

Patient self-report and medical records

Measuring agreement for binary data

Angela M. Barbara MSc Mark Loeb MD Lisa Dolovich MSc PharmD Kevin Brazil MA PhD Margaret L. Russell MD PhD

Research data in family medicine often comes from 2 sources: self-report and medical record review. Frequently, the quality of these data sources is assumed to be high, but measuring the reproducibility of these data is essential to evaluating the quality of the information collected. In ideal circumstances, data obtained from either data source would be equivalent. However, no source of data is without error. In cases of low agreement between data sources, research findings differ depending on the method of data collection used¹ and leave the researcher with questions about which estimate is correct. Comparing data from different sources can give family medicine researchers insight into which data source is most appropriate to answer a specific research question or can direct efforts to improve the collection and recording of health data.²

Imagine that we are interested in the prevalence of fever or cough in outpatients over the past influenza season. Neither the medical record nor patient self-report is considered the true criterion standard for symptoms. We are not assessing the accuracy of one data source compared with another; rather, we are examining agreement between the sources of data. The presence or absence of patient symptoms is considered a binary variable—a categorical variable in which there are 2 possible conditions (eg, yes or no, positive or negative). This paper describes indicators for determining agreement between binary variables: total agreement, κ , and positive and negative agreement.

Interpreting the value of κ

Table 1 displays data from the Hutterite Influenza Prevention Study in 2×2 contingency tables.³ Symptoms

reported by Hutterite community members were compared with documentation in the medical records. Total agreement is the number of concordant pairs divided by the total sample. In Table 1A, total agreement is 74%, which is the number of concordant yes's for fever (18) plus the concordant no's (112) divided by 176 participants. However, this simple measure does not take into account that a certain amount of agreement between medical charts and self-report is expected by chance alone⁴; assessment of κ , on the other hand, measures the strength of agreement beyond what we expect solely by chance. The calculation for κ is as follows:

$$\kappa = \frac{\text{total agreement} - \text{chance agreement}}{1 - \text{chance agreement}}$$

The answer falls on a scale of -1 to 1, where 0 equals chance agreement and 1 equals perfect agreement. In 1977, Landis and Koch proposed the following guidelines for understanding κ values: less than 0 equals no agreement, 0.01 to 0.20 equals slight agreement, 0.21 to 0.40 equals fair agreement, 0.41 to 0.60 equals moderate agreement, 0.61 to 0.80 equals substantial agreement, and 0.81 to 1.0 equals almost perfect agreement.⁵ While these guidelines are widely used and cited, the cutoffs are not universally accepted and have been criticized for being arbitrary divisions based on personal opinion rather than evidence.^{6,7}

The value of κ is not simple to interpret because it is influenced by the prevalence of the variable being measured.⁸ Table 1A and 1C have similar total agreements (as do 1B and 1D), but κ values differ according to distributions. The κ value represents the proportion of total

Table 1. Contingency tables of data from the Hutterite Influenza Prevention Study³: A) fever, B) earache, C) cough, and D) chills.

		MEDICAL RECORD				MEDICAL RECORD			
		YES	NO			YES	NO		
A)	SELF-REPORT	YES	18	6	B)	SELF-REPORT	YES	15	4
		NO	40	112			NO	18	139
Total agreement = 0.74, κ = 0.31, Positive agreement = 0.44 Negative agreement = 0.83				Total agreement = 0.88, κ = 0.51, Positive agreement = 0.58 Negative agreement = 0.93					
		MEDICAL RECORD				MEDICAL RECORD			
		YES	NO			YES	NO		
C)	SELF-REPORT	YES	82	20	D)	SELF-REPORT	YES	2	16
		NO	30	44			NO	11	147
Total agreement = 0.72, κ = 0.41, Positive agreement = 0.77 Negative agreement = 0.64				Total agreement = 0.85, κ = 0.05, Positive agreement = 0.13 Negative agreement = 0.92					

variance that is not attributable to chance or random error. Because total variance is minimal in a uniform (homogeneous) population where there is a relatively high (or low) prevalence, κ will be low even though total agreement might be high (Table 1D). Because chance agreement is smallest in a mixed (heterogeneous) population, κ will be higher when prevalence is closer to 50% (Table 1B and 1C). This makes it difficult to compare κ values between patient symptoms or other variables and different prevalences.⁹

Calculation of κ is also influenced by bias or the disagreement in the proportion of positive or negative cases (number of discordant responses)⁶; that is, the mismatch of positive or negative cases or disagreements are not random but go in one direction rather than another,^{8,10} which tends to happen when the prevalence of a symptom is high or low. This might result in a low κ value even though agreement is substantial (Table 1A and 1D); the value of κ is higher when there is a large bias and lowest when bias is absent.¹¹

The κ value does not distinguish between various types and sources of agreement and disagreement.^{6,8,12,13} The aim of measuring agreement is to discover the bases of differences and reduce them if possible, rather than, for example, simply quantifying the degree of disagreement.⁹ In fact, it might be that no single agreement statistic can adequately capture agreement.¹¹

Calculating positive and negative agreement


To help interpret κ values, calculating both positive and negative agreement has been recommended.^{11,14} The formula for calculating positive agreement is as follows:

$$\frac{2 \times \text{concordant positives}}{(\text{positive pair} + \text{positive pair})}$$

Negative agreement is calculated as follows:

$$\frac{2 \times \text{concordant negatives}}{(\text{negative pair} + \text{negative pair})}$$

Using these indices also provides insight into the agreement and imbalance in the proportion of positive or negative responses. This information is useful in determining where the focus should be to improve data quality depending on what is most important, which would be missed by calculating solely the κ value and total agreement.^{2,11,14} Low positive agreement indicates there is poor concordance between both sources in reporting the presence of the symptom (Table 1D), whereas high negative agreement means there is good concordance between both sources in identifying that the symptom was not experienced¹⁴ (Table 1A, 1B, 1D).

Family medicine practitioners should consider these concepts when evaluating various aspects of clinical care, such as data collection for a new practice quality assurance process. Although total agreement and the value of κ are commonly reported in agreement studies, we recommend the additional calculation of positive agreement and negative agreement. 

Ms Barbara is a researcher in the Department of Pathology and Molecular Medicine and a doctoral candidate in the Health Research Methodology Program of the Department of Clinical Epidemiology and Biostatistics at McMaster University in Hamilton, Ont. **Dr Loeb** is Professor in the Department of Pathology and Molecular Medicine and a joint member of the Department of Clinical Epidemiology and Biostatistics at McMaster University. **Dr Dolovich** is Research Director and Associate Professor in the Department of Family Medicine and the Centre for Evaluation of Medicines at McMaster University. **Dr Brazil** is Director of St Joseph's Health System Research Network and Professor in the Department of Clinical Epidemiology and Biostatistics at McMaster University. **Dr Russell** is Associate Professor in the Department of Community Health Sciences at the University of Calgary in Alberta.

Competing interests

None declared

Correspondence

Angela Barbara, Infectious Diseases Research Unit, Pathology and Molecular Medicine, 1280 Main St W, MDCL 3200, Hamilton, ON L8N 4K1; telephone 905 525-9140, extension 21478; fax 905 389-5822; e-mail barbara@mcmaster.ca

References

- Zhu K, McKnight B, Stergachis A, Daling JR, Levine RS. Comparison of self-report data and medical records data: results from a case-control study on prostate cancer. *Int J Epidemiol* 1999;28(3):409-17.
- Westbrook JI, McIntosh JH, Rushworth RL, Berry G, Duggan JM. Agreement between medical record data and patients' accounts of their medical history and treatment for dyspepsia. *J Clin Epidemiol* 1998;51(3):237-44.
- Loeb M, Russell ML, Moss L, Fonseca K, Fox J, Earn DJ, et al. Effect of influenza vaccination of children on infection rates in Hutterite communities: a randomized trial. *JAMA* 2010;303(10):943-50.
- Banerjee M, Capozzoli M, McSweeney L, Sinha D. Beyond kappa: a review of interrater agreement measures. *Can J Stat* 1999;27(1):3-23.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33(1):159-74.
- Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;85(3):257-68.
- Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992;304(6840):1491-4.
- Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990;43(6):543-9.
- Sargeant JM, Martin SW. The dependence of kappa on attribute prevalence when assessing the repeatability of questionnaire data. *Prev Vet Med* 1998;34(2-3):115-23.
- Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol* 1993;46(5):423-9.
- Chen G, Faris P, Hemmelgarn B, Walker RL, Quan H. Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. *BMC Med Res Methodol* 2009;9:5.
- Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol* 1988;41(10):949-58.
- Lantz CA, Nebenzahl E. Behavior and interpretation of the kappa statistic: resolution of the two paradoxes. *J Clin Epidemiol* 1996;49(4):431-4.
- Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 1990;43(6):551-8.

Hypothesis is a quarterly series in *Canadian Family Physician*, coordinated by the Section of Researchers of the College of Family Physicians of Canada. The goal is to explore clinically relevant research concepts for all CFP readers. Submissions are invited from researchers and nonresearchers. Ideas or submissions can be submitted online at <http://mc.manuscriptcentral.com/cfp> or through the CFP website www.cfp.ca under "Authors."