Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/22964

Please be advised that this information was generated on 2018-07-07 and may be subject to change.

How Good Is That Agreement?

To the Editor:

Cohen's kappa is commonly used as a measure of chance-adjusted agreement. Warnings have been sounded[1–4] about the difficulties in interpreting kappa because of its dependence on the prevalence of the attribute being measured and the bias between observers. Nevertheless, researchers frequently appeal to the standards suggested by Landis and Koch[5] in discussing interrater reliability: <0.00 (Poor), 0.00–0.20 (Slight), 0.21–0.40 (Fair), 0.41–0.60 (Moderate), 0.61–0.80 (Substantial), 0.81–1.00 (Almost perfect).

With a small, implied criticism of the lower levels of this scale, Posner *et al*[6] comment, "One would view the finding of a negative kappa value, or a level of agreement less than expected by chance alone, as surprising and indeed serious."

The descriptions have been slightly modified by Altman[7]: <0.20 (Poor), 0.21–0.40 (Fair), 0.41–0.60 (Moderate), 0.61–0.80 (Good), 0.81–1.00 (Very good).

Alternatively, writers have used the suggestion of Fleiss[8]: <0.40 (Poor agreement beyond chance), 0.40–0.75 (Fair to good agreement beyond chance), >0.75 (Excellent agreement beyond chance), which has been interpreted in at least one study[9] to mean: <0.40 (Poor), 0.41–0.57 (Fair), 0.58–0.75 (Good), >0.75 (Excellent).

A kappa value of 0.59 is moderate according to one "standard" and good according to another; a value of 0.77 is substantial in one case and excellent in another. Clearly, these explanatory terms are arbitrary and are subject to the difficulties referred to above. It would be better if those who reported values of kappa and those who read the reports could manage without explanatory terms such as fair, good, etc, and develop a feeling for the coefficient itself, paying respect to the prevalence and bias. To do so, however, would be difficult for those who have little experience with kappa.

It can be helpful, in trying to attach a meaning to a calculated value of kappa, to consider a range of agreement tables free of bias and prevalence effects (in which case, kappa = PABAK, the prevalence-adjusted bias-adjusted kappa, as defined by Byrt *et al*[3]). For example, in Table 1, values of kappa are given for six tables, in which two observers, without bias, have rated 100 subjects as belonging to one or the other of two categories with equal prevalences.

The kappa values given by these tables can be used to define limits that are much the same as those that have been previously suggested but have the advantage that they provide an easy means of comparing a calculated value of kappa with a situation that is not complicated by prevalence and bias effects (Table 2).

Such a standard would not remove the need for researchers to pay heed to the issues of prevalence and bias, but it may make the interpretation of kappa more meaningful for nonexperts.

TABLE 1. Reference Tables

48	2	Kappa = 0.92	45	5	Kappa = 0.80
2	48		5	45	
40	10	Kappa = 0.60	35	15	Kappa = 0.40
10	40		15	35	
30	20	Kappa = 0.20	25	25	Kappa = 0.00
20	30		25	25	

TABLE 2. Proposed Kappa Description

0.93–1.00	Excellent agreement
0.81–0.92	Very good agreement
0.61–0.80	Good agreement
0.41–0.60	Fair agreement
0.21–0.40	Slight agreement
0.01–0.20	Poor agreement
0.00, or less	No agreement

Ted Byrt
Clinical Epidemiology and Biostatistics Unit,
Royal Children's Hospital,
Parkville, Victoria 3052, Australia
(address for correspondence)

References

1. Kraemer HC. Ramifications of a population model for kappa as a coefficient of reliability. Psychometrika 1979;44:461–472.
2. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ 1992;304:1491–1494.
3. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423–429.
4. Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. Epidemiology 1996;7:199–202.
5. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
6. Posner KL, Sampson PD, Caplan RA, Ward RJ, Cheney FW. Measuring interrater reliability among multiple raters: an example of methods for nominal data. Stat Med 1990;9:1103–1115.
7. Altman DG. Practical Statistics for Medical Students. London: Chapman and Hall, 1991.
8. Fleiss JL. Statistical Methods for Rates and Proportions. 2nd ed. New York: John Wiley and Sons, 1981;218.
9. Pinto J, Paneth N, Kazam E, Kairam R, Wallenstein S, Rose W, Rosenfeld D, Schonfeld S, Stein I, Witomski T. Interobserver variability in neonatal cranial ultrasonography. Paediatr Perinat Epidemiol 1988;2:43–58.

Seasonality Bias in Poor Reproductive Outcome

To the Editor:

Recently, Basso *et al*[1] warned about bias in studies of seasonality in poor reproductive outcome, stemming from a seasonal variation in pregnancy planning. Such variation in pregnancy planning was found in a survey in five European countries during the 1970s and 1980s.[1] For the same reason, bias can be expected in studies of seasonally bound exposure and poor reproductive outcome. Examples of these forms of exposure are the use of pesticides and herbicides, the occurrence of influenza, exposure to sunlight, and the consumption of vitamin C. Bias will occur if (a) the exposure has a seasonal distribution, and (b) the proportion of conceptions that end as a poor reproductive outcome varies throughout the season because of seasonal variation in pregnancy planning (as illustrated by Basso *et al*). The result can be a spurious relation between exposure and reproductive outcome, or, in the case of a real effect of exposure, over- or underestimation of the strength of the relation.

We studied the impact of this kind of bias in the association between exposure and poor reproductive outcome by means of a simulation study. Using the data presented by Basso *et al*,[1] we defined three subpopulations, which differed in probability of conception and probability of spontaneous abortion (Table 1). We defined the distribution of the proportion of women who started per month as a cosine function with a period of 1 year

TABLE 1. Simulations of Seasonality in Spontaneous Abortion (SAB) Risk Due to Exposure to *X* in a Population with Variation in Distributions of Fecundability and Abortion Risk (Conditional on the Null Hypothesis of No Real Effect of *X* on Abortion Risk)

Model	Subpopulation 1: Fecund, Low Risk of SAB		Subpopulation 2: Intermediate		Subpopulation 3: Subfecund, High Risk of SAB		RR‡ (95% CI)	Range of SAB/ Month (%)
	P_C*	P_{SAB}†	P_C	P_{SAB}	P_C	P_{SAB}		
1. Basso *et al*[1]	36.7	6.3	21.7	8.8	9.2	14.5	1.03 (1.01–1.05)	9.5–10.8
2. More variation in P_{SAB}	36.7	3.1	21.7	8.8	9.2	29.1	1.06 (1.04–1.08)	12.5–16.5
3. Large fecund population	36.7	3.1	36.7	3.1	9.2	29.1	1.10 (1.08–1.12)	10.2–15.7

* P_C = probability of conception per month (in percentages).
† P_{SAB} = probability of spontaneous abortions per month (in percentages).
‡ RR = relative spontaneous abortion risk for exposure to *X*, relative to nonexposed.

and a shift of 7.5 months. This definition resulted in a seasonal pattern in pregnancy planning, with a maximum probability of 16.6% for starting in August to try to conceive and a minimum of 2.7% for starting in February. For simplicity, we used the moment of conception as the etiologic moment for spontaneous abortion. To maximize the overestimation, we defined the seasonal pattern in the probability of exposure as a cosine function with a period of 1 year and a shift of 3 months. We set the proportion of women exposed at 10%. These assumptions led to variation in the probability of exposure, with a maximum of 22.6% for conceptions in March and April and a minimum of 4.1% for conceptions in September and October. Note that the probability of exposure did not vary among the three subpopulations and thus was not related to the degree of fecundity, nor to the probability of spontaneous abortion. The simulation was based on 100,000 women per year who planned to become pregnant and continued for a period of 20 years. As the model was not stable during the first years of simulation, we analyzed only the results of the last 10 years.

The results are shown in Table 1. Model 1 used the data presented by Basso *et al* and resulted in a relative risk of spontaneous abortion for exposed *vs* unexposed women of 1.03 [95% confidence interval (CI) = 1.01–1.05]. We found the largest bias when we defined a large fecund population (Subpopulations 1 and 2 combined) with extreme variation in the probabilities of spontaneous abortion (Model 3). This model resulted in a relative risk of 1.10 (95% CI = 1.08–1.12). In all of the simulations, we found the highest probability of spontaneous abortions after conceptions in March or April and the lowest after conceptions in September.

We conclude that bias in the relation between seasonally bound exposure and poor reproductive outcome does occur because of seasonal variation in pregnancy planning, but, for practical purposes, this bias will be negligible.

Annette M. Stolwijk
Huub Straatman
Gerhard A. Zielhuis
Department of Medical Informatics, Epidemiology and Statistics,
University of Nijmegen, P.O. Box 9101,
NL-6500 HB Nijmegen, The Netherlands
(address for correspondence)

Reference

1. Basso O, Olsen J, Bisanti L, Juul S, Boldsen J, and the European Study Group on Infertility and Subfecundity. Are seasonal preferences in pregnancy planning a source of bias in studies of seasonal variation in reproductive outcomes? Epidemiology 1995;6:520–524.

LETTERS

Letters may be submitted by e-mail (epidemiol@aol.com), facsimile transmission (617-244-9669), or regular mail (Kenneth J. Rothman, Editor, Epidemiology, One Newton Executive Park, Newton Lower Falls, MA 02162-1450, USA). Include full names, postal and e-mail addresses, signatures, and daytime telephone and FAX numbers. Letters should not exceed 600 words. Letters sent by mail or facsimile transmission should be triple-spaced in 12-point courier font.