Accepted Manuscript
Version of Record at: https://doi.org/10.1093/jrsssc/qlad063

A novel agreement statistic using data on uncertainty in ratings

Jarcy Zee

University of Pennsylvania Perelman School of Medicine, Department of Biostatistics, Epidemiology, and Informatics, and Children's Hospital of Philadelphia, Philadelphia, PA, USA

E-mail: Jarcy.Zee@Pennmedicine.upenn.edu

Laura Mariani

University of Michigan, Department of Internal Medicine, Ann Arbor, MI, USA.

Laura Barisoni

Duke University, Department of Pathology, Durham, NC, USA.

Parag Mahajan

University of Delaware, Department of Economics, Newark, DE, USA.

Brenda Gillespie

University of Michigan, Department of Biostatistics, Ann Arbor, MI, USA.

Summary. Many existing methods for estimating agreement correct for chance agreement by adjusting the observed proportion agreement by the probability of chance agreement based on different assumptions. These assumptions may not always be appropriate, as demonstrated by pathologists' ratings of kidney biopsy descriptors. We propose a novel agreement statistic that accounts for the empirical probability of chance agreement, estimated by collecting additional data on rater uncertainty for each rating. A standard error estimator for the proposed statistic is derived. Simulation studies show that in most cases, our proposed statistic is unbiased in estimating the probability of agreement after removing chance agreement.

Keywords: Agreement; Chance Agreement; Rater Assessment; Reproducibility; Uncertainty

1. Introduction

In the absence of a gold standard or reference standard, an instrument is typically evaluated by estimating inter-rater (and intra-rater) agreement. Inter-rater agreement measures the degree to which different raters similarly rate an observation using the instrument, such as a scoring system or protocol. Many agreement statistics have been proposed to evaluate reproducibility between two raters giving dichotomous responses. The most commonly used are proportion agreement and several statistics that correct for agreement by chance, which can be defined as having no systematic assessment of an observation. While an extreme example of chance occurs when raters flip coins to determine their ratings, a more realistic and common example occurs if raters are uncertain or unsure of their ratings—e.g., due to difficulty in interpreting how an instrument applies to an observation or even due to fatigue, haste, or variations in environment or technology.

Commonly used existing agreement statistics that correct for chance include Cohen's κ, Scott's π, Krippendorff's α, Prevalence- and Bias-Adjusted Kappa (PABAK), and Gwet's AC_1. Assuming no missing data, all of these agreement statistics are based on rating results drawn from a 2×2 table of possible responses (Table 1), where 0 and 1 represent two response categories such as absence and presence of a feature (respectively), and the entry in each cell represents the probability, denoted by p_{ab}, that rater A gives response $a \in \{0, 1\}$ and rater B gives response $b \in \{0, 1\}$. Values in the margins represent the

		Rater B 0	Rater B 1	Total
Rater A	0	p_{00}	p_{01}	$p_{0.}$
	1	p_{10}	p_{11}	$p_{1.}$
	Total	$p_{.0}$	$p_{.1}$	1

Table 1: 2×2 table of rating results

probabilities of responses of 0 or 1 by each rater, with a dot to indicate a value of either 0 or 1 for the other rater. These marginal probabilities can also be interpreted as the rater-specific prevalences of each response. All probabilities in the table can be empirically estimated by $\hat{p}_{ab} = n_{ab}/N$ where n_{ab} is the number of observations for which rater A gave response a and rater B gave response b and N is the total number of observations. While several existing agreement statistics were originally defined in terms of population parameters, they are calculated using sample-based estimates and we therefore use the sample estimates in their definitions below.

Among the chance-corrected statistics, almost all are calculated by adjusting the observed proportion agreement, $\hat{P}_o = \hat{p}_{00} + \hat{p}_{11}$, by an estimated probability of chance agreement, $\hat{P}_c$, and take the form:

$$\frac{\hat{P}_o - \hat{P}_c}{1 - \hat{P}_c}.$$

The numerator represents the agreement in excess of chance, while the denominator is a normalizing factor. The primary differences among existing agreement statistics are in the way the probability of chance agreement P_c is estimated. These statistics are based on different assumptions and thus result in different agreement estimates.

For example, disease diagnoses for patients with nephrotic syndrome, a rare kidney disorder, are conventionally made based on pathologists' interpretive assessments of kidney biopsy tissue. In the Nephrotic Syndrome Study Network (NEPTUNE), a part of the National Institutes of Health (NIH) Rare Disease Clinical Research Network (RDCRN), pathologists established a new scoring system to aid biopsy evaluation using digital biopsy

Accepted Manuscript
Version of Record at: https://doi.org/10.1093/jrsssc/qlad063

images. Using a list of structural features called morphology "descriptors," pathologists indicate whether each descriptor is present or absent (Barisoni et al., 2013). Given the novelty of the scoring system, the descriptors were first assessed for inter-pathologist agreement before testing their utility in differentiating disease diagnoses. These preliminary reproducibility studies found that some descriptors had very low prevalence (among the samples included in the studies) and that different existing agreement statistics yielded widely different results (Barisoni et al., 2016; Zee et al., 2018). To evaluate whether assumptions made by different existing agreement statistics were reasonable for our study, we collected additional data on pathologists' certainty in ratings. We found that pathologists demonstrated little uncertainty even with imperfect agreement (Zee et al., 2018). These unique data were not only useful in evaluating the suitability of existing agreement statistics, but also allowed for the development of a novel agreement statistic.

We propose a study design in which data on both the rating itself and uncertainty about the rating are collected for each item rated for a study with human raters. By asking raters which observations were rated with uncertainty, we can identify the observations that may be susceptible to chance agreement due to rater uncertainty. We propose a novel agreement statistic that utilizes these uncertainty data to accurately estimate chance-corrected agreement, with chance agreement defined as agreement on uncertain ratings.

The rest of this article is organized as follows. We first briefly review some of the most commonly used existing agreement statistics (Section 2). We then describe our theoretical framework, introduce our proposed agreement statistic and its standard error estimator, and provide practical considerations for how uncertainty data can be collected (Section 3). We conducted a series of simulation studies to test the performance of our proposed statistic in estimating the population parameter of interest, the design and results of which are described in Section 4. This is followed by an application of our proposed method to evaluate inter-pathologist agreement on NEPTUNE morphology descriptors (Section 5). Finally, we summarize our findings and describe situations where our proposed statistics are particularly useful (Section 6). This study received a waiver for Institutional Review Board approval by Ethical & Independent Review Services, #16034-01.

2. Existing Agreement Statistics

2.1. Non-chance-corrected Agreement Statistics

The proportion (or percent) of agreement, $\hat{P}_o$, is perhaps the simplest and most intuitive agreement statistic. However, it is often criticized for not taking chance agreement into account. If the two raters decided on ratings by independently flipping unbiased coins, the proportion agreement would equal 0.50 on average. This agreement estimate might be interpreted as moderate agreement even though it was generated purely by chance. Although some authors believe chance correction is unnecessary (Fleiss et al., 2003), this extreme case demonstrates the motivation to account for agreement by chance.

Several authors have advised caution when comparing agreement estimates across studies with different prevalences of positive ratings (Armitage et al., 1966; Grove et al., 1981; Spitznagel and Helzer, 1985; Uebersax, 1987). Citing the parallels between agreement eval-

Accepted Manuscript
Version of Record at: https://doi.org/10.1093/jrsssc/qlad063

uation and diagnostic test evaluation, one solution is to separately estimate the proportion of positive agreement as the proportion of concordant positive responses divided by the average (across raters) prevalence of positive responses, i.e., $\frac{\hat{p}_{11}}{(\hat{p}_{1\cdot}+\hat{p}_{\cdot 1})/2}$, and similarly for the proportion of negative agreement, $\frac{\hat{p}_{00}}{(\hat{p}_{0\cdot}+\hat{p}_{\cdot 0})/2}$ (Cicchetti and Feinstein, 1990). However, these two agreement measures are also not corrected for chance.

2.2. Chance-corrected Agreement Statistics

Cohen's κ (Cohen, 1960) is perhaps the most widely used chance-corrected agreement statistic. It is calculated for two raters and two response categories by

$$\kappa = \frac{\hat{P}_o - \hat{P}_\kappa}{1 - \hat{P}_\kappa},$$
$$\text{where } \hat{P}_\kappa = \hat{p}_{0.}\hat{p}_{.0} + \hat{p}_{1.}\hat{p}_{.1}$$

and $\hat{P}_o$ is the proportion agreement, as defined above. The probability of chance agreement is thus estimated by assuming raters are independent and summing the estimated probability that both rate 0 with the estimated probability that both rate 1. In other words, $\hat{P}_\kappa$ assumes ratings for all observations are based on a random biased coin flip, where the bias of the coin is based on the marginal proportions of observed ratings (Grove et al., 1981; Uebersax, 1987). Therefore, κ assumes that all observations are susceptible to chance agreement (Gwet, 2014). In fact, if both raters were merely independently flipping coins to decide on their ratings, then κ would appropriately estimate agreement to be 0. However, this extreme case is not realistic for most reproducibility studies. Furthermore, several authors have found that κ is sensitive to the prevalence of positive responses and to the bias (difference in prevalence) between raters (Spitznagel and Helzer, 1985; Feinstein and Cicchetti, 1990; Byrt et al., 1993; Gwet, 2014).

Scott's π (Scott, 1955) is calculated by

$$\pi = \frac{\hat{P}_o - \hat{P}_\pi}{1 - \hat{P}_\pi},$$
$$\text{where } \hat{P}_\pi = \left(\frac{\hat{p}_{0.} + \hat{p}_{.0}}{2}\right)^2 + \left(\frac{\hat{p}_{1.} + \hat{p}_{.1}}{2}\right)^2.$$

Scott's π is similar to Cohen's κ but makes the slightly more stringent assumption that the prevalences of positive responses are equal across the two raters (Cohen, 1960; Banerjee et al., 1999; Gwet, 2014). Therefore, if the marginal proportions of observed ratings are identical, then Cohen's κ and Scott's π will be equivalent. Krippendorff's α (Krippendorff,

Accepted Manuscript
Version of Record at: https://doi.org/10.1093/jrsssc/qlad063

1970) is similar to Scott's π:

$$\alpha = \frac{\hat{P}_o' - \hat{P}_\alpha}{1 - \hat{P}_\alpha},$$
$$\text{where } \hat{P}_o' = \left(1 - \frac{1}{2N}\right)\hat{P}_o + \frac{1}{2N},$$
$$\hat{P}_\alpha = \hat{P}_\pi = \left(\frac{\hat{p}_{0.} + \hat{p}_{.0}}{2}\right)^2 + \left(\frac{\hat{p}_{1.} + \hat{p}_{.1}}{2}\right)^2$$

except that α makes a correction for small samples, since $\hat{P}_o'$ is a weighted average of the proportion agreement and its maximum value of 1. Thus, with moderate sample size, α is similar in value to π.

Several agreement statistics have been independently proposed based on the assumption that all categories have equal probability of use by both coders, including Bennett's S (Bennett et al., 1954), Brennan-Prediger (Brennan and Prediger, 1981), and PABAK (Byrt et al., 1993). For two raters and two response categories, all three statistics are calculated by

$$PABAK = \frac{\hat{P}_o - 0.5}{1 - 0.5}$$
$$= 2\hat{P}_o - 1.$$

This statistic is not sensitive to prevalence or bias, since the marginal proportions of observations are not used in the agreement estimate calculation. However, it assumes that the probability of chance agreement is always 0.5.

Gwet developed the AC_1 statistic for two or more nominal response categories (Gwet, 2014). With two response categories,

$$AC_1 = \frac{\hat{P}_o - \hat{P}_{AC1}}{1 - \hat{P}_{AC1}}$$
$$\text{where } \hat{P}_{AC1} = (\hat{p}_{0.} + \hat{p}_{.0})\left(1 - \frac{\hat{p}_{0.} + \hat{p}_{.0}}{2}\right).$$

Unlike the previous statistics, Gwet's AC_1 assumes that some observations have deterministic ratings (i.e., not random or by chance) and other observations have nondeterministic ratings (i.e., rated with some randomness or chance) (Gwet, 2014). Therefore, only a proportion rather than all observations are susceptible to agreement by chance. AC_1 thus estimates the probability of chance agreement as the product of the estimated probability of agreement given the observation had a nondeterministic rating and the propensity for a rater to give a nondeterministic rating for that observation. AC_1 further assumes that if the observation was given a nondeterministic rating, then the rating is determined by an unbiased coin flip. Gwet argues that the probability of chance agreement should not exceed 0.5, and demonstrates that $\hat{P}_{AC1}$ can only take values between 0 and 0.5 whereas $\hat{P}_\kappa$ and $\hat{P}_\pi$ can both take values between 0.5 and 1 (Gwet, 2002).

The various assumptions made by existing agreement statistics can result in extremely different agreement estimates, especially given the sensitivity of several agreement statistics to prevalence. Such variability motivates the need to meticulously assess whether assumptions about chance agreement made by agreement statistics are valid for specific applications. Because these stringent assumptions may not be applicable to some situations, we propose a new agreement statistic that uses uncertainty data to empirically estimate and account for chance agreement due to rater uncertainty.

3. Novel Proposed Statistic

3.1. Theoretical Framework

Since agreement statistics are used when there is no gold standard or ground truth, we emphasize that we do not seek to assess "accuracy" or "correctness" of ratings compared to the ground truth. Instead, we seek to assess the extent to which different raters agree on their ratings. We assume each rater has an underlying, unobserved probability of each response, and it is both the sampling of observations from the population and the existence of chance, as defined by rater uncertainty, that causes the actual observed ratings to differ from the underlying population model. Agreement statistics based on observed ratings must therefore account for chance to accurately estimate the underlying population-based true agreement between raters.

We define a theoretical framework that assumes some observations to be rated are susceptible to chance agreement and some are not. Other authors have labeled these as "hard" and "easy" cases (Gwet, 2014) or used the terms "dishonest" and "honest" coding (Zhao et al., 2013). We contend that in most research studies, raters are inherently honest. We also recognize that a case that is difficult to rate by one rater may be easy to rate by another. Thus, we allow each rater to have ratings that are "certain" or "uncertain." As Grove (Grove et al., 1981) notes, having such information would allow certain cases to be treated separately from uncertain cases. We then define any observation rated with uncertainty by either rater to be susceptible to chance agreement. We note that uncertainty in a rating is a separate notion from error (compared to the unavailable ground truth), so a rater can theoretically have 100% certainty in a completely "incorrect" rating. As described above, the assessment of agreement concerns deviations from raters' underlying probabilities of responses due to uncertainty or chance and not deviations from the ground truth due to error.

To illustrate the framework more concretely, we classify the probabilities for each rating by each of two raters as in Table 2. Each cell of the table, $p_{ab}^{\mathcal{AB}}$, represents the probability that rater A gives response $a \in \{0, 1\}$, rater B gives response $b \in \{0, 1\}$, rater A has certainty (C) or uncertainty (U), $\mathcal{A} \in \{C, U\}$, and rater B has certainty or uncertainty, $\mathcal{B} \in \{C, U\}$. The margins show the rater-specific true probabilities of each response, or true rater prevalences of each response, where a dot represents all possible values of a, b, $\mathcal{A}$, or $\mathcal{B}$ by the other rater. We also note that adding values from corresponding cells can

Accepted Manuscript
Version of Record at: https://doi.org/10.1093/jrsssc/qlad063

			Rater B				
			Certain		Uncertain		
			0	1	0	1	Total
Rater A	Certain	0	p_{00}^{CC}	p_{01}^{CC}	p_{00}^{CU}	p_{01}^{CU}	$p_{0.}^{C.}$
		1	p_{10}^{CC}	p_{11}^{CC}	p_{10}^{CU}	p_{11}^{CU}	$p_{1.}^{C.}$
	Uncertain	0	p_{00}^{UC}	p_{01}^{UC}	p_{00}^{UU}	p_{01}^{UU}	$p_{0.}^{U.}$
		1	p_{10}^{UC}	p_{11}^{UC}	p_{10}^{UU}	p_{11}^{UU}	$p_{1.}^{U.}$
		Total	$p_{.0}^{.C}$	$p_{.1}^{.C}$	$p_{.0}^{.U}$	$p_{.1}^{.U}$	1

Table 2: Probability of each rating and uncertainty indication

collapse Table 2 to the simpler Table 1, i.e., $p_{ab} = \sum_{\mathcal{A},\mathcal{B}} p_{ab}^{\mathcal{AB}}$. Because any observation rated with uncertainty by either rater is considered susceptible to chance, the shaded cells in the table are those susceptible to chance.

A similar table is presented by Gwet (2008, 2014), where observations may be described as easy or hard rather than ratings being certain and uncertain. We therefore consider our theoretical framework to be different from that of Gwet's, since it separates observations by an attribute related to both the rater and the case rather than separating observations based only on attributes of the case itself. Gwet also assumes observations identified as easy by both raters can only lead to agreement, whereas we allow for raters to disagree even though both rated with certainty. In the NEPTUNE study of inter-rater agreement on morphology descriptors, for example, two pathologists who have been trained differently may have different interpretations of the same descriptor. Although each is absolutely certain in what he/she sees, they may disagree on whether the descriptor is present. Therefore, it is not chance that caused disagreement in this case but an underlying bias between the raters. An analagous example of disagreement with certainty is from the picture of a dress that became popular on the internet because some identified the dress as blue and black and others identified it as white and gold. Neuroscientists postulated that the large differences in color perception were caused by different experiences and adaptations of individuals' brains (Lafer-Sousa et al., 2015). Our framework therefore accommodates such situations where there can be certainty and disagreement.

Cohen defines κ as "the proportion of agreement *after* chance agreement is removed from consideration" (Cohen, 1960). Based on this concept, Gwet interpreted the population quantity of interest for the AC_1 statistic explicitly (Gwet, 2014). To define our population quantity of interest, we let γ represent a similar quantity as Gwet's using our notation and framework from Table 2. As Gwet established, the probability of agreement can be represented by a fraction with probability of agreement in the numerator and probability of agreement or disagreement in the denominator. The probability of chance agreement is $\sum_{i=0}^{1} \left(p_{ii}^{UU} + p_{ii}^{UC} + p_{ii}^{CU}\right)$, so we remove this quantity from both the numerator and

Accepted Manuscript
Version of Record at: https://doi.org/10.1093/jrsssc/qlad063

denominator of the fraction:

$$\gamma = \frac{\sum_{i=0}^{1} p_{ii}^{CC}}{1 - \sum_{i=0}^{1} \left(p_{ii}^{UU} + p_{ii}^{UC} + p_{ii}^{CU}\right)}$$
$$= \frac{\sum_{i=0}^{1} p_{ii}^{CC}}{\sum_{i=0}^{1} p_{ii}^{CC} + \sum_{j \neq i} \sum_{i=0}^{1} \sum_{\mathcal{A},\mathcal{B}} p_{ij}^{\mathcal{AB}}}.$$

γ is therefore the probability of agreement with certainty divided by the sum of the probability of agreement with certainty and the probability of disagreement regardless of certainty or uncertainty, i.e., the probability of agreement after chance agreement is removed from consideration. If we have observed information on certainty and uncertainty of ratings and the ratings themselves, then each probability in Table 2 and γ above could be empirically estimated by its corresponding proportion. Specifically, $p_{ab}^{\mathcal{AB}}$ can be replaced with $\hat{p}_{ab}^{\mathcal{AB}}$, where $\hat{p}_{ab}^{\mathcal{AB}} = n_{ab}^{\mathcal{AB}}/N$. This is the basis of our proposed statistic.

3.2. *Proposed Statistic*

Our proposed agreement statistic, ζ, aims to estimate the probability of agreement after chance agreement due to rater uncertainty is removed from consideration. It does so by empirically estimating the probability of chance agreement, $\hat{P}_\zeta$, as the proportion of observations in which the two raters agreed and at least one rater indicated uncertainty in their rating. This probability can also be expressed as a product of the probability of uncertainty and the probability of agreement given uncertainty; thus, $\hat{P}_\zeta$ is the product of the proportion of observations rated with uncertainty by either rater and the proportion of observations for which ratings agreed given that they were rated with uncertainty. Therefore, our proposed agreement statistic ζ can be calculated similarly to other existing agreement statistics by

$$\zeta = \frac{\hat{P}_o - \hat{P}_\zeta}{1 - \hat{P}_\zeta},$$

$$\text{where } \hat{P}_o = \hat{p}_{00} + \hat{p}_{11} = \sum_{i=0}^{1} \sum_{\mathcal{A},\mathcal{B}} \hat{p}_{ii}^{\mathcal{AB}}$$
$$= \hat{p}_{00}^{CC} + \hat{p}_{11}^{CC} + \hat{p}_{00}^{UC} + \hat{p}_{11}^{UC} + \hat{p}_{00}^{CU} + \hat{p}_{11}^{CU} + \hat{p}_{00}^{UU} + \hat{p}_{11}^{UU},$$

$$\text{and } \hat{P}_\zeta = \sum_{i=0}^{1} \sum_{j=0}^{1} \left(\hat{p}_{ij}^{CU} + \hat{p}_{ij}^{UC} + \hat{p}_{ij}^{UU}\right) \cdot \frac{\sum_{i=0}^{1} \sum_{\mathcal{A},\mathcal{B}} \hat{p}_{ii}^{\mathcal{AB}} - \sum_{i=0}^{1} \hat{p}_{ii}^{CC}}{\sum_{i=0}^{1} \sum_{j=0}^{1} \left(\hat{p}_{ij}^{CU} + \hat{p}_{ij}^{UC} + \hat{p}_{ij}^{UU}\right)}$$
$$= \sum_{i=0}^{1} \sum_{\mathcal{A},\mathcal{B}} \hat{p}_{ii}^{\mathcal{AB}} - \sum_{i=0}^{1} \hat{p}_{ii}^{CC}$$
$$= \hat{p}_{00}^{UC} + \hat{p}_{11}^{UC} + \hat{p}_{00}^{CU} + \hat{p}_{11}^{CU} + \hat{p}_{00}^{UU} + \hat{p}_{11}^{UU}.$$

Accepted Manuscript
Version of Record at: https://doi.org/10.1093/jrsssc/qlad063

Subtracting $\hat{P}_\zeta$ from $\hat{P}_o$ yields $\sum_{i=0}^{1} \hat{p}_{ii}^{CC}$, so $\hat{P}_\zeta$ must be less than or equal to $\hat{P}_o$. Therefore, like proportion agreement, ζ can only be non-negative. By algebraic manipulation, the statistic can also be rewritten as

$$\zeta = \frac{\sum_{i=0}^{1} n_{ii}^{CC}}{\sum_{i=0}^{1} n_{ii}^{CC} + \sum_{j \neq i} \sum_{i=0}^{1} \sum_{A,B} n_{ij}^{AB}}.$$

ζ is therefore an empirical estimator of γ.

3.3. Standard Error

To derive a standard error for ζ, we re-write ζ as

$$\zeta = \frac{\hat{p}_X}{\hat{p}_Y} = \frac{\left(\sum_{i=0}^{1} n_{ii}^{CC}\right)/N}{\left(\sum_{i=0}^{1} n_{ii}^{CC} + \sum_{j \neq i} \sum_{i=0}^{1} \sum_{\mathcal{A},\mathcal{B}} n_{ij}^{\mathcal{AB}}\right)/N},$$

a ratio of two binomial proportions, $\hat{p}_X$ and $\hat{p}_Y$, both with denominator N. Similarly, we re-write $\gamma = \frac{p_X}{p_Y}$, where $p_X < p_Y$ always. Let each of the binomial proportions be expressed as a sum of N Bernoulli random variables, X_k and Y_k, respectively, where k represents an individual observation. Therefore, $X_k = 1$ when observation k is rated with certainty by both raters and the raters agree, and $Y_k = 1$ when $X_k = 1$ or when raters disagree (regardless of certainty). Then

$$\mathrm{E}\begin{bmatrix} X_k \\ Y_k \end{bmatrix} = \begin{bmatrix} p_X \\ p_Y \end{bmatrix}$$
$$\mathrm{Var}\begin{bmatrix} X_k \\ Y_k \end{bmatrix} = M = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}$$

where $\sigma_X^2 = p_X(1-p_X)$, $\sigma_Y^2 = p_Y(1-p_Y)$ and $\sigma_{XY} = \mathrm{E}(X_kY_k) - \mathrm{E}(X_k)\mathrm{E}(Y_k)$. Since $X_k = 1$ implies $Y_k = 1$, then $X_kY_k = X_k$, so $\sigma_{XY} = \mathrm{E}(X_k) - \mathrm{E}(X_k)\mathrm{E}(Y_k) = p_X - p_Xp_Y = p_X(1-p_Y)$.

By the Law of Large Numbers, $\hat{p}_X \xrightarrow{p} p_X$ and $\hat{p}_Y \xrightarrow{p} p_Y$. By the multivariate Central Limit Theorem and the fact that $\bar{X} = \hat{p}_X$ and $\bar{Y} = \hat{p}_Y$,

$$\sqrt{N}\left(\begin{bmatrix} \hat{p}_X \\ \hat{p}_Y \end{bmatrix} - \begin{bmatrix} p_X \\ p_Y \end{bmatrix}\right) \xrightarrow{D} N(0, M).$$

By the Continuous Mapping Theorem, $\log\left(\frac{\hat{p}_X}{\hat{p}_Y}\right) \xrightarrow{p} \log\left(\frac{p_X}{p_Y}\right)$. Then using the multivariate

Delta Method,

$$\sqrt{N}\left([\log(\hat{p}_X)-\log(\hat{p}_Y)]-[\log(p_X)-\log(p_Y)]\right)\xrightarrow{D}N(0,\tilde{M})$$

$$\begin{aligned}
\text{where } \tilde{M} &= \begin{pmatrix} \frac{1}{p_X} & -\frac{1}{p_Y} \end{pmatrix} M \begin{pmatrix} \frac{1}{p_X} & -\frac{1}{p_Y} \end{pmatrix}' \\
&= \begin{pmatrix} \frac{\sigma_X^2}{p_X}-\frac{\sigma_{XY}}{p_Y} & \frac{\sigma_{XY}}{p_X}-\frac{\sigma_Y^2}{p_Y} \end{pmatrix}\begin{pmatrix} \frac{1}{p_X} \\ -\frac{1}{p_Y} \end{pmatrix} \\
&= \frac{\sigma_X^2}{p_X^2}+\frac{\sigma_Y^2}{p_Y^2}-\frac{2\sigma_{XY}}{p_Xp_Y} \\
&= \frac{p_X(1-p_X)}{p_X^2}+\frac{p_Y(1-p_Y)}{p_Y^2}-\frac{2p_X(1-p_Y)}{p_Xp_Y} \\
&= \frac{1-p_X}{p_X}-\frac{1-p_Y}{p_Y}.
\end{aligned}$$

Therefore, using similar arguments as in Katz et al. (1978), $\log(\zeta)=\log(\hat{p}_X/\hat{p}_Y)=\log(\hat{p}_X)-\log(\hat{p}_Y)$ is approximately normally distributed with mean $\mu=\log(\gamma)$ and asymptotic variance

$$\sigma^2=\frac{1-p_X}{Np_X}-\frac{1-p_Y}{Np_Y}.$$

Finally, since $\log(\zeta)$ is approximately normally distributed, ζ is approximately log-normally distributed with estimated standard error

$$\widehat{\text{SE}}(\zeta)=\sqrt{(\exp(\sigma^2)-1)\exp(2\mu+\sigma^2)}.$$

In practice, γ, p_X, and p_Y can be estimated with ζ, $\hat{p}_X$, and $\hat{p}_Y$, respectively.

3.4. Collection of Uncertainty Data

To assess certainty, we can ask raters to indicate whether each of their ratings was made with certainty or uncertainty. To minimize the extra time that such an indication would require, a default can be set *a priori*. For example, all observations can be assumed to be rated with certainty unless the rater checks a box indicating that they were uncertain. Alternatively, all observations can be assumed to be rated with uncertainty unless indicated otherwise. Another way to think about certainty is to ask raters about their confidence in giving the same rating for an identical observation at a later time. This method uses the idea of intra-rater agreement without conducting a formal intra-rater agreement study.

The decision to collect these additional uncertainty data should be made when the study protocol is designed so raters can indicate uncertainty at the time of rating each observation. Asking raters to indicate uncertainty after all observations have been rated may not be feasible, and the uncertainty data collected in this way may suffer from recall bias. We also emphasize that our proposed methods are suitable for studies with human raters where such uncertainty data can be ascertained, whereas many rater reliability studies involving non-human raters would not be able to provide such data.

Accepted Manuscript
Version of Record at: https://doi.org/10.1093/jrsssc/qlad063

4. Simulation Studies

4.1. Simulation Setup

We tested the performance of our proposed statistic in estimating the population quantity of interest γ defined in Section 3.1 using a series of simulation studies assuming two raters and two possible responses. The true rater prevalence is defined as the rater-specific true probability of a positive rating (not the ground truth prevalence, since there is no gold standard or ground truth). We first generated data assuming that rater A and rater B both rated with certainty. We varied the true rater prevalence for rater B when rated with certainty, $p_{.1}^{\cdot C} \in \{0.1, 0.2, \cdots, 0.9\}$ and varied the bias between rater B and rater A, $p_{.1}^{\cdot C} - p_{1.}^{C\cdot} \in \{0, 0.1, 0.2, 0.3\}$, which determined the true rater prevalence for rater A when rated with certainty, $p_{1.}^{C\cdot}$.

Given $p_{.1}^{\cdot C}$ and $p_{1.}^{C\cdot}$, we first calculated all possible values of p_{11}^{CC} based on the following constraints:

$$
\begin{aligned}
p_{11}^{CC} &\leq p_{.1}^{\cdot C} \\
p_{11}^{CC} &\leq p_{1.}^{C\cdot} \\
p_{11}^{CC} &\geq p_{1.}^{C\cdot} + p_{.1}^{\cdot C} - 1 \\
0 \leq p_{11}^{CC} &\leq 1
\end{aligned}
$$

or equivalently,

$$\max\{p_{1.}^{C\cdot} + p_{.1}^{\cdot C} - 1, 0\} \leq p_{11}^{CC} \leq \min\{p_{.1}^{\cdot C}, p_{1.}^{C\cdot}\}$$

To conduct a finite number of simulations, we selected possible values of $p_{11}^{C}C$ in this range in 0.1 increments. Then based on each value of p_{11}^{CC}, we calculated

$$
\begin{aligned}
p_{10}^{CC} &= p_{1.}^{CC} - p_{11}^{CC} \\
p_{01}^{CC} &= p_{10}^{CC} + p_{.1}^{\cdot C} - p_{1.}^{C\cdot} \\
p_{00}^{CC} &= 1 - (p_{10}^{CC} + p_{01}^{CC} + p_{11}^{CC}).
\end{aligned}
$$

The true amount of agreement under certainty was calculated as $p_{00}^{CC} + p_{11}^{CC}$. Ratings for raters A and B were then randomly sampled from a multinomial distribution with

$$P(a, b|\mathcal{A} = \mathcal{B} = C) = p_{ab}^{CC}.$$

All possible combinations of prevalence, bias, and true agreement were simulated, i.e., those where all probabilities p_{ab} were between 0 and 1.

We then varied the probability that observations were rated with uncertainty by each rater, first assuming equal uncertainty across raters, $\psi = P(\mathcal{A} = U) = P(\mathcal{B} = U) \in \{0\%, 20\%, 50\%, 80\%, 100\%\}$. Uncertainty for each observation was then sampled from a Bernoulli distribution with probability ψ. Ratings for observations with uncertainty were either randomly sampled from a Bernoulli distribution with

$$P(a|\mathcal{A} = U) = P(b|\mathcal{B} = U) = 0.5$$

Accepted Manuscript
Version of Record at: https://doi.org/10.1093/jrsssc/qlad063

to mimic an unbiased coin flip or randomly sampled from a Bernoulli distribution with

$$P(a|\mathcal{A} = U) = p^{C\cdot}_{1\cdot}$$
$$P(b|\mathcal{B} = U) = p^{\cdot C}_{\cdot 1}$$

to mimic a biased coin flip.

The number of observations to be rated was also varied, with $N \in \{25, 50, 100, 200, 500\}$, and each set of simulation parameters was repeated $M = 10,000$ times, each time with different draws. For each simulation repetition, we calculated agreement estimates using our proposed statistic ζ and compared them to the population quantity of interest γ. We also estimated proportion of agreement, Cohen's κ, Gwet's AC_1, Scott's π, PABAK, and Krippendorff's α for comparison of magnitudes and properties, since these other statistics are based on different definitions of chance agreement. For our proposed agreement statistic ζ, the simulated ratings and the uncertainty indications for each rating were used for calculation (i.e., by filling in the corresponding cells in the 4×4 Table 2). Because existing statistics do not use uncertainty data in their calculations, these statistics were estimated using only the simulated ratings (i.e., by collapsing the 4×4 table to the 2×2 Table 1). R code used to implement these simulations is included in Supplemental Section A.1. Agreement estimates from Scott's π and Krippendorff's α were close to those of Cohen's κ in almost all simulations, so the former two are not presented in simulation results and any differences are described below. Agreement estimates were averaged across simulation repetitions for display in figures and tables.

4.2. Simulation Results with Unbiased Coin Flips for Uncertain Ratings

Figure 1 shows simulation results with 0 bias between raters, unbiased coin flips for uncertain ratings, $N = 200$, and true agreement under certainty of 0.80. We observed similar results at different levels of true agreement under certainty. Estimates from all existing agreement statistics, including proportion of agreement and chance-corrected statistics, generally decreased as the proportion of rater uncertainty increased from 0% (Figure 1, leftmost panel) to 100% (Figure 1, rightmost panel). Even though proportion of agreement, Cohen's κ, and PABAK are not defined by and do not account for uncertainty as our proposed statistic does or as Gwet's AC_1 accounts for difficulty of cases, the fact that they change with the uncertainty level reflects the extent to which their assumptions regarding chance agreement fit the underlying data and the amount of rater uncertainty. Gwet's AC_1 statistic demonstrated a concave up curve, with higher agreement as the prevalence approached 0 or 1. Cohen's κ had a concave down curve, with lower agreement as the prevalence approached 0 or 1. As expected, PABAK was flat across different prevalences and always in between Gwet's AC_1 and Cohen's κ. Our proposed statistic ζ was unbiased in estimating γ (true agreement after removing chance agreement) and had a flat curve, showing less sensitivity to prevalence than Gwet's AC_1 or Cohen's κ. ζ decreased as the amount of uncertainty increased but always remained unbiased in estimating γ.

In the extreme case where all observations were rated with uncertainty, all of the existing chance-corrected agreement statistics and our proposed statistic ζ estimated agreement

Accepted Manuscript
Version of Record at: https://doi.org/10.1093/jrsssc/qlad063

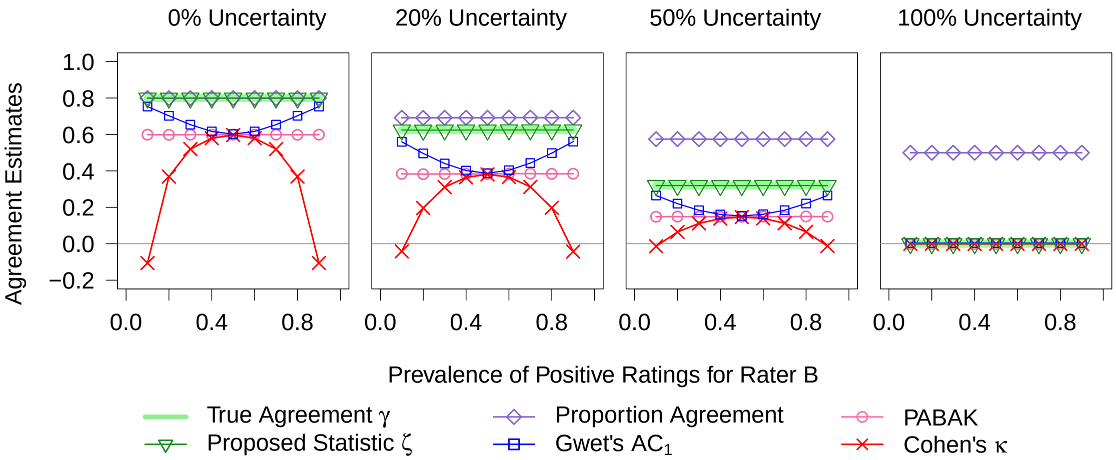

Fig. 1. Agreement estimates by uncertainty of both raters for 0.80 true agreement

to be 0, which is also the value of γ in this situation. In other words, all of the chance-corrected agreement statistics removed all chance agreement and estimated true agreement to be 0. This property is often desirable, since the agreement statistics are able to account for uncertainty in their estimates by giving lower values with increasing uncertainty. In contrast, proportion of agreement was 0.5 with 100% uncertainty, showing moderate agreement even though ratings were purely by chance. We also conducted simulations that allowed the probability of uncertainty to be different between the two raters, e.g., 0.2 for rater A and 0.5 for rater B. Results were similar, with agreement estimates of existing statistics and of our proposed statistic ζ falling between the agreement estimates when both raters had probability of uncertainty 0.2 and both raters had probability of uncertainty 0.5.

When positive bias between rater B and rater A was introduced, results were similar, but all curves were shifted to the right. An example comparing no bias between raters to bias of 0.2 between raters, with true agreement under certainty of 0.80, 50% rater uncertainty, unbiased coin flips for uncertain ratings, and $N = 200$ is shown in Supplemental Figure S1. Having positive bias between raters was also the one situation where Scott's π and Krippendorff's α were not as close to Cohen's κ as they were when there was no bias between raters. Across simulations with bias of 0.2 between raters, differences between Cohen's κ and Scott's π ranged from 0 to 0.11, and differences between Cohen's κ and Krippendorff's α ranged from 0 to 0.08. Results show that agreement statistics that have flat relationships with prevalence are unchanged with bias between raters, as for our proposed statistic ζ, the proportion agreement, and PABAK. For other existing agreement statistics that have curved relationships with prevalence, the curves were shifted right by 0.2 when bias between raters was increased from 0 to 0.2.

4.3. Simulation Results showing Standard Error Estimates over Different Sample Sizes

We compared the precision of agreement estimates for different sample sizes. For all agreement statistics, we calculated empirical standard errors (standard deviations of agreement estimates over simulation repetitions) and separately estimated asymptotic standard errors in each simulation repetition (then averaged across simulation repetitions). For existing agreement statistics, we calculated estimated standard errors using established asymptotic expressions, i.e., the normal approximation of binomial proportions for proportion agreement and PABAK, Fleiss et al.'s standard error estimate for Cohen's κ (Fleiss et al., 1969; Gwet, 2014), and Gwet's standard error expressions for all others (Gwet, 2008, 2014). For our proposed agreement statistic, we estimated standard errors as described in Section 3.3 and also used those standard errors to calculate 95% confidence interval coverage using a standard normal approximation, $\zeta \pm z_{0.975} \cdot \widehat{\mathrm{SE}}(\zeta)$.

Table 3 shows empirical standard errors (standard deviations of agreement estimates across simulation repetitions) and average estimated standard errors using asymptotic variance expressions of agreement estimates with prevalence 0.5, 0 bias, true agreement under certainty of 0.80, 50% rater uncertainty, and unbiased coin flips for uncertain ratings. These parameters resulted in true agreement after removing chance agreement of $\gamma = 0.32$. We first observe that our proposed statistic ζ is close to γ for all sample sizes. We also see that all estimated asymptotic standard errors are close to the empirical standard errors, especially as N becomes large. For our proposed agreement statistic, this result confirms the accuracy of the asymptotic standard error expression developed in Section 3.3. As expected, standard errors decrease as N increases. Proportion agreement had the smallest estimated standard error overall, followed by the proposed statistic ζ. 95% confidence interval coverage for our proposed statistic was 96.8% for $N = 25$, 94.4% for $N = 50$, and 94.6% for $N \geq 100$.

4.4. Simulation Results with Biased Coin Flips for Uncertain Ratings

For biased coin flips, ratings for observations rated with uncertainty were simulated based on the probability of a positive rating for each rater. Simulation results with 0 bias between raters, $N = 200$, true agreement under certainty of 0.80, and biased coin flips for uncertain ratings are shown in Figure 2. Agreement estimates from Cohen's κ were similar to those when using unbiased coin flips. Notably, agreement estimates for our proposed statistic ζ and proportion agreement show a curve rather than a straight line over different values of prevalance, demonstrating some sensitivity to prevalence. Unlike proportion agreement, however, our proposed statistic ζ still retains its property of accounting for uncertainty by giving lower agreement estimates as uncertainty increases. Gwet's AC_1 statistic also demonstrated a much steeper curve, but also often had non-zero values even with 100% uncertainty. This may be due to the fact that Gwet's AC_1 explicitly assumes unbiased coin flips only for hard ratings (Gwet, 2014). PABAK had a steep curve with biased coin flips and closely resembled Gwet's AC_1, with values just slightly lower or equal to Gwet's AC_1.

Accepted Manuscript
Version of Record at: https://doi.org/10.1093/jrsssc/qlad063

N	Statistic	Mean of Agreement Estimates	Empirical Standard Error	Mean of Asymptotic Standard Errors
25	Proportion Agreement	0.578	0.100	0.097
	Gwet's AC_1	0.173	0.200	0.197
	PABAK	0.155	0.200	0.193
	Cohen's κ	0.151	0.198	0.201
	Proposed Statistic ζ	0.322	0.122	0.131
50	Proportion Agreement	0.573	0.069	0.069
	Gwet's AC_1	0.156	0.139	0.140
	PABAK	0.146	0.139	0.138
	Cohen's κ	0.144	0.138	0.141
	Proposed Statistic ζ	0.319	0.084	0.087
100	Proportion Agreement	0.574	0.049	0.049
	Gwet's AC_1	0.152	0.098	0.099
	PABAK	0.148	0.098	0.098
	Cohen's κ	0.147	0.098	0.099
	Proposed Statistic ζ	0.319	0.059	0.060
200	Proportion Agreement	0.575	0.035	0.035
	Gwet's AC_1	0.152	0.071	0.070
	PABAK	0.150	0.071	0.070
	Cohen's κ	0.149	0.071	0.070
	Proposed Statistic ζ	0.320	0.042	0.042
500	Proportion Agreement	0.575	0.022	0.022
	Gwet's AC_1	0.151	0.045	0.044
	PABAK	0.150	0.045	0.044
	Cohen's κ	0.150	0.045	0.044
	Proposed Statistic ζ	0.320	0.027	0.026

Table 3: Simulation results of agreement estimates and empirical and estimated standard errors.

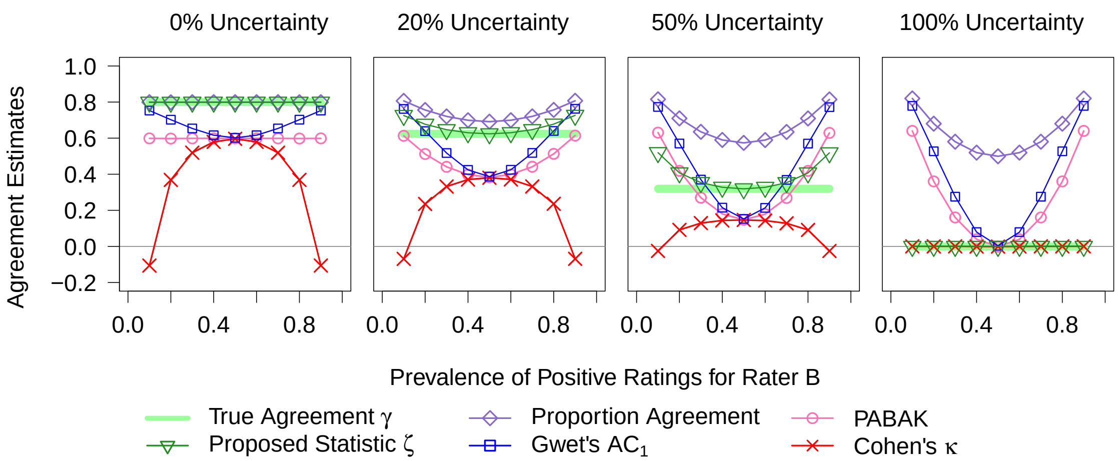

Fig. 2. Agreement estimates with biased coin flips for uncertain ratings

4.5. Simulation Results with Inacccurate Uncertainty Indications

Our proposed statistic is advantageous because it does not make strong assumptions about the probability of chance agreement, instead using uncertainty data to account for chance agreement. However, one might suspect that uncertainty indications may also be inaccurate, whereas our proposed statistic relies on the assumption that uncertainty data are accurate. We believe this is a plausible assumption, because we assume that raters are inherently rating with maximal effort and rigor. While ratings may be susceptible to chance due to reasons outside of the rater's control (e.g. blurry images, lack of previous experience, or having to dichotomize a continuous measure close to the cutpoint), we place trust in the rater's ability to assess their own certainty.

Still, we conducted additional simulations to test the robustness of our proposed agreement statistic when indications of uncertainty were inaccurate. Although we expect inaccurate uncertainty of 10% or less in most situations, we present simulations where we randomly selected 0, 10, 30, and 50 percent of certainty or uncertainty indications to be wrong. Figure 3 shows results from these simulations for true agreement under certainty of 0.8, $N = 200$, unbiased coin flips for uncertain ratings, and probability of uncertainty for each rater of 0.2, 0.5, and 0.8.

When the probability of uncertainty was low (Figure 3, top row), ζ decreased as the percent of inaccurate indications of uncertainty increased. With 30% or less inaccurate uncertainty indications, ζ still had little bias in estimating γ. With 50% inaccurate uncertainty indications, however, the bias was larger. When the probability of uncertainty was high, ζ increased slightly as the percent of inaccurate indications of uncertainty increased. When the probability of uncertainty was 0.5, ζ had little bias for estimating γ even with 50% inaccurate uncertainty indications. The behavior of ζ is expected, since

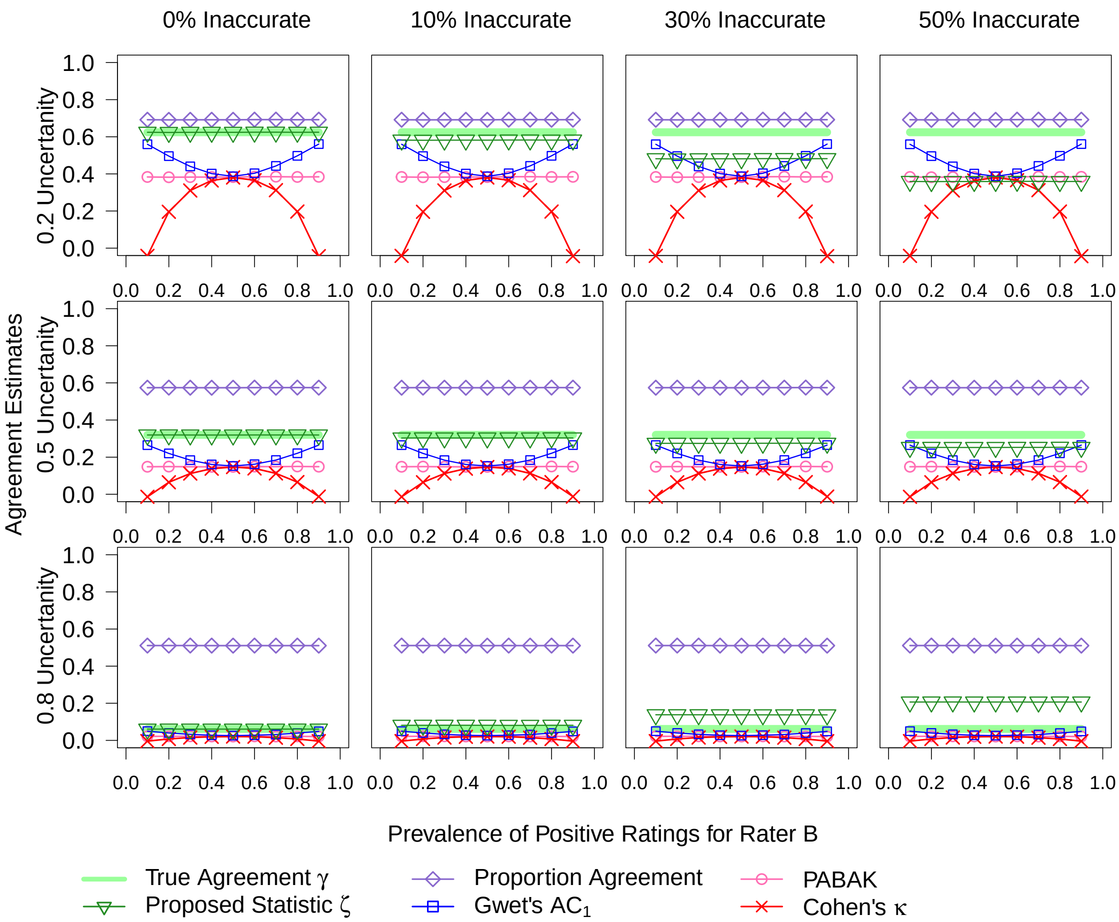

Fig. 3. Agreement estimates with inaccurate indications of uncertainty

the chance-correction in ζ contains the proportion of observations rated with uncertainty. When the proportion of uncertainty is indicated to be high (although it is truly low at 0.2), ζ will correct more for chance. When the proportion of uncertainty is indicated to be low (although it is truly high at 0.8), ζ will correct less for chance.

5. Data Application in the NEPTUNE Study

NEPTUNE is a multicenter prospective cohort study of children and adults with nephrotic syndrome, a rare type of kidney disease diagnosed by evaluation of a kidney biopsy (Gadegbeku et al., 2013). Kidney biopsy tissue is collected according to the NEPTUNE Digital Pathology Protocol (NDPP), including digital electron microscopy (EM) images (Barisoni et al., 2013). The NDPP includes a scoring system for evaluating biopsies, in which pathologists rate descriptors representing structural lesions. Although some of these descriptors are used in conventional diagnoses for nephrotic syndrome, most have never before been collected in such a quantitative manner. Therefore, it was important to identify the set of reproducible descriptors to bring to the next stage of validation.

For the current study, we analyzed pathologists' ratings of six descriptors from assessment of EM images of kidney tissue from $N = 196$ NEPTUNE patients and one descriptor from assessment of whole slide images from $N = 157$ NEPTUNE patients. Two pathologists rated each descriptor as present or absent and additionally used a check box to indicate whether their rating was made with any uncertainty. The first pathologist indicated uncertainty in the range of 0.3% to 12.7% of observations across the six EM descriptors, whereas the second pathologist indicated uncertainty in the range of 0.0% to 16.6% of observations. Pathologists attributed some differences in uncertainty to different experiences, e.g., different interpretations of image quality based on images usually seen in clinical practice by different pathologists.

We calculated the inter-pathologist agreement for each descriptor using proportion agreement, Cohen's κ, Gwet's AC_1, PABAK, and our proposed statistic ζ. Agreement estimates for the seven descriptors with prevalences between 11.8% and 73.8% are plotted in Figure 4, along with their standard errors. After removing chance agreement (i.e., agreement due to uncertainty), inter-rater agreement (standard error) using our proposed statistic ζ ranged between 0.58 (0.04) for Sinechia and 0.82 (0.02) for Electron Densities/Hyaline Material. Proportion agreement yielded the highest agreement estimates, followed by ζ. For descriptors with a low level of uncertainty in ratings, proportion agreement and ζ were very similar. The descriptor with the highest uncertainty by both raters was Sinechia, which also had the largest difference between proportion agreement and ζ. The large corrections for chance agreement implemented by other existing statistics resulted in agreement estimates that were lower.

As described in Section 3.4, it is ideal if uncertainty data can be collected at the same time that the ratings are made, as in the NEPTUNE data example. If uncertainty indications for each observation were not collected when ratings were made, or if only summarized uncertainty data are available, simulations may be useful to approximate ζ. An additional data example illustrating this idea is provided in Supplemental Section A.3.

Accepted Manuscript
Version of Record at: https://doi.org/10.1093/jrsssc/qlad063

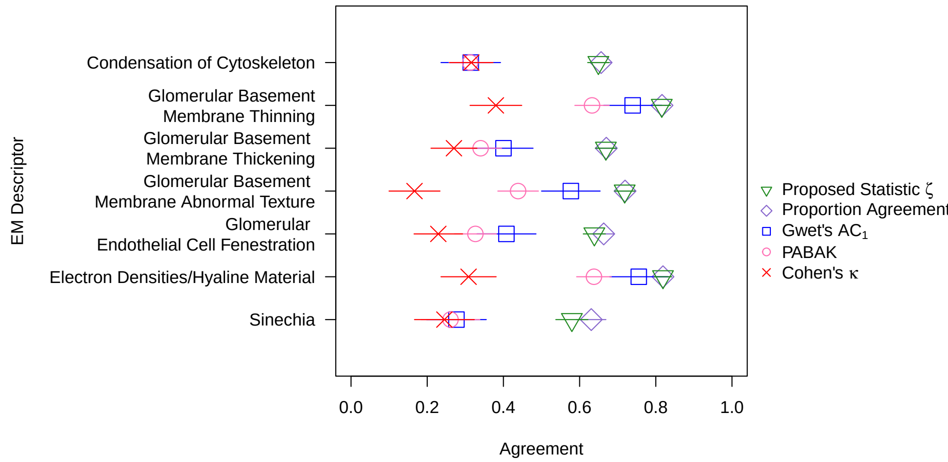

Fig. 4. Agreement estimates $\pm$ 1 standard error for seven pathology descriptors identified in kidney biopsies from the Nephrotic Syndrome Study Network (NEPTUNE)

6. Conclusion

Given the wide variability in estimates from existing agreement statistics, it can be challenging for researchers to know which agreement statistic is appropriate for their data. Many default to Cohen's κ, but its strong penalty for chance agreement from assuming all observations are susceptible to chance may be too stringent. Other existing statistics are less conservative but still make assumptions for estimating chance agreement that may not hold in many situations, such as assuming an unbiased coin flip determines ratings for observations rated with chance. In response to these challenges, we propose that researchers collect rater-reported data on uncertainty and use our novel agreement statistic, ζ, to estimate agreement corrected for chance due to rater uncertainty.

Our proposed statistic ζ can accommodate a wide range of assumptions about uncertainty—both in terms of which observations are rated by chance and the probability that they are rated by chance. It possesses the desirable ability to account for uncertainty in its estimate by decreasing in value as the amount of uncertainty increases. With these additional uncertainty data, we can separate agreement (disagreement) between raters due to chance from agreement (disagreement) between raters due to true agreement (disagreement). Further, ζ has little bias and little sensitivity to prevalence for estimating the probability of agreement after removing chance agreement.

With very low levels of uncertainty, one could argue that proportion of agreement is sufficient to estimate agreement without having to correct for chance agreement at all. However, we must first understand the level of rater uncertainty to justify this approach and therefore must collect the uncertainty data. Calculation of ζ can then be considered, along with any existing agreement statistic. Another potential strategy to use uncertainty

Accepted Manuscript
Version of Record at: https://doi.org/10.1093/jrsssc/qlad063

data would be to use proportion agreement to estimate agreement only among cases rated with certainty by both raters. Among such cases with no uncertainty, there is no chance agreement and therefore no need to correct for chance. However, the resulting proportion agreement estimate should be interpreted only among cases rated with certainty. Only if the cases rated with certainty are sampled completely at random from the entire population could this agreement estimate be broadly generalizable.

There are two limitations of our proposed statistic worth noting. First, the calculation of ζ requires additional data collection and thus more resources. However, we estimate the additional burden to be small relative to potentially large benefits of being able to calculate this chance-corrected agreement statistic. Second, we assume that raters provide reasonably accurate indications of uncertainty. We believe this is a plausible assumption, since research-based agreement studies are generally conducted with experienced raters following rigorous scientific principles. To minimize the risk of inaccurate indications of uncertainty, we recommend well-designed protocols for rating and assessing uncertainty and adequate training to carry out the protocol design. Our simulation study results also showed robustness of our proposed statistic when these inaccuracies were low or moderate. Specifically, ζ had little bias in estimating the probability of agreement after removing chance agreement with modest (30% or less) inaccurate indications of uncertainty. However, when the true amount of uncertainty is thought to be close to 0 or 1 and the amount of inaccurate indications of uncertainty are thought to be 50% or more, ζ may not be appropriate.

It is also important to note that there are studies in which more conservative agreement estimates would be preferred, e.g., when the goal is to flag poorly performing raters. Alternatively, established agreement statistics may be preferred when it is important to compare to previous studies. On the other hand, e.g., in our NEPTUNE data example in which the goal is to choose reproducible descriptors for further validation, we want to be very cautious about removing useful descriptors and therefore want a more liberal agreement statistic. ζ being insensitive to prevalence also has the benefit of being comparable across studies with different prevalence levels. Such studies should have similar uncertainty levels for appropriate comparison, although differing uncertainties could uncover important information about the raters' processes and/or interpretation of observations. In some cases, it may be useful to calculate multiple agreement statistics and discuss the differences between the estimates.

Currently, our proposed method is applicable to agreement studies with two raters, two response categories, and a dichotomous indication of uncertainty. Natural extensions of our method would allow for more than two raters and ordinal or more than two nominal response categories. To accommodate ordinal responses, we could follow strategies similar to those used for existing statistics (Cohen, 1968; Gwet, 2014), i.e., incorporating weights that are highest when the raters agree perfectly and that decrease as the difference between the two chosen ratings increases. Both the proportion agreement and the estimated probability of chance agreement would be calculated using a weighted sum of all observed proportions. Similarly, uncertainty may sometimes exist on an ordinal or continuous scale rather than dichotomous. If we apply our proposed statistic ζ by dichotomizing uncertainty in this case, ζ may be conservative in its correction for chance agreement since even small degrees of uncertainty are assumed to lead to chance ratings. An ordinal scale of uncertainty could therefore improve agreement estimates in these cases and could be incorporated by weighting ratings for each observation by the degree of uncertainty.

As new biological markers for nephrotic syndrome and other chronic conditions become available, it is crucial to evaluate their reproducibility. Our proposed statistic can provide accurate assessments of rater agreement by accounting for rater uncertainty. We thus recommend that future inter-rater reproducibility studies also collect uncertainty data for agreement estimation.

Acknowledgments

We thank Regina Baker for discussions on the assumptions made by existing agreement statistics and John Gillespie for insightful comments on the manuscript. We also thank Stephen Hewitt, Jeffrey Hodgin, Matthew Palmer, and Joseph Gaut for the collection of descriptor scores and uncertainty data in the NEPTUNE study and Matthias Kretzler for clinical and operational support in using the NEPTUNE data for this project. We thank Gisele Neiva, Dennis Fasbinder, Ronald Heys, and Donald Heys for the dental study data and uncertainty estimates. This project was funded in part by a NEPTUNE Career Development Fellowship award to Dr. Zee (NEPTUNE grant U54DK083912).

Data Availability

The data in this study were obtained from the Nephrotic Syndrome Study Network (NEPTUNE) where data sharing requires ancillary study approval and a data use agreement. The dataset may be requested from the NEPTUNE Data Analysis and Coordinating Center (DACC), `https://www.neptune-study.org/`.

Conflict of Interest

The authors declare that there is no conflict of interest.

Funding

The Nephrotic Syndrome Study Network Consortium (NEPTUNE), U54-DK-083912, is a part of the National Institutes of Health (NIH) Rare Disease Clinical Research Network (RDCRN), supported through a collaboration between the Office of Rare Diseases Research (ORDR), NCATS, and the National Institute of Diabetes, Digestive, and Kidney Diseases. Additional funding and/or programmatic support for this project has also been provided by the University of Michigan, the NephCure Kidney International, the Halpin Foundation, and Arbor Research Collaborative for Health.

References

Armitage, P., Blendis, L. M. and Smyllie, H. C. (1966) The measurement of observer disagreement in the recording of signs. *Journal of the Royal Statistical Society. Series A (General)*, **129**, 98–109.

Banerjee, M., Capozzoli, M., McSweeney, L. and Sinha, D. (1999) Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, **27**, 3–23.

Barisoni, L., Nast, C. C., Jennette, J. C., Hodgin, J. B., Herzenberg, A. M., Lemley, K. V., Conway, C. M., Kopp, J. B., Kretzler, M., Lienczewski, C., Avila-Casado, C., Bagnasco, S., Sethi, S., Tomaszewski, J., Gasim, A. H. and Hewitt, S. M. (2013) Digital pathology evaluation in the multicenter nephrotic syndrome study network (neptune). *Clin J Am Soc Nephrol*, **8**, 1449–1459.

Barisoni, L., Troost, J. P., Nast, C., Bagnasco, S., Avila-Casado, C., Hodgin, J., Palmer, M., Rosenberg, A., Gasim, A., Liensziewski, C., Merlino, L., Chien, H.-P., Chang, A., Meehan, S. M., Gaut, J., Song, P., Holzman, L., Gibson, D., Kretzler, M., Gillespie, B. W. and Hewitt, S. M. (2016) Reproducibility of the neptune descriptor-based scoring system on whole-slide images and histologic and ultrastructural digital images. *Mod Pathol*, **29**, 671–684.

Bennett, E. M., Alpert, R. and Goldstein, A. C. (1954) Communications through limited-response questioning. *Public Opinion Quarterly*, **18**, 303–308.

Brennan, R. L. and Prediger, D. J. (1981) Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, **41**, 687–699.

Byrt, T., Bishop, J. and Carlin, J. B. (1993) Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, **46**, 423–429.

Cicchetti, D. V. and Feinstein, A. R. (1990) High agreement but low kappa: resolving the paradoxes. *Journal of Clinical Epidemiology*, **43**, 551–558.

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

— (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*, **70**, 213–220.

Feinstein, A. R. and Cicchetti, D. V. (1990) High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, **43**, 543–549.

Fleiss, J. L., Cohen, J. and Everitt, B. S. (1969) Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, **72**, 323–327.

Fleiss, J. L., Levin, B. and Paik, M. C. (2003) *The Measurement of Interrater Agreement*, book section Ch. 18, 598–626. John Wiley & Sons, Inc.

Gadegbeku, C. A., Gipson, D. S., Holzman, L. B., Ojo, A. O., Song, P. X., Barisoni, L., Sampson, M. G., Kopp, J. B., Lemley, K. V., Nelson, P. J., Lienczewski, C. C., Adler, S. G., Appel, G. B., Cattran, D. C., Choi, M. J., Contreras, G., Dell, K. M., Fervenza, F. C., Gibson, K. L., Greenbaum, L. A., Hernandez, J. D., Hewitt, S. M., Hingorani, S. R., Hladunewich, M., Hogan, M. C., Hogan, S. L., Kaskel, F. J., Lieske, J. C., Meyers, K. E., Nachman, P. H., Nast, C. C., Neu, A. M., Reich, H. N., Sedor, J. R., Sethna, C. B., Trachtman, H., Tuttle, K. R., Zhdanova, O., Zilleruelo, G. E. and Kretzler, M. (2013) Design of the nephrotic syndrome study network (neptune) to evaluate primary glomerular nephropathy by a multidisciplinary approach. *Kidney International*, **83**, 749–756.

Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B. and Shapiro, R. W. (1981) Reliability studies of psychiatric diagnosis: Theory and practice. *Archives of General Psychiatry*, **38**, 408–413.

Gwet, K. L. (2002) Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Series: Statistical Methods For Inter-Rater Reliability Assessment*, **1**, 1–5.

— (2008) Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, **61**, 29–48.

— (2014) *Handbook of Inter-Rater Reliability.* Gaithersburg: Advanced Analytics, LLC.

Katz, D., Baptista, J., Azen, S. and Pike, M. (1978) Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics*, **34**, 469–474.

Krippendorff, K. (1970) Bivariate agreement coefficients for reliability of data. *Sociological Methodology*, **2**, 139–150.

Lafer-Sousa, R., Hermann, K. L. and Conway, B. R. (2015) Striking individual differences in color perception uncovered by 'the dress' photograph. *Current Biology*, **25**, R545–R546.

Scott, W. A. (1955) Reliability of content analysis:the case of nominal scale coding. *Public Opinion Quarterly*, **19**, 321–325.

Spitznagel, E. L. and Helzer, J. E. (1985) A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry*, **42**, 725–728.

Uebersax, J. S. (1987) Diversity of decision-making models and the measurement of interrater agreement. *Psychological bulletin*, **101**, 140.

Zee, J., Hodgin, J. B., Mariani, L. H., Gaut, J. P., Palmer, M., Bagnasco, S. M., Rosenberg, A. Z., Hewitt, S. M., Holzman, L. B., Gillespie, B. W. and Barisoni, L. (2018) Reproducibility and feasibility of strategies for morphologic assessment of renal biopsies using the nephrotic syndrome study network (neptune) digital pathology scoring system. *Archives of Pathology and Laboratory Medicine*, **142**, 613–625.

Zhao, X., Liu, J. S. and Deng, K. (2013) Assumptions behind intercoder reliability indices. *Annals of the International Communication Association*, **36**, 419–480.

Figure Legends

1. Agreement estimates by uncertainty of both raters for 0.80 true agreement
2. Agreement estimates with biased coin flips for uncertain ratings
3. Agreement estimates with inaccurate indications of uncertainty
4. Agreement estimates ± 1 standard error for seven pathology descriptors identified in kidney biopsies from the Nephrotic Syndrome Study Network (NEPTUNE)

Legends for Supporting Information

A.1 R code to implement simulation study

A.2 Additional simulation results

A.3 Dental study data example using summarized uncertainty data