Crop Protection 101 (2017) 68–75

Contents lists available at ScienceDirect

Crop Protection

journal homepage: www.elsevier.com/locate/cropro

Bland-Altman comparison of two methods for assessing severity of Verticillium wilt of potato

S.K.R. Yellareddygari, N.C. Gudmestad*

Department of Plant Pathology, North Dakota State University, Fargo, 58105, USA

ARTICLE INFO

Article history:
Received 24 April 2017
Received in revised form
19 July 2017
Accepted 23 July 2017
Available online 29 July 2017

Keywords:
Canopeo
Correlation
Mixed effects model

ABSTRACT

The agreement between two disease assessment approaches is important to know prior to replacing or interchanging the use of an established method with a recently developed method of measurement. Frequently used statistical methods to compare two different disease rating methods is the Pearson correlation coefficient or the ordinary least square regression (OLS), but they have their shortcomings. Bland-Altman proposed an alternative method for studying agreement between methods using simple graphs and basic statistics. Traditionally, when disease management strategies are being evaluated in the field, the severity of the disease is estimated using a visual assessment. Canopeo, designed by the Oklahoma State University app center, is a smart phone app designed for measuring green canopy cover. Thus, the aim of this study was to explain the Bland-Altman method with examples of visual and Canopeo methods of wilt measurement. Symptoms of Verticillium wilt in potato were estimated (repeated measures) in two trials using Canopeo and a traditional visual assessment method. Complete wilt data (repeated measures) were considered for studying the agreement between visual and Canopeo assessments. A preset cutoff limit of ≤5% bias (total allowable) between rating methods was considered acceptable prior to using the Bland-Altman comparison. The Bland-Altman method for determining the agreement in wilt severity methods in trial 1 and trial 2 estimated that the mean difference between rating methods were 5.10 and 5.91%, respectively. A mean difference greater than five indicates that the methods of measuring wilt are not in agreement. The study reported here demonstrates that Pearson correlation and OLS regression are inappropriate for assessing the agreement between two methods of measurement.

1. Introduction

Plant pathology research often encounters two methods of measurement that assess the same quantity. For example, studies where pustules are counted with the naked eye and magnification hand lens or estimating cell concentration using hemocytometer and spectrophotometer methods. One method is traditionally used before the introduction of another method for replacement or interchangeable use. In this scenario, the new method is compared with an established method rather than the measured variables of each subject (Bland and Altman, 1999). Some level of disagreement is allowed between methods because each method is subjected to random measurement error, knowing the amount of agreement between methods is important prior to a researcher replacing an old with a new method (Bland and Altman, 1999). Agreement is quantified by appraising the differences and cause of these differences between two quantitative methods of measurement (Giavarina, 2015). Most commonly a Pearson correlation coefficient (r) is used to compare two analytical methods (Altman and Bland, 1983; Ludbrook, 2002) and based on the magnitude of r the degree of association between two methods is determined. For example, correlation coefficient was used to compare the Assess (software) disease quantification method to that of visual method for counting maize rust lesions (Bade and Carmona, 2011). Also, correlation was used to compare common bean leaf area index (LAI) measurements by a LAI-2000 plant canopy analyzer to that of central leaflet width method (de Jesus Junior et al., 2001). However, correlation coefficient measures linear association rather than agreement between methods (Bland and Altman, 1986, 2010; Hopkins, 2004). Simply stated, correlation is used to measure the strength of the relationship between variables and it is inappropriate for quantifying systematic differences between two

* Corresponding author.
E-mail address: neil.gudmestad@ndsu.edu (N.C. Gudmestad).

http://dx.doi.org/10.1016/j.cropro.2017.07.019
0261-2194/

methods.

Another popular method, ordinary least square (OLS) regression is also not appropriate for measuring agreement because the measurement data sets are typically subject to random errors (Linnet, 1998; Ludbrook, 2002, 2010). However, OLS analysis assumes only *y* variables are subject to random errors and *x* variables as fixed and results in biased estimation of slope and inaccurate testing of hypotheses (Cornbleet and Gochman, 1979; Linnet, 1998; Ludbrook, 2010; Parvin, 1984). The Concordance correlation coefficient (CCC) developed by Lin is another method for assessing agreement between measurements made by multiple methods, observers, or instruments (Barnhart et al., 2007a; Lin, 1989). However, CCC is dependent on between subject variability and high CCC values imply large variability even if the individual difference between measurements remain the same (Atkinson and Nevill, 1997; Barnhart et al., 2007b). Other statistical methods, such as a paired *t*-test, are occasionally used by researchers for agreement measurement. However, there is a chance that paired t-tests may overestimate lower range values and underestimate higher range values when overall mean differences of the results of two methods are included (van Stralen et al., 2008).

The above statistical methods are prone to erroneous conclusions in quantifying the agreement between two methods. For this reason, an alternative approach has been proposed based on graphical techniques and simple calculations for the comparison of different methods (Bland and Altman, 1986). The Bland-Altman method is widely accepted and highly cited (over 36,000 times) across various disciplines of peer-reviewed research (Giavarina, 2015). However, to our knowledge this statistical approach has not been used in plant pathological research. It is important to assess agreement between measurements made by two different methods and previous phytopathological studies have not provided the equivalency tests between different disease assessment methods or disease assessments from multiple raters (Bardsley and Ngugi, 2013; Yi et al., 2008).

Verticillium wilt, caused by the fungus *Verticillium dahliae,* is an important soilborne disease of potato. Verticillium wilt is the most economically damaging disease of the French fry processing sector in the U.S.A. (Rowe and Powelson, 2002). It causes a vascular wilt in host plants by blocking xylem elements and disrupting water movement (Johnson and Dung, 2010). Typical symptoms on potato include wilting, chlorosis, and necrosis which progress upward from the base of the plant (Dung et al., 2012; Johnson and Dung, 2010). The net result of which is a gradual loss of green canopy and ground cover. The fungus survives in soil for prolonged periods (14 years) as microsclerotia (Davis et al., 2001; Powelson et al., 1993; Wilhelm, 1955). *V. dahliae* propagule levels in soil are known to influence wilt severity (Ben-Yephet and Szmulewich, 1985; Gudmestad et al., 2007; Nicot and Rouse, 1987). Severe wilt occurs under favorable cultural and environmental factors and can cause yield losses up to 30% (Cappaert et al., 1992; Dung et al., 2012).

Traditionally, Verticillium wilt severity is assessed visually by rating the percentage of plants affected at a single stage or multiple stages of plant growth (Pasche et al., 2013, 2014; Taylor et al., 2005). For example, the individual responsible for rating directly observes the experimental plot and estimates the percentage of plants affected. The reliability and accuracy of visual assessment of plant disease has been discussed in detail (Bardsley and Ngugi, 2013; Bock et al., 2010; Nita et al., 2003). Advances in technology for assessing disease severity include computerized disease rating methods and rapid detection technologies, which are being increasingly implemented in plant pathology research. Computerized plant disease quantification methods such as Assess (Lamari, 2002) and Digital Image Processing (Barbedo, 2014) are frequently used in plant pathology. The use of technology has extended to smartphone operated apps delivering relevant plant disease information and diagnostic tools to a user at the site of pathogen origin (Pethybridge and Nelson, 2015). Canopeo is a smart phone app developed and released by the Oklahoma State University App Center for rapid measurement of percentage green canopy cover. Canopeo is easy to use and can be downloaded for free on Android and IOS operated smartphones. This app has been increasingly adopted by growers for estimating percentage green plant canopy during the growing season (Gudmestad, personal observation).

Our aim is to explain the Bland-Altman method using examples of visual and Canopeo methods of wilt measurement. The first objective of this study is to provide step by step details for design, analysis, and interpretation of the Bland-Altman method for plant pathology research. Another objective is to provide guidance on the use of Bland-Altman method for measuring systemic differences between visual and Canopeo disease rating methods.

2. Materials and methods

Two field trials were established near Park Rapids, MN during May 4–6, 2015. Each study was established separately to test novel chemical and biological treatments for the management of Verticillium wilt of potato and henceforth are referred to as a trial 1 and a trial 2, respectively. The experimental design for both trials was a randomized complete block with six replications per treatment and each experimental unit consisted of four 9.1 m long by 0.9 m wide rows. The required fertility and weed control program to grow a French fry processing crop using cv. Russet Burbank was applied by the grower cooperator as required. Verticillium wilt was allowed to develop naturally throughout the crop growing season. The percentage wilt was rated in each experimental unit by visual observation followed by a smart phone operated Canopeo app. For both trials, percentage wilt rating (visual and canopeo) was performed by a single evaluator (first author). Wilt severity was estimated on the two middle rows in four 1-square meter sampling areas (2 per row). These sampling areas were approximately 1m from the beginning and 1m from the end of the paired center rows and ignores the plot edges to reduce interplot interference. Nine weeks after planting, recurrent wilt rating (weekly basis) was performed over a seven week period from tuber initiation-early bulking (week 1–5) and late-bulking-tuber maturation (week 6–7) growth stages of potato (Yellareddygari et al., 2016).

2.1. Canopeo app

The Canopeo app was downloaded from Android Google Play App store onto a smartphone having a current software upgrade. Calibration of the Canopeo smartphone application was performed (Patrignani and Ochsner, 2015). Although the app does not estimate wilt symptoms directly, percentage healthy (green) canopy is calculated which in turn is used to obtain wilt rating (disease severity = 100- healthy canopy). The app accesses the camera on the phone to estimate percentage green canopy cover. Through the entire crop season a single user (1.78 m tall first author) operated the app and the photographs were taken approximately from shoulder height (1.5 m). A digital photograph was taken by holding the camera parallel to the ground and an 'OK' on screen was accepted to obtain a percentage green canopy estimation. This process was repeated for each photograph taken in four sampling areas of an experimental unit. All the wilt ratings were performed approximately between 8.00 am and 3.00 pm on sunny to partially cloudy days. Photographs obtained on sunny and partial cloudy days were shown previously to have no influence on the Canopeo

image analysis (Patrignani and Ochsner, 2015).

2.2. Bland and Altman method

The first step is to plot the difference between two methods of measurement against the mean of the two methods on the x-y scatter plot (Bland and Altman, 1986). Assuming two wilt rating methods as M1 and M2, the difference between two measurements (M1-M2) is plotted as the y-coordinate and the mean of measurements ((M1+M2)/2) is plotted as the x-coordinate. The bias between two methods is a measure of the lack of agreement and is estimated by the mean difference ($\overline{d}$) and the variation around the bias is estimated as standard deviation (*sd*). Assuming the differences are normally distributed, the variation of the results is calculated as ± 1.96x*sd* and referred as limits of agreement (LOA). Although the violation of normal distribution assumption may not be a serious problem, logarithmic transformation of original data can be applied for skewness correction (Bland and Altman, 1999; Giavarina, 2015; Grilo and Grilo, 2012). If skewness remains after the log transformation, a regression approach to evaluate the agreement is suggested (Bland and Altman, 1999; Grilo and Grilo, 2012). LOA values indicate that 95% of the data points range between the limits of mean difference ($\overline{d} \pm 1.96x sd$) and was used for visual examination of good agreement between two rating methods. Visual inspection of Bland-Altman plots is needed to identify bias and this step is important to find systematic differences between methods before replacing a reference method with a new method. Fixed bias is observed when one method gives a continuous amount of high or low values compared to the other method and proportional bias observed when one method gives high or low values that is proportional to the other measured variable (Ludbrook, 1997).

2.3. Repeated measures

Repeated measures analysis is common in plant pathology because recurring disease ratings are performed over time on the same experimental units (Harveson and Rush, 2002; Lipps and Madden, 1992; Shah and Madden, 2004; Xiao and Subbarao, 2000). For accurate comparison of methods it is appropriate to use all repeated measurements for the Bland-Altman plot. One problem with repeated measures is that methods may seem high in agreement due to an underestimation of *sd*, resulting in lower estimation of random errors (van Stralen et al., 2008). To overcome this, a mixed effects model is proposed (Carstensen et al., 2008; Myles and Cui, 2007). This model estimates within-subject variation, where each subject has a different intercept and slope over the observation period (Laird and Ware, 1982; Myles and Cui, 2007).

2.4. Statistical analysis

All plots for the Bland-Altman method comparisons were graphed using SAS SGPLOT in version 9.3. Final disease ratings from trial 1 and trial 2 were used to demonstrate the step by step Bland-Altman method design for comparing two methods. The Pearson correlation coefficient and OLS regression association between two different disease assessments was estimated using SAS PROC CORR and SAS PROC REG, respectively. The correlation coefficient between rating methods were obtained for lower wilt rating (wilt ≥20 and ≤ 50%), higher wilt rating (wilt ≥40%) and across the entire range of wilt measures (%) being assessed. Also, concordance correlation coefficient strength of agreement (p_c) between Canopeo and visual measurements was estimated for both trials. Concordance measures the correspondence between two wilt readings by measuring the variation of the fitted linear relationship from the 45° line through the origin (equality line) and precision by measuring how far each observation deviates from the fitted line (Lin et al., 2012). The CCC agreement index is defined as poor ($p_c < 0.90$), moderate ($0.90 \leq p_c < 0.95$), good ($0.95 \leq p_c \leq 0.99$), or excellent ($p_c > 0.99$) (McBride, 2005). Since the current study estimated the wilt at weekly intervals (repeated measure), a mixed effects model (SAS PROC MIXED) was fitted for the complete (1–7 weeks disease severity) data set. A preset cutoff limit of ≤5% (wilt) heterogeneity between rating methods was considered acceptable prior to Bland-Altman comparison. The mixed effects model was designed with method of rating as a fixed effect and time (week) of the disease measurement as the random effect. The estimated variance components within each method were used for creating Bland-Altman plots (Carstensen et al., 2008). During the wilt measurement, visual assessment was assumed as the reference and therefore, Canopeo measurements are compared to the reference. For example, negative or positive bias (underestimation or overestimation) between two methods of wilt measurements is compared to the reference (visual assessment).

3. Results

Correlation was not constant within the range of data assessed (Fig. 1) and an inability to measure systematic differences (Fig. 2) between methods was demonstrated. For trial 1, the correlation coefficient between visual and Canopeo ratings was higher ($r = 0.96, P = < 0.0001$) when the entire broad range of values were selected and lower when specific populations were selected (higher range of values ($r = 0.85, P = < 0.0001$) and lower range of values ($r = 0.79, P = < 0.0001$)) (Fig. 1). Similarly for trial 2, correlation coefficients varied across the entire broad ($r = 0.98, P = < 0.0001$), lower ($r = 0.84, P = < 0.0001$) and higher ($r = 0.89, P = < 0.0001$) range of wilt measurements (Fig. 1). This indicates that Pearson correlation coefficient is sensitive to the range of values that are in the study. For this study (both trials), the visual observation of equality line demonstrates that one method overestimates Verticillium wilt at the beginning and underestimates towards the end of the crop growing season (Fig. 2). This information is not visualized when correlation alone is plotted (regression line). A systematic difference between the methods is visually observed from plots based on the difference between the regression line and the equality line (Fig. 2). Two separate X_{est} (visual) and Y_{est} (Canopeo) lines representing the minimization of sums of the squares of the deviations of the *x* values and sums of the squares of the deviations of the *y* values, respectively, were calculated for OLS regression. The coefficients in the OLS regression model for trial 1 and trial 2 were $Y_{can} = 4.37 + 0.87visual$, $X_{vis} = -0.54 + 1.08canopeo$ and $Y_{can} = 1.69 + 0.89visual$, $X_{vis} = 1.72 + 1.06canopeo$, respectively. For perfect agreement between methods the Y_{est} and X_{est} lines are identical, however, OLS results indicate that the two lines are not identical. In this context, the lines were distinctly separate because more often both methods in comparison studies are subjected to random error and OLS assumption (values of *x* variable (fixed) and *y* variable (random)) is rarely satisfied.

The difference between Canopeo and visual assessment, average of two measurements, bias ($\overline{d}$), standard deviation and LOA were calculated for the final wilt data (Table 1 and Table 2) and were used to demonstrate the design of Bland-Altman method. Bland-Altman analysis (final wilt severity) scatter plots graphed the bias and LOA's ($\overline{d} = -3.4$ (−10.11 to 3.32) and $\overline{d} = -1.66$ (−6.34 to 3.01) for trial 1 and trial 2, respectively (%)) between the two methods of rating (Fig. 3). The positive bias indicates that values measured by one method are higher than the other and negative bias indicates otherwise. The negative bias (in both trials) indicates that Canopeo

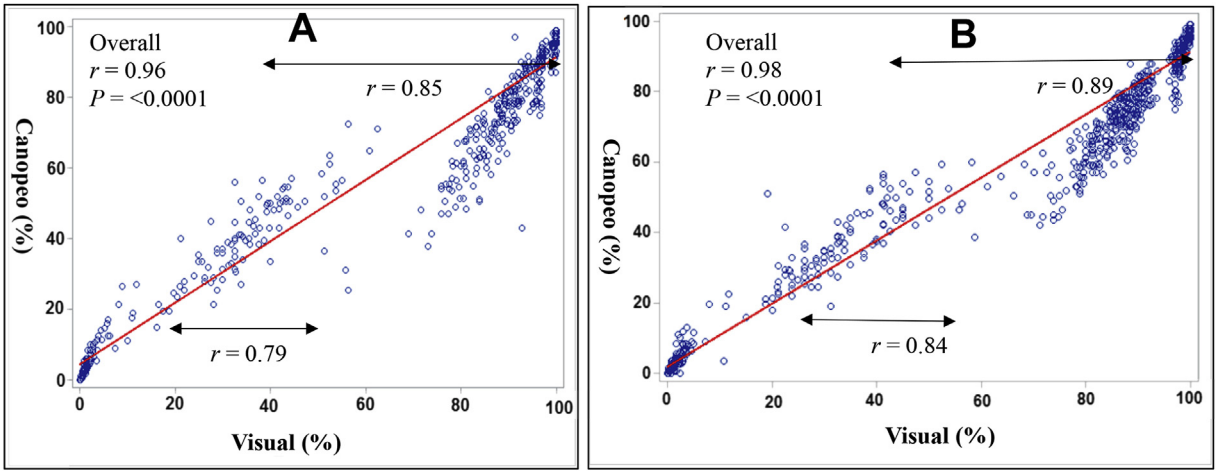

Fig. 1. Scatter plots demonstrating the correlations between two methods is dependent on the range of values. **A**, Correlations for trial 1: Lower range (wilt ≥20 and ≤ 50%) is $r = 0.79$, *P*-value <0.0001; Higher range (wilt ≥40%) is $r = 0.85$, *P*-value <0.0001. **B**, Correlations for trial 2: Lower range (wilt ≥20 and ≤ 50%) is $r = 0.84$, *P*-value <0.0001; Higher range (wilt ≥40%) is $r = 0.89$, *P*-value <0.0001.

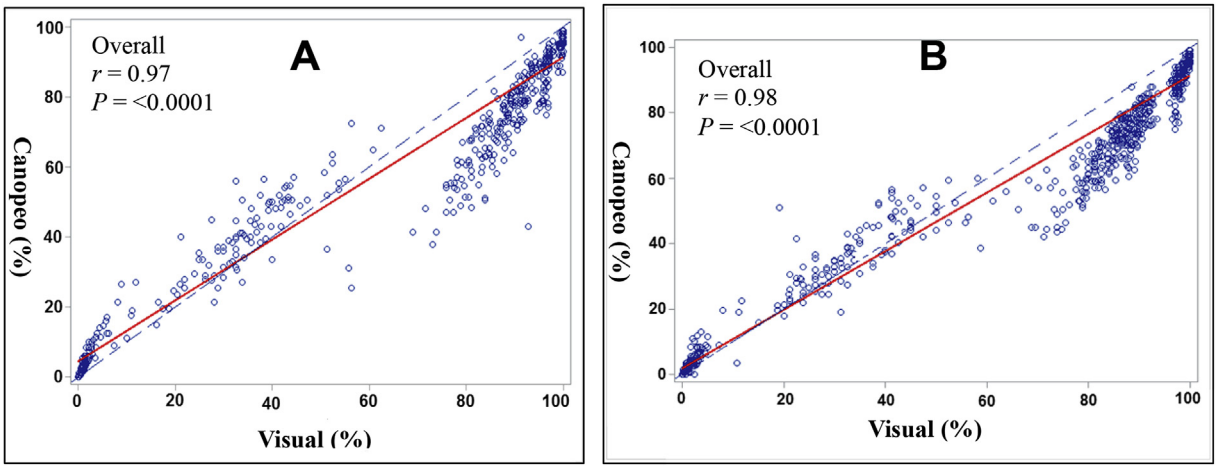

Fig. 2. Scatter plots demonstrating that correlation coefficient is not appropriate for measuring systematic difference between methods. **A**, trial 1 and **B**, trial 2. The solid line represents the regression line, where most data points are clustered for a high correlation and the broken line represents the 45° equality or agreement line, where data points should align for perfect agreement between methods. For both trials, visual observation show distinct difference between the lines. For lower range values the regression line is higher than equality line and vice versa.

final wilt ratings were lower when compared to the reference (visual assessment). Similarly, 95% LOA quantifies whether methods agree sufficiently for use in wilt assessment. For example, in Fig. 3 and 95% LOA between two rating methods of (−10.11 to 3.32) indicate that for 95% of observations, wilt measurement made by one method (new) was between −10.11% less and 3.32% more than a measurement made by the reference method.

Repeated measures wilt assessment (random effect model) by Canopeo and visual methods were used for assessing final agreement between methods. For trial 1, the mean difference (bias), standard deviation and LOA of method comparison were calculated as −5.07, 9.13, and −22.98 to 12.82 (lower to upper LOA), respectively (Fig. 4). Similarly, mean difference, standard deviation, and LOA of method agreement for trial 2 were estimated as −5.91, 8.05, and −21.71 to 9.88 (lower to upper LOA), respectively (Fig. 5). For both trials, the estimated bias is negative and higher than preset cutoff limit (5%), indicating that values measured by Canopeo are lower than the reference. Also, CCC was used as an alternative to the Bland-Altman method for the agreement measurement between two methods. For trial 1 and trial 2, the concordance (p_c) between Canopeo and visual assessments were 0.95 (95% CI: 0.947, 0.959) and 0.96 (CI: 0.952, 0.962), respectively. The CCC strength of agreement between two methods is good but does not approach unity required for perfect agreement.

4. Discussion

The primary purpose of this study was to make plant pathologists aware that alternative and more appropriate methods are available for comparing disease assessment methods compared to conventional correlation coefficients and OLS regression models. The Bland-Altman method is easy to calculate and interpret which is a probable explanation for its wide acceptance and use in other disciplines (Giavarina, 2015). Although difficult to execute, other alternative statistical analyses can be used such as major axis regression (Deming's method), bivariate least median squares method and ordinary least product regression (geometric mean regression) for method comparison study (Ludbrook, 2010). For methods estimating qualitative variables, the Kappa coefficient is recommended for method comparison studies (Ludbrook, 2010).

Table 1
Design of Bland-Altman method for final Verticillium wilt disease severity for trial 1.

Obs	Canopeo (M1) (%)	Visual (M2) (%)	M1-M2 (%)	Mean ((M1+M2)/2) (%)
1	98.92	99.58	−0.67	99.25
2	96.17	98.54	−2.38	97.35
3	94.75	97.96	−3.21	96.35
4	92.67	97.04	−4.38	94.85
5	95.08	98.50	−3.42	96.79
6	95.83	98.63	−2.79	97.23
7	90.33	96.29	−5.96	93.31
8	94.17	97.79	−3.63	95.98
9	95.17	98.38	−3.21	96.77
10	95.17	98.38	−3.21	96.77
11	90.83	96.54	−5.71	93.69
12	91.33	96.79	−5.46	94.06
13	95.42	98.58	−3.17	97.00
14	96.83	99.04	−2.21	97.94
15	94.67	97.79	−3.13	96.23
16	96.92	98.79	−1.88	97.85
Mean ($\bar{d}$)			**−3.40**	
Standard deviation (*sd*)			**1.41**	
LOA ($\bar{d} \pm 1.96^{*}sd$)			**(-0.64, -6.16)**	

*Note: The replication data were averaged to fit the table; LOA, represent limits of agreement.

Table 2
Design of Bland-Altman method for final Verticillium wilt disease severity for trial 2.

Obs	Canopeo (M1) (%)	Visual (M2) (%)	M1-M2 (%)	Mean ((M1+M2)/2) (%)
1	99.00	99.29	−0.29	99.15
2	95.50	97.46	−1.96	96.48
3	98.17	98.96	−0.79	98.56
4	97.33	98.58	−1.25	97.96
5	97.33	98.58	−1.25	97.96
6	97.42	98.71	−1.29	98.06
7	95.92	97.63	−1.71	96.77
8	95.33	98.17	−2.83	96.75
9	98.08	99.08	−1.00	98.58
10	96.92	98.67	−1.75	97.79
11	97.92	98.42	−0.50	98.17
12	95.58	98.25	−2.67	96.92
13	94.08	97.08	−3.00	95.58
14	94.83	98.04	−3.21	96.44
15	97.00	98.04	−1.04	97.52
16	97.25	98.67	−1.42	97.96
17	97.33	98.42	−1.08	97.88
18	94.50	97.25	−2.75	95.88
19	97.33	97.33	0.00	97.33
20	95.33	97.38	−2.04	96.35
21	95.75	98.04	−2.29	96.90
22	95.83	98.29	−2.46	97.06
Mean ($\bar{d}$)			**−1.66**	
Standard deviation (*sd*)			**0.91**	
LOA ($\bar{d} \pm 1.96^{*}sd$)			**(3.02, -6.34)**	

*Note: The replication data were averaged to fit the table; LOA, represent limits of agreement.

This study demonstrated two reasons that the Pearson's correlation coefficient is inappropriate for comparing agreement between two methods (van Stralen et al., 2008). First, correlation is influenced by a range of values when compared to the overall correlation obtained from a plant population. For example, wilt severity measurements between 20 and 50% had a lower correlation (0.79) than the overall correlation (0.97) over the entire data set. This indicates that even after there is strong association among overall wilt infected plants, there are differences among individuals or subgroups within the population which an overall correlation misses. Secondly, correlation fails to show the systematic difference between the two methods because perfect correlation is not same as perfect agreement. Perfect correlation between two methods occurs when data points align close to any regression line and perfect agreement is observed only when data points cluster along the equality line (the linear line drawn from zero) (Bland and Altman, 1986).

OLS regression is another method used frequently to compare agreement between two methods (Linnet, 1998; Ludbrook, 2002, 2010). The OLS regression model for a population of Y on X can be written as $Y = \beta 0 + \beta 1X + \varepsilon$, where the parameters $\beta 0$,$\beta 1$ and ε are intercept, slope, and random error, respectively. From the above equation, the error term represents the random measurement error of the measuring instrument or the effects of random variation in *Y* (Ludbrook, 1997). It is clear that only the values of *y* variable are attended by random error and the values of *x* variable are fixed and without random measurement error. Since methods in comparison are usually subjected to some random measurement error, plotting

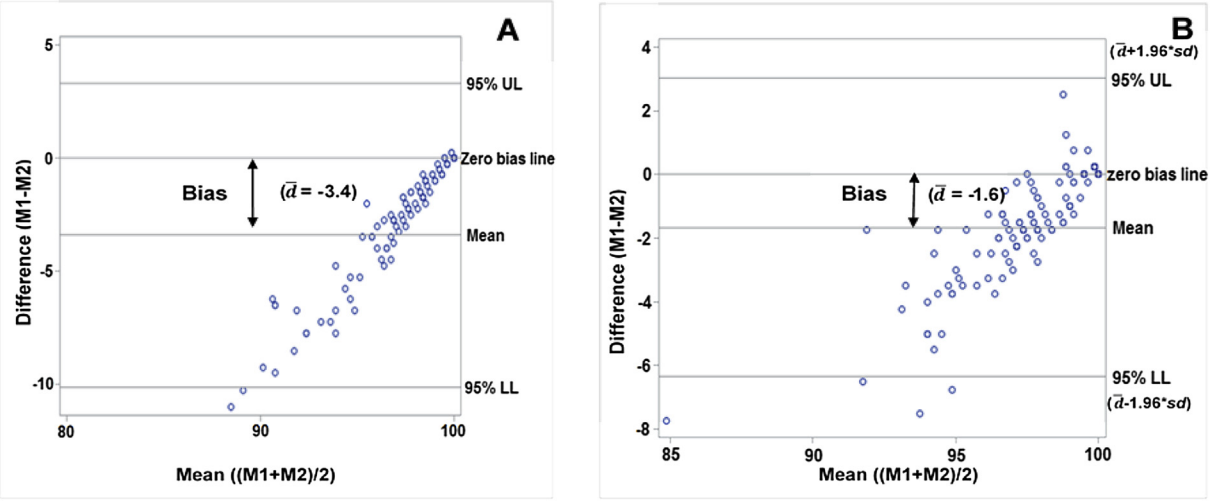

Fig. 3. Bland-Altman scatter plot design for final disease severity. **A**, **Trial 1**, $\bar{d}$, sd, and LOA were estimated as −3.4, 3.42, and −10.11 to 3.32 (lower to upper LOA), respectively. **B**, **Trial 2**, the mean difference ($\bar{d}$), standard deviation (*sd*), limits of agreement (LOA) were estimated as −1.66, 2.38, and −6.34 to 3.01 (lower to upper LOA), respectively. The estimated negative bias for both trials indicates that Canopeo app is reading wilt rating lower than visual assessment.

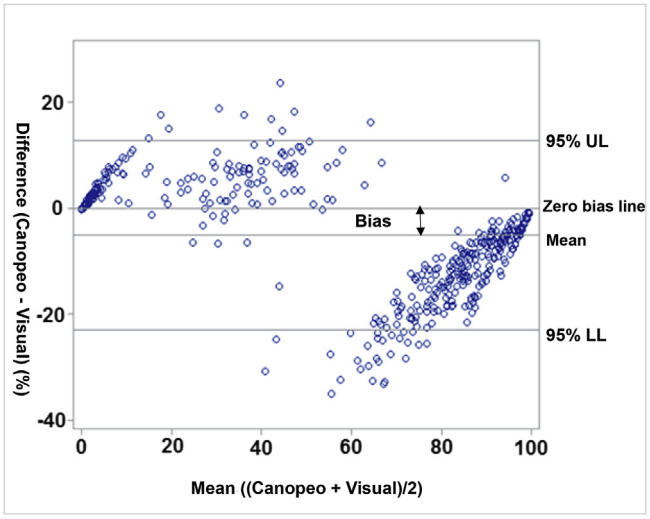

Fig. 4. Bland-Altman scatter plot (trial 1) design for repeated measures. The mean difference ($\bar{\bar{d}}$), standard deviation (*sd*), limits of agreement (LOA) were estimated as −5.07, 9.13, and (−22.98, 12.82) (lower, upper LOA), respectively. The estimated bias (−5.07) is greater than preset cutoff limit 5%, indicating that the two method measurements are not in agreement.

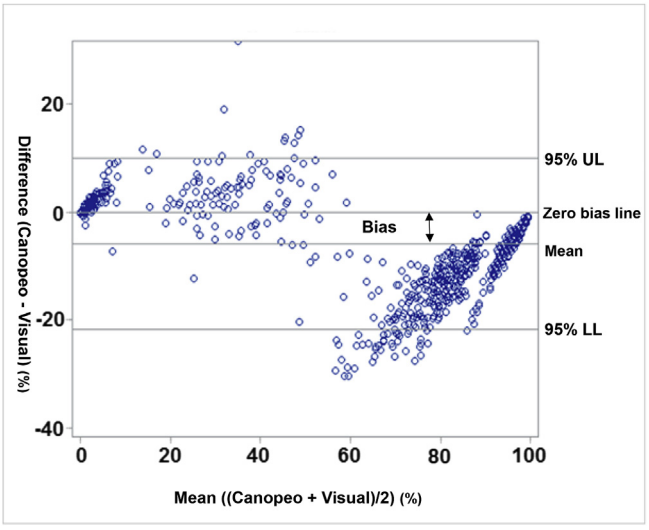

Fig. 5. Bland-Altman scatter plot (trial 2) design for repeated measures. The mean difference ($\bar{\bar{d}}$), standard deviation (*sd*), limits of agreement (LOA) were estimated as −5.91, 8.05, and (−21.71, 9.88) (lower, upper LOA), respectively. The estimated bias (−5.91) is greater than preset cutoff limit 5%, indicating that the two method measurements are not in agreement.

two method measurements as x-coordinate and y-coordinate of OLS regression results in biased estimation of slope. In the present study, we clearly demonstrated that Pearson correlation and OLS regression were inadequate to show that a new method disagreed with an older method of disease assessment. Another method, CCC measured the strength of agreement between visual and Canopeo measurements as moderate to good. However, CCC index has limitations regarding the assessment of individual differences between multiple observer's measurements on the same subject and inconsistent interpretation for the same variable across different populations in multiple projects (Barnhart et al., 2014).

Before starting the Bland-Altman analysis, the cutoff limit for heterogeneity between two study methods was set at five percent. The cutoff limit is set low because, even 5% of wilt not accounted for in potato may cause adverse effects on potential yield. Variable yield losses have been reported with Verticillium wilt on potato (Johnson and Dung, 2010). The cutoff limit hypothesizes that when estimated bias between visual and Canopeo methods is larger than the cutoff limit, the new method is rejected. When the bias is small and consistent, the estimated mean difference can be adjusted by addition or deletion from the results obtained with a new method (Bland and Altman, 2010). The cutoff limit is not bounded by any statistical method and can be set by the investigator when a new method is compared to the standard method. For example, when compared to established disease severity evaluator, more than 25–50% of variation of results with a new rating method is unacceptable. Some applicable methods measuring sensitive disease influencing parameters like temperature effect, moisture level, and pathogen inoculum rate may need precise error cutoff limits for minimizing error. For example, when a specific concentration of conidial suspension is required for the development of adequate infection, the variation of results between two methods measuring

the cell concentration should be zero.

Visual inspection of the scatter plot is key for identifying bias and random error between two methods. When both methods are in agreement (highly unlikely) the wilt data point should line with the zero bias line on a scatter plot. Simply stated, the mean difference of zero indicates that two methods were identical in assessing percentage wilt. Repeated wilt ratings are subjected to random error and this is estimated as limits of agreement around the bias. Ideally the Bland-Altman results should have very small bias and narrow limits of agreement (Cecconi et al., 2009). The current study results showed lower bias (5–6%) and restricted LOA (−22.9 lower and 12.8 upper limit). However, the results from both trials demonstrated a greater bias compared to the set cutoff limit and, therefore, the Canopeo results were not in agreement with standard visual assessment. Visual observation of scatter plots (Figs. 4 and 5) indicated that agreement (data pattern from figure) decreases as Verticillium wilt severity increased. This indicates that towards final wilt rating, one method (new) is constant in assessing lower wilt development when compared to the reference method. Although the Bland-Altman statistical analysis demonstrated a lack of agreement between two rating methods, a researcher's judgement is still needed to decide whether a new method can be used in place of reference method of wilt assessment. Also, visual examination of plots reveal different phases of wilt data (clustered as phases) and separate Bland-Altman analysis may have been more appropriate. The clustering could be due to uneven progress of wilt over time, the disease intensity is slow to begin and peaks rapidly towards final stages of crop growth. However, we rationalize the results because the objective was to demonstrate the methodology rather than determining which method is more reliable for wilt measurement.

In the future, it is likely that with the availability of knowledge-based information and sensing techniques that plant disease detection and quantification will be significantly influenced by smartphone and mobile phone solutions (Mahlein, 2016). However, the number of downloads (free/paid) and usage of these apps ultimately depends on potential performance and quality of the crop disease management information provided. Prior to purchase of most disease diagnostic apps, their potential accuracy is unknown and it can be extremely difficult for a user to determine their practical utility prior to usage (Rodrigues et al., 2013). For most new methods, the true measurement value is unknown and should be compared to a standard method (if available) prior to replacement or for interchangeable use (Bland and Altman, 2010). We think the Canopeo app was appropriate to evaluate since the progression of a vascular wilt, such as the one caused by *V. dahliae* in potato, causes the plant canopy to prematurely senesce leading to reductions in plant canopy and ground cover over time (Pasche et al., 2013; Taylor et al., 2005).

The primary goal of the study reported here was to raise awareness of the Bland-Altman comparison of agreement for use in pest management studies and specifically, between two methods of disease severity assessment. For studying agreement between methods, simple graphs and a hand calculation of means and standard deviations can suffice and be a substitute for complicated statistical programming. We also want to raise awareness that some current and commonly used statistical analyses are inadequate to compare disease severity assessment methods and that alternatives exist that should be explored by the plant pathology research community. Additionally, studies performed under different conditions and multiple raters are needed for determining the potential of the Canopeo app. In the current era of smart technology, the future for phytopathology apps is optimistic and comparison studies are needed for new apps.

Acknowledgements

This work was supported largely by chemical companies (BASF, Bayer, Monsanto, Syngenta, and Valent) and partially by potato growers from North Dakota and Minnesota. The authors thank Canopeo app center for technical support.

References

Rowe, R.C., Powelson, M.L., 2002. Potato early dying: management changes in a changing production environment. Plant Dis. 86, 1184–1193.
Altman, D.G., Bland, J.M., 1983. Measurement in medicine: the analysis of method comparison studies. Statistician 32, 307–317.
Atkinson, G., Nevill, A., 1997. Comment on the use of concordance correlation to assess the agreement between two variables. Biometrics 53, 775–777.
Bade, C.A., Carmona, M.A., 2011. Comparison of methods to assess severity of common rust caused by *Puccinia sorghi* in maize. Trop. Plant Pathol. 36, 264–266.
Barbedo, J.G.A., 2014. An automatic method to detect and measure leaf disease symptoms using digital image processing. Plant Dis. 98, 1709–1716.
Bardsley, S.J., Ngugi, H.K., 2013. Reliability and accuracy of visual methods to quantify severity of foliar bacterial spot symptoms on peach and nectarine. Plant Pathol. 62, 460–474.
Barnhart, H.X., Haber, M., Lokhnygina, Y., Kosinski, A.S., 2007a. Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement. J. Biopharm. Stat. 17, 721–738.
Barnhart, H.X., Kosinski, A.S., Haber, M., 2007b. Assessing individual agreement. J. Biopharm. 17, 697–719.
Barnhart, H.X., Yow, E., Crowley, A.L., Daubert, M.A., Rabineau, D., Bigelow, R., Pencina, M., Douglas, P.S., 2014. Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. Stat. Methods Med. Res. 0, 1–20.
Ben-Yephet, Y., Szmulewich, Y., 1985. Inoculum levels of *Verticillium dahliae* in the soils of the hot semi-arid Negev regions of Israel. Phytoparasitica 13, 193–200.
Bland, J.M., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1, 307–310.
Bland, J.M., Altman, D.G., 1999. Measuring agreement in method comparison studies. Stat. Methods Med. Res. 8, 135–160.
Bland, J.M., Altman, D.G., 2010. Statistical methods for assessing agreement between two methods of clinical measurement. Int. J. Nurs. Stud. 47, 931–936.
Bock, C.H., Poole, G.H., Parker, P.E., Gottwald, T.R., 2010. Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. Crit. Rev. Plant Sci. 29, 59–107.
Cappaert, M.R., Powelson, M.L., Christensen, N.W., Crowe, F.J., 1992. Influence of irrigation on severity of potato early dying and tuber yield. Phytopathology 82, 1448–1453.
Carstensen, B., Simpson, J., Gurrin, L.C., 2008. Statistical models for assessing agreement in method comparison studies with replicate measurements. Int. J. Biostat. 4, 1–26.
Cecconi, M., Rhodes, A., Poloniecki, J., Roca, D.G., Grounds, R.M., 2009. Bench-to-bedside review: the importance of the precision of the reference technique in method comparison studies-with specific reference to the measurement of cardiac output. Crit. Care 13, 201–207.
Cornbleet, P.J., Gochman, N., 1979. Incorrect least-squares regression coefficients in method comparison analysis. Clin. Chem. 25, 432–438.
Davis, J.R., Huisman, O.C., Everson, D.O., Schneider, A.T., 2001. Verticillium wilt of potato: a model of key factors related to disease severity and tuber yield in southeastern Idaho. Am. J. Potato Res. 78, 291–300.
de Jesus Junior, W.C., do Vale, F.X.R., Coelho, R.R., Costa, L.C., 2001. Comparison of two methods for estimating leaf area index on common bean. Agron. J. 93, 989–991.
Dung, J.K.S., Ingram, J.T., Cummings, T.F., Johnson, D.A., 2012. Impact of seed lot infection on the development of black dot and Verticillium wilt of potato in Washington. Plant Dis. 96, 1179–1184.
Giavarina, D., 2015. Understanding Bland altman analysis. Biochem. Medica 25, 141–151.
Grilo, L.M., Grilo, H.L., 2012. Comparison of clinical data based on limits of agreement. Biom. Lett. 49, 45–56.
Gudmestad, N.C., Taylor, R.J., Pasche, J.S., 2007. Management of soilborne diseases of potato. Australas. Plant Path 36, 109–115.
Harveson, R.M., Rush, C.M., 2002. The influence of irrigation frequency and cultivar blends on the severity of multiple root diseases in sugar beets. Plant Dis. 86, 901–908.
Hopkins, W.G., 2004. Bias in bland-altman but not regression validity analyses. Sportscience 8, 42–46.
Johnson, D.A., Dung, J.K.S., 2010. Verticillium wilt of potato-the pathogen, disease and management. Can. J. Plant Pathol. 32, 58–67.
Laird, N.M., Ware, J.H., 1982. Random effects models for longitudinal data. Biometrics 38, 963–974.
Lamari, L., 2002. ASSESS: Image Analysis Software for Plant Disease Quantification. American Phytopathological Society, St. Paul, MN.
Lin, L.I., 1989. A concordance correlation coefficient to evaluate reproducibility.

Biometrics 45, 225–268.
Lin, H.M., Kim, H.Y., Williamson, J.M., Lesser, V.M., 2012. Estimating agreement coefficients from sample survey data. Surv. Methodol. 38, 63–72.
Linnet, K., 1998. Performance of Deming regression analysis in case of mispecified analytical error ratio in method comparison studies. Clin. Chem. 44, 1024–1031.
Lipps, P.E., Madden, L.V., 1992. Effects of plot size and border width on assessment of powdery mildew of winter wheat. Plant Dis. 76, 299–303.
Ludbrook, J., 1997. Comparing methods of measurement. Clin. Exp. Pharmacol. P. 24, 193–203.
Ludbrook, J., 2002. Statistical techniques for comparing measurers and methods of measurement: a critical review. Clin. Exp. Pharmacol. P. 29, 527–536.
Ludbrook, J., 2010. Linear regression analysis for comparing two measures or methods pf measurement: but which regression? Clin. Exp. Pharmacol. P. 37, 692–699.
Mahlein, A., 2016. Plant disease detection by imaging sensors-parallel and specific demands for precision agriculture and plant phenotyping. Plant Dis. 100, 41–51.
McBride, G.B., 2005. A proposal for strength-of-agreement criteria for Lin's Concordance Correlation Coefficient. NIWA Client Rep. HAM2005–H2062.
Myles, P.S., Cui, J., 2007. Using the Bland-Altman method to measure agreement with repeated measures. Brit. J. Anaesth. 99, 309–311.
Nicot, P.C., Rouse, D.I., 1987. Relationship between soil inoculum density of Verticillium dahliae and systemic colonization of potato stems in commercial fields over time. Phytopathology 77, 1346–1355.
Nita, M., Ellis, M.A., Madden, L.V., 2003. Reliability and accuracy of visual estimation of Phomopsis leaf blight of strawberry. Phytopathology 93, 995–1005.
Parvin, C.A., 1984. A direct comparison of two slope-estimation techniques used in method-comparison studies. Clin. Chem. 30, 751–754.
Pasche, J.S., Thompson, A.L., Gudmestad, N.C., 2013. Quantification of field resistance to *Verticillium dahliae* in eight russet-skinned potato cultivars using real-time PCR. Am. J. Potato Res. 90, 158–170.
Pasche, J.S., Taylor, R.J., David, N., Gudmestad, N.C., 2014. Effect of soil temperature, injection depth, and metam sodium rate on the management of Verticillium wilt of potato. Am. J. Potato Res. 91, 277–290.
Patrignani, A., Ochsner, T.E., 2015. Canopeo: a powerful new tool for measuring fractional green canopy cover. Agron. J. 107, 2312–2320.
Pethybridge, S.J., Nelson, S.C., 2015. Leaf doctor: a new portable application for quantifying plant disease severity. Plant Dis. 99, 1310–1316.
Powelson, M.L., Johnson, K.B., Rowe, R.C., 1993. Management of diseases caused by soilborne pathogens. In: Rowe, R.C. (Ed.), Potato Health Management. American Phytopathological Society, St. Paul, pp. 149–151.
Rodrigues, M.A., Visvanathan, A., Murchison, J.T., Brady, R.R., 2013. Radiology smartphone applications; current provision and cautions. Insights Imaging 4, 555–562.
Shah, D.A., Madden, L.V., 2004. Nonparametric analysis of ordinal data in designed factorial experiments. Phytopathology 94, 33–43.
Taylor, R.J., Pasche, J.S., Gudmestad, N.C., 2005. Influence of tillage and method of metam sodium application on distribution and survival of Verticillium dahliae in the soil and the development of potato early dying disease. Am. J. Potato Res. 82, 451–461.
van Stralen, K.J., Jager, K.J., Zoccali, C., Dekker, F.W., 2008. Agreement between methods. Kidney Int. 74, 1116–1120.
Wilhelm, S., 1955. Longevity of Verticillium wilt fungus in the laboratory and field. Phytopathology 45, 180–181.
Xiao, C.L., Subbarao, K.V., 2000. Effects of irrigation and *Verticillium dahliae* on cauliflower root and shoot growth dynamics. Phytopathology 90, 995–1004.
Yellareddygari, S.K.R., Pasche, J.S., Taylor, R.J., Gudmestad, N.C., 2016. Individual participant data meta-analysis of foliar fungicides applied for potato early blight management. Plant Dis. 100, 200–206.
Yi, Q., Wang, P.P., He, Y., 2008. Reliability analysis for continuous measurements: equivalence test for agreement. Stat. Med. 27, 2816–2825.