1*", "*observer2*"), dataset2=*inputdata2*, raters2 =c("*observer1*", "*observer2*"), alpha=0.05, bootci="Y", bootstrap="B", boot.seed=123, bs=2000) out
Hypothetical Output	N_1 N_2 R CCC_1 CCC_2 CCC_DIFF SE_DIFF 100 100 2 0.75 0.25 0.50 0.10 LCL UCL PVALUE BS BOOT_LCL BOOT_UCL 0.30 0.70 0.005 2000 0.30 0.70

by each assay at two independent trials conducted 1 week apart was recorded for 10 materials of varying toxicity. We are interested in assessing the agreement of the measurements produced over the two independent trials for each assay, as well as whether the agreement between the two assays differs. The SAS code and output for this example are located in Appendix A.

For the estimation of the assay-specific CCC measured over two trials, we can use Lin's CCC [1] presented in Eq. (1). For assay ATP-76, the estimate for the CCC is 0.97 with a 95% asymptotic confidence interval of (0.89, 0.99). As an option, a BCa bootstrap confidence interval of (0.92, 0.99) is also produced. For assay CLA-74, the estimate for the CCC is much lower at 0.28. The asymptotic 95% confidence interval is (−0.23, 0.67) and a bootstrap confidence interval is (−0.20, 0.88). Calculating a bootstrap confidence interval for the difference in dependent CCCs, we find that the difference between ATP-76 and CLA-74 is 0.69 (0.07, 1.15), or that the reproducibility of assay ATP-76 is significantly greater than the reproducibility of assay CLA-74.

4.2. *Carotid stenosis*

A carotid stenosis screening study was conducted at Emory University from 1994 to 1996 [2,3,17]. Three observers, each using three diagnostic methods, assessed the stenosis of the left and right carotid arteries of 55 patients. The three methods were magnetic resonance angiography two-dimensional time of flight (MRA 2D), magnetic resonance angiography three-dimensional time of flight (MRA 3D), and intra-arterial angiogram (IA), where IA is considered the current gold standard. We are interested in assessing the agreement of the three observers within each method, and then comparing this agreement between each of the methods. We are also interested in whether the agreement between each of the experimental methods, MRA 2D and MRA 3D, and the gold standard are different. We can explore each of these questions separately for the left and right arteries. The SAS code and output for the example are located in Appendix B, but is restricted to the left side.

We estimated the overall concordance correlation coefficient for each method with three raters, as shown in Eq. (2)

Table 4 – Testing for the difference between two dependent CCCs with two or more raters

Summary	Performs hypothesis testing of the difference in dependent CCCs for a set of raters evaluating two methods over the same group of subjects, where the CCC for each dependent group is calculated using eq. (1) for two raters and eq. (2) for more than two raters.
Example Input Dataset	DATASET: INPUTDATA ID meth1rater1 meth1rater2 meth2rater1 meth2rater2 1 x_{111} x_{112} x_{121} x_{122} 2 x_{211} x_{212} x_{221} x_{222} 3 x_{311} x_{312} x_{321} x_{322}
Example SAS Code	%include 'cccmacro.sas'; %ccc(analysis=4, dataset1=*inputdata1*, raters1=*meth1rater1 meth1rater2*, raters2=*meth2rater1 meth2rater2*, alpha=0.05, out=*outputdata*, bootstrap='B', boot_seed=789, bs=2000);
Example R Code	out <- f.analysis(analysis=4, dataset1=*inputdata1*, raters1=c("*meth1rater1*","*meth1rater2*"), raters2=c("*meth2rater1*","*meth2rater2*"), alpha=0.05, bootstrap="B", boot.seed=789, bs=2000) out
Hypothetical Output	N R CCC_1 CCC_2 CCC_DIFF BS BOOT_LCL BOOT_UCL 100 2 0.25 0.75 -0.50 2000 -0.70 -0.30

[1,2,11,12], as well as a 95% BCa bootstrap confidence interval based upon 2000 bootstrap replicates. For the left artery, the CCCs for MRA 2D, MRA 3D and the gold standard were 0.62 (0.44, 0.77), 0.64 (0.45, 0.79), and 0.88 (0.69, 0.94), respectively. Similarly, the CCCs for the right artery were 0.61 (0.43, 0.75), 0.62 (0.45, 0.76), and 0.92 (0.85, 0.95), respectively. These estimates are all significantly greater than zero. The estimates of the CCC for MRA 2D and MRA 3D appear to be very similar while the estimates for IA appear to be larger, indicating that the agreement among raters for the gold standard may be greater than the agreement among raters for the two experimental methods. We can test the difference in these overall CCCs using a bootstrap confidence interval [3]. The differences in estimated CCCs, as well as the bootstrap confidence intervals based upon 2000 bootstrap samples, showed that the CCCs for MRA 2D and IA and the CCCs for MRA 3D and IA were significantly different for the left [MRA 2D − IA = −0.26 (−0.44, −0.12); MRA 3D − IA = −0.24 (−0.44, −0.09)] and right [MRA 2D − IA = −0.31 (−0.48, −0.17); MRA 3D − IA = −0.30 (−0.48, −0.15)] arteries. These results indicate that the agreement among raters for the experimental methods is significantly worse than the agreement for the gold standard.

We can further assess the agreement of the three raters between two methods, the experimental method and IA, for both MRA 2D and MRA 3D by applying the formula for the overall CCC for two methods found in Eq. (3) [3] and calculating a 95% BCa bootstrap confidence interval. In the left artery, the CCC for MRA 2D and IA is 0.56 (0.37, 0.72) while the CCC for MRA 3D and IA is 0.48 (0.27, 0.65). In the right artery, the CCCs are 0.63 (0.46, 0.75) and 0.56 (0.38, 0.75), respectively. We can also test the difference between these overall CCCs for the left and right arteries in order to see whether the agreement among raters over MRA 2D and IA is greater than that for MRA 3D and IA. Using the bootstrap confidence interval for dependent CCCs, we find that the difference for the left artery is 0.08 (−0.02, 0.22) and the difference for the right artery is 0.07 (−0.08, 0.23). These results indicate that the differences in overall agreement are not statistically significant.

5. Macro availability and run time

The CCC macro written in SAS [13], R [15], or S-PLUS [18] can be obtained by directly contacting the authors or by accessing the following websites: http://www.statisticaldisplays.org or http://www.personal.psu.edu/hxl28/research/CCCprogram. The SAS macro was written in v9, but because some of the IML functions available in SAS v9 are not available in SAS v8, a v8

Table 5 – Testing for the difference between two dependent CCCs for two methods with two or raters

Summary	Performs hypothesis testing of the difference in two dependent CCCs, where each CCC is measuring the agreement between an experimental method and a gold standard. Each dependent CCC is calculated using eq. (3).
Example Input Dataset	DATASET: INPUTDATA ID M1rater1 M1rater2 M2rater1 M2rater2 G_rater1 G_rater2 1 x_{111} x_{112} x_{121} x_{122} x_{1G1} x_{1G2} 2 x_{211} x_{212} x_{221} x_{222} x_{2G1} x_{2G2} 3 x_{311} x_{312} x_{321} x_{322} x_{3G1} x_{3G2}
Example SAS Code	%include 'cccmacro.sas'; %ccc(analysis=5, dataset1=*inputdata1*, raters1=*M1rater1 M1rater2*, raters2=*M2rater1 M2rater2*, raters_gold=*G_rater1 G_rater2*, alpha=0.05, out=*outputdata*, bootstrap='B', boot_seed=789, bs=2000);
Example R Code	out <- f.analysis(analysis=5, dataset1= *inputdata1*, raters1=c("*M1rater1*","*M1rater2*"), raters2=c("*M2rater1*","*M2rater2*"), raters.gold=c("*G_rater1*","*G_rater2*"), alpha=0.05, bootstrap="B", boot.seed=789, bs=2000) out
Hypothetical Output	N R CCC_1 CCC_2 CCC_DIFF BS BOOT_LCL BOOT_UCL 100 2 0.50 0.50 0.00 2000 -0.15 0.15

macro also was created. This macro performs the same functions, but requires more computing time, and can be obtained following the above instructions. The program for R and S-PLUS was designed to be a self-contained function, does not require bootstrap packages, and should run in all versions of the R or S-PLUS software. It was tested for R version 1.9.0 and S-PLUS version 6.1.2 and runs faster in R than in S-PLUS if bootstrap computations are requested. For the examples in SAS, the macro was run in SAS v9 in the PC environment on a desktop computer with an Intel® Pentium® 4 processor of 1.80 GHz speed and 512 MB of RAM. For the examples in R, the function was also run in the PC environment on a desktop computer with an Intel® Pentium 4 processor with 2.60 GHz speed and 512 MB of RAM.

The amount of CPU time required to run the macro will vary considerably based upon the number of subjects, the number of raters, and the number of bootstrap samples specified (where applicable). For the carotid stenosis example, when estimating the overall CCC and computing a bootstrap confidence interval based on 2000 bootstrap samples, the amount of required CPU time for the SAS macro was approximately 20–21 s. The amount of CPU time required for the R function for the same example was approximately 1.5 s. The estimation of the difference in dependent CCCs and the computation of a bootstrap confidence interval based on 2000 samples took 40–41 s of CPU time in SAS and 3 s in R. When exploring the CCC for two methods in the carotid stenosis example (such as for an experimental method and a gold standard), the estimation of the CCC and a bootstrap confidence interval took 17–19 s in SAS and 3.5 s in R, and the estimation of the difference in dependent CCCs and a bootstrap confidence interval took 34–35 s in SAS and 7 s in R.

Acknowledgements

Sara Crawford's research is supported by an appointment to the Research Participation Program at the Centers for Disease Control and Prevention, National Center for Infectious Diseases, Division of Parasitic Diseases administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and CDC. The research of Andrzej Kosinski and Huiman Barnhart is supported by the National Institutes of Health Grant R01 MH70028. We thank two anonymous referees for helpful comments in regards to the preparation of this manuscript.

Appendix A. Biochemical in vitro assay example

The following output displays a snapshot of the dataset *assays.sas7bdat* for the first three subjects only. The variable *id* represents the subject number. *Trial1A* and *trial2A* represent the results for each trial for assay A while *trial1C* and *trial2C* represent the results for each trial for assay C. The incomplete SAS code and output follow. Note that a unique example for each type of call is given, while similar replications were excluded.

id	trial1A	trial2A	trial1C	trial2C
1	1.2	-4.5	54.1	51.9
2	6.5	4.2	102.6	53.7
3	45.9	30.3	100.6	68.3

```
%inc 'C:/CCC macro v9.sas';

*ESTIMATING THE OVERALL CCC FOR EACH ASSAY ALONG WITH A BCa BOOTSTRAP
CONFIDENCE INTERVAL;
%ccc(analysis=1, dataset1=assays, raters1=trial1A trial2A, alpha=0.05,
out=outdata, bootCI='Y', bootstrap='B', boot_seed=1001, bs=2000);
```

One Method with Two Raters

Concordance Correlation Coefficient and SE (Lin, 1989)

95.0% Asymptotic Confidence Interval, Z-transformation

95.0% BCa Bootstrap Confidence Interval

N	R	CCC	SE	LCL	UCL
10	2	0.9686	0.021	0.8862	0.9916

BS	BOOTSTRAP_LCL	BOOTSTRAP_UCL
2000	0.9174	0.991

```
*ESTIMATING THE DIFFERENCE IN DEPENDENT CCCS ALONG WITH A BCa BOOTSTRAP
CONFIDENCE INTERVAL;
%ccc(analysis=4, dataset1=assays, raters1=trial1A trial2A, raters2=trial1C
trial2C, alpha=0.05, out=outdata, bootstrap='B', boot_seed=3001, bs=2000);
```

```
 Two Dependent Methods, Each with Two Raters (Lin, 1989)
Difference in Overall Concordance Correlation Coefficients
         95.0% BCa Bootstrap Confidence Interval

    N        R    CCC_1    CCC_2   CCC_DIFF
   10        2   0.9686   0.2829     0.6857

     BS   BOOTSTRAP_LCL   BOOTSTRAP_UCL
   2000          0.0715          1.1462
```

Appendix B. Carotid stenosis example

The following output displays a snapshot of the dataset *ica_left.sas7bdat* for the first three subjects. The variable *id* represents the subject number. The variables *m1r1*, *m1r2*, and *m1r3* represent the values for the three raters evaluating method MRA 2D, the variables *m2r1*, *m2r2*, and *m2r3* represent the values for the three raters evaluating method MRA 3D, and the variables *m3r1*, *m3r2*, and *m3r3* represent the values for the three raters evaluating the gold standard IA. The incomplete SAS code and output follow. Note that a unique example for each type of call is given, while similar replications were excluded. The dataset, SAS code, and output are restricted to the left side only.

id	m1r1	m1r2	m1r3	m2r1	m2r2	m2r3	m3r1	m3r2	m3r3
1	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000
2	60.410	72.214	69.472	62.340	61.898	58.955	66.575	77.186	73.077
3	0.000	0.000	49.324	12.348	0.000	0.000	0.000	3.191	0.000

```
%inc 'C:/CCC macro v9.sas';

*OVERALL CCC FOR EACH METHOD AND BCa BOOTSTRAP CONFIDENCE INTERVAL;
%ccc(analysis=1, dataset1=ica_left, raters1=m1r1 m1r2 m1r3, alpha=0.05,
out=outdata, bootCI='Y', bootstrap='B', boot_seed=400, bs=2000);

                    One Method with Multiple Raters
          Overall Concordance Correlation Coefficient (Barnhart, 2002)
                  95.0% BCa Bootstrap Confidence Interval

    N        R      CCC       BS   BOOTSTRAP_LCL   BOOTSTRAP_UCL
```

55	3	0.6226	2000	0.4377	0.766

*ESTIMATING THE DIFFERENCE IN DEPENDENT CCCS ALONG WITH A BCa BOOTSTRAP CONFIDENCE INTERVAL, PAIRWISE;

%*ccc*(analysis=4, dataset1=ica_left, raters1=m1r1 m1r2 m1r3, raters2=m3r1 m3r2 m3r3, alpha=0.05, out=outdata, bootstrap='B', boot_seed=500, bs=2000);

Two Dependent Methods, Each with Multiple Raters (Barnhart, 2002)

Difference in Overall Concordance Correlation Coefficients

95.0% BCa Bootstrap Confidence Interval

N	R	CCC_1	CCC_2	CCC_DIFF
55	3	0.6226	0.8818	-0.259

BS	BOOTSTRAP_LCL	BOOTSTRAP_UCL
2000	-0.443	-0.123

*ESTIMATION OF THE OVERALL CCC FOR EACH METHOD VERSUS THE GOLD STANDARD AND BCa CONFIDENCE INTERVAL;

%*ccc*(analysis=2, dataset1=ica_left, raters1=m1r1 m1r2 m1r3, raters_gold=m3r1 m3r2 m3r3, alpha=0.05, out=outdata, bootstrap='B', boot_seed=600, bs=2000);

Experimental Method and Gold Standard with Two or More Raters

Overall Concordance Correlation Coefficient (Williamson, accepted)

95.0% BCa Bootstrap Confidence Interval

N	R	CCC	BS	BOOTSTRAP_LCL	BOOTSTRAP_UCL
55	3	0.5602	2000	0.3698	0.7203

* ESTIMATION OF THE DIFFERENCE IN DEPENDENT CCCS FOR TWO METHODS (GOLD) AND A BCa CONFIDENCE INTERVAL;

```
%ccc(analysis=5, dataset1=ica_left, raters1=m1r1 m1r2 m1r3, raters2=m2r1 m2r2
m2r3, raters_gold=m3r1 m3r2 m3r3, alpha=0.05, out=outdata, bootstrap='B',
boot_seed=700, bs=2000);
```

Agreement Between Two Dependent Experimental Methods and a Gold Standard

Each with Two or More Raters (Williamson, accepted)

Difference in Overall Concordance Correlation Coefficients

95.0% BCa Bootstrap Confidence Interval

N	R	CCC_1	CCC_2	CCC_DIFF
55	3	0.5602	0.4765	0.0837

BS	BOOTSTRAP_LCL	BOOTSTRAP_UCL
2000	-0.022	0.2231

REFERENCES

[1] L.I-K. Lin, A concordance correlation coefficient to evaluate reproducibility, Biometrics 45 (1989) 255–268.

[2] H.X. Barnhart, M. Haber, J.L. Song, Overall concordance correlation coefficient for evaluating agreement among multiple observers, Biometrics 58 (2002) 1020–1027.

[3] J.M. Williamson, S.B. Crawford, H.M. Lin, Resampling dependent concordance correlation coefficients, J. Biopharm. Stat. 17 (2007) 685–696.

[4] J.L. Fleiss, The Design Analysis of Clinical Experiments, John Wiley & Sons, New York, 1986.

[5] H. Quan, W.J. Shih, Assessing reproducibility by the within-subject coefficient of variation with random effects models, Biometrics 52 (1996) 1195–1203.

[6] J. Lee, D. Koh, C.N. Ong, Statistical evaluation of agreement between two methods for measuring a quantitative variable, Comput. Biol. Med. 19 (1989) 61–70.

[7] C.A.E. Nickerson, A note on "a concordance correlation coefficient to evaluate reproducibility", Biometrics 53 (1997) 1503–1507.

[8] J.L. Carrasco, L. Jover, Estimating the generalized concordance correlation coefficient through variance components, Biometrics 59 (2003) 849–858.

[9] H.X. Barnhart, M.J. Haber, L.I. Lin, An overview on assessing agreement with continuous measurement, J. Biopharm. Stat. 17 (2007) 529–569.

[10] L.I-K. Lin, Assay validation using the concordance correlation coefficient, Biometrics 48 (1992) 599–604.

[11] L.I-K. Lin, A note on the concordance correlation coefficient, Biometrics 56 (2000) 324–325.

[12] T.S. King, V.M. Chinchilli, A generalized concordance correlation coefficient for continuous and categorical data, Stat. Med. 20 (2001) 2131–2147.

[13] Statistical Analysis Software, SAS Institute Inc., Cary, NC, 1995.

[14] B. Efron, R. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.

[15] R language and environment for statistical computing and graphics, 2004 (http://www.r-project.org/).

[16] A.C. Davison, D.V. Hinkley, Bootstrap Methods and their Applications, Cambridge University Press, Cambridge, 1997.

[17] H.X. Barnhart, J.M. Williamson, Modeling concordance correlation via GEE to evaluate reproducibility, Biometrics 57 (2001) 931–940.

[18] S-PLUS Statistical Package, Insightful Corp., Seattle, WA, 1998.