In a *Journal* article, Yawn and Wollan (1) reported interrater agreement to determine the reliability of data abstracted from the ambulatory care and inpatient medical records of 1,200 patients with heart disease. Six to nine nurses had abstracted three categories of data from the same medical records at five intervals over 2 years. The data abstraction procedure involved the nurses making a provisional diagnosis of heart disease based on the abstracted data. The authors found “very good to excellent” agreement.

They declared that there were no standard methods for assessing interrater reliability (1). Among several authors, Fleiss (2) has described several methods of measuring interrater agreement. Contrasting observed with expected agreement and correcting or adjusting for chance-expected agreement was a common factor in these methods (2). The kappa statistic (3) or a similar interclass correlation coefficient is generally accepted as a quantitative measure of interrater agreement. Medical records are commonly used in medical care or outcomes research, but, unlike in Yawn and Wollan’s study (1), the data abstractors are not required to interpret symptoms and other data and make diagnoses. The abstractors usually copy the diagnoses recorded by the attending physicians, or the researchers use standard algorithms or indices to convert abstracted data elements into diagnoses (e.g., Ellerbeck et al. (4), Oei et al. (5)), thereby avoiding misclassification.

Surprisingly, the authors chose not to use the kappa statistic to assess interrater agreement because it was irrelevant to their study (1). All raters in their study were well-qualified nurses familiar with the medical records, the terminology used therein, and the data elements abstracted. For these reasons, they were expected to agree with each other (6). Furthermore, agreement between two or more raters may occur simply by chance, and there has to be a way of separating observed from expected-chance agreement (7). The kappa statistic corrects or adjusts for this expected-chance agreement. Because of the multiple ratings per subject, a weighted average of kappa or some other formula, such as those described by Fleiss (8), was necessary in Yawn and Wollan’s study (1). Consequently, they could not tell how much of the observed “excellent” agreement was due to expected-chance agreement.

**ACKNOWLEDGMENTS**

The views expressed in this letter do not represent the views and policies of the Centers for Medicare & Medicaid Services or the United States.

Conflict of interest: none declared.

**REFERENCES**


Kazim Sheikh
US Department of Health and Human Services, Centers for Medicare & Medicaid Services, Kansas City, MO 64106

DOI: 10.1093/aje/kwi287; Advance Access publication September 8, 2005