Simulation-based demonstration of the kappa prevalence problem

Istvan Janosi – 16 Sep 2026

MRMC simulation, inter-reader agreement, and comparison with Gwet's AC1

Key result. The simulator produced readers with approximately 85% sensitivity and 95% specificity on average, and about 89% inter-reader raw agreement. In the fixed analysis dataset, observed inter-reader agreement was 0.883, yet Fleiss' kappa was only 0.511. Applying Gwet's AC1 to exactly the same ratings yielded 0.846. The discrepancy arose from the different definitions of expected chance agreement, not from any change in the data.

1. Why this simulation was created

The purpose of the simulation was to create a transparent multi-reader, multi-case (MRMC) setting in which the underlying diagnostic performance of the readers was known. This is important because agreement statistics are often interpreted without knowing whether a low coefficient reflects genuinely poor reader performance or a property of the coefficient itself. A simulation allows these two possibilities to be separated.

The specific methodological question was whether a conventional chance-corrected agreement coefficient can produce a surprisingly low value even when readers have good diagnostic accuracy and a high observed level of agreement. This question is directly relevant to diagnostic studies with unbalanced outcomes, particularly settings in which positive findings are relatively uncommon.

Cohen introduced kappa as a chance-corrected measure of agreement for two raters [1], and Fleiss later generalized the idea to multiple raters [2]. Kappa has become standard in clinical agreement studies, but its dependence on marginal category frequencies has been recognized for decades. Feinstein and Cicchetti described situations in which high observed agreement coexists with low kappa [3], and Byrt, Bishop and Carlin showed explicitly how prevalence and rater bias affect kappa [4].

2. The MRMC data-generating model

The simulated study used a fully crossed binary MRMC design: every reader evaluated every case, and every case was evaluated twice. The final design contained 500 cases, 5 readers, and 2 reading sessions, giving 5,000 individual ratings.

Component	Specification	Purpose
Cases	500	Provides a sufficiently large case sample while remaining easy to inspect and reproduce.
Readers	5	Allows a genuine multi-reader rather than pairwise agreement analysis.

Component	Specification	Purpose
Reading sessions	2	Supports later intra-reader analysis.
Outcome	Binary	Positive / negative diagnostic classification.
Target prevalence	10%	Creates an intentionally imbalanced diagnostic setting.
Mean reader sensitivity	approximately 85%	Controls diagnostic accuracy among truly positive cases.
Mean reader specificity	approximately 95%	Controls diagnostic accuracy among truly negative cases.
Design	Fully crossed	Every reader evaluates every case in both sessions.

The data-generating mechanism included several distinct sources of variation. Cases had different levels of intrinsic difficulty; readers had slightly different sensitivities and specificities; a stable reader-by-case component allowed a particular reader to interpret a particular case systematically differently from another reader; and small session-specific and assessment-specific fluctuations were added to avoid unrealistically deterministic repeated readings.

A key refinement was required for the repeated readings. When two completely independent Bernoulli draws were used for the same reader-case combination, the model introduced too much random intra-reader variation. Conversely, using an identical decision threshold in both sessions produced unrealistically high intra-reader agreement (approximately 0.99). The final implementation therefore used correlated, but non-identical, latent decision thresholds for the two readings. This produced a more plausible distinction between inter-reader and intra-reader reproducibility.

3. Validation of the simulator

Before using a single fixed dataset for the blog analysis, the generator was checked across 100 Monte Carlo replications. The purpose of this step was not to force the simulator to reproduce exact values to several decimal places, but to verify that it represented the intended diagnostic environment consistently.

Validation metric	Monte Carlo mean	Interpretation
Sensitivity	0.848	Close to the intended 0.85 reader sensitivity.
Specificity	0.951	Close to the intended 0.95 reader specificity.
Inter-reader raw agreement	0.891	High, but not near-perfect, agreement between different readers.
Intra-reader raw agreement	0.929	Higher than inter-reader agreement, as expected for repeated readings by the same reader.

These values are Monte Carlo averages over repeated simulated datasets. The subsequent agreement comparison used one fixed dataset generated with a predefined seed. Therefore, the observed agreement in that particular dataset need not be numerically identical to the Monte Carlo mean. This distinction is important: the validation describes the behavior of the simulator, whereas the fixed dataset provides one reproducible example for the methodological demonstration.

4. Inter-reader agreement analysis in the fixed dataset

For the initial inter-reader analysis, only the first reading session was used. This produced a conventional 500-case by 5-reader rating matrix. The same matrix was analyzed with several agreement coefficients. No data were changed between methods; only the mathematical definition of the agreement coefficient changed.

For Fleiss' kappa, observed agreement (P_o) is corrected by an expected chance agreement (P_e):

kappa = (P_o - P_e) / (1 - P_e)

The important point is that P_e is not an empirically observed amount of random agreement. It is a model-based reference quantity derived from the marginal category proportions. When one category dominates, the kappa chance-agreement term can become very large.

Gwet's AC1 uses the same general chance-correction structure but a different definition of expected chance agreement. Gwet proposed AC1 specifically to obtain a more stable coefficient in situations where kappa behaves paradoxically under high agreement and imbalanced margins [5]. For a binary outcome with pooled positive rating proportion p, the conceptual contrast is especially transparent: kappa uses a chance agreement term based on p^2 + (1-p)^2, whereas AC1 uses a term proportional to 2p(1-p). As p becomes extreme, these quantities move in opposite directions.

5. Results: high observed agreement, moderate kappa

The fixed dataset produced an observed inter-reader agreement of 0.883. Despite this high raw agreement, Fleiss' kappa was only 0.511. Conger's kappa and Krippendorff's alpha were almost identical in this particular complete, balanced nominal design. Gwet's AC1, in contrast, was 0.846.

Method	Observed agreement	Expected chance agreement	Coefficient	95% CI
Percent agreement	0.883	N/A	0.883	0.866 to 0.900
Fleiss' kappa	0.883	0.760	0.511	0.441 to 0.581
Conger's kappa	0.883	0.760	0.511	0.441 to 0.581
Gwet's AC1	0.883	0.240	0.846	0.819 to 0.873
Krippendorff's alpha	0.883	0.760	0.511	0.441 to 0.582
Brennan-Prediger	0.883	0.500	0.766	0.731 to 0.800

The most important comparison is therefore not 0.511 versus 0.846 in isolation, but the decomposition that produced these values. Both Fleiss' kappa and Gwet's AC1 started from exactly the same observed agreement, P_o = 0.883. The difference was the expected chance agreement: approximately 0.760 for Fleiss' kappa versus approximately 0.240 for AC1.

Numerical illustration. For Fleiss' kappa: (0.883 - 0.760) / (1 - 0.760) = approximately 0.51. Thus, the coefficient is low not because observed reader agreement is low, but because the kappa model regards about 76% agreement as already expected from the marginal distribution.

6. What does the result mean?

The result should not be interpreted as saying that the readers agree only 51% of the time. They agree approximately 88% of the time. Kappa = 0.51 means that, under the kappa-specific chance model, the

observed agreement accounts for about 51% of the maximum possible improvement above the expected chance agreement.

This distinction is particularly important in diagnostic settings with low prevalence. If negative classifications dominate, two readers will both assign many negative ratings. Kappa treats a large part of this concordance as expected from the marginal proportions, so the correction can be severe. This is the mechanism underlying the classic prevalence paradox described by Feinstein and Cicchetti [3] and further analyzed by Byrt and colleagues [4].

The simulation makes the phenomenon unusually easy to interpret because the underlying reader performance is known. The readers were generated to have good diagnostic accuracy: approximately 85% sensitivity and 95% specificity on average. The simulator also produced high inter-reader raw agreement. Nevertheless, the fixed example yielded a Fleiss kappa of approximately 0.51. Thus, a moderate kappa value can arise even in a setting in which neither individual reader accuracy nor raw agreement is poor.

7. Does Gwet's AC1 'solve' the problem?

Gwet's AC1 was developed precisely to reduce the instability of kappa in situations of high agreement and unbalanced category frequencies [5]. In the present example, AC1 = 0.846 is much closer to the observed agreement of 0.883 and to the qualitative impression suggested by the known reader accuracy. This is a strong indication that AC1 is less distorted by the prevalence structure in this dataset.

However, AC1 should not be described as an automatically correct replacement for kappa in every study. The two statistics embody different definitions of chance agreement. Choosing between them is therefore not simply a computational decision; it reflects what null model of agreement is considered scientifically meaningful. A defensible analysis should report the observed agreement itself, describe the category distribution, and explain why a particular chance-corrected coefficient was selected.

For this reason, the most informative reporting strategy in a methodological demonstration is to show several quantities together: observed agreement, the chance agreement implied by each coefficient, and the resulting corrected coefficient. Reporting only kappa conceals the mechanism by which the low value was obtained.

8. Why did several other coefficients equal kappa here?

In this particular simulated dataset, Fleiss' kappa, Conger's kappa and Krippendorff's alpha all produced values near 0.51. This does not mean that these methods are generally identical. The current data are complete, fully crossed, binary and nominal, with the same raters assessing every case. Under such a simple design, several nominal agreement estimators can become numerically very similar because they rely on closely related marginal or disagreement structures.

Krippendorff's alpha is especially valuable in more general settings because it naturally accommodates multiple raters, missing observations and different measurement levels. Hayes and Krippendorff advocated alpha as a broadly applicable reliability coefficient for coding data [6]. Its similarity to kappa in this example is therefore a feature of this specific design, not a reason to regard the methods as interchangeable in all applications.

9. Next analysis: deliberately changing prevalence

The current example provides a single demonstration at approximately 10% disease prevalence. The next step is more informative: hold the reader performance mechanism essentially constant while changing only

prevalence, for example across 50%, 30%, 20%, 10%, 5%, 2% and 1%. For each scenario, the same agreement measures can be recalculated.

This experiment will allow the key methodological question to be visualized directly. If sensitivity, specificity and the underlying reader mechanism remain stable while kappa changes strongly with prevalence, then the coefficient is responding not only to reader agreement but also to the marginal category distribution. If AC1 remains more stable, the contrast will provide a direct simulation-based illustration of why prevalence-sensitive agreement coefficients may be difficult to compare across clinical populations with different disease frequencies.

10. Interim conclusion

The simulation has achieved its initial purpose. It created a realistic MRMC setting with good reader-level diagnostic performance and high observed inter-reader agreement. In the fixed dataset, however, Fleiss' kappa was only 0.51 because its expected chance agreement rose to approximately 0.76 under the imbalanced category distribution. Gwet's AC1 applied to the same data produced 0.85 because its chance-agreement model behaved very differently. The example therefore demonstrates that a moderate kappa coefficient does not necessarily imply moderate raw agreement or poor diagnostic performance.

The main methodological message is not that kappa is universally wrong or that AC1 is universally right. Rather, chance-corrected agreement coefficients can answer subtly different statistical questions, and prevalence can materially affect their numerical values. Agreement studies should therefore report observed agreement, category frequencies and the rationale for the chosen coefficient, rather than interpreting a single kappa value in isolation.

11. Notes

Analysis note: Agreement estimates shown above are from the fixed simulated dataset used for the inter-reader analysis. Simulator performance values are Monte Carlo averages across 100 independently generated datasets. The analysis was performed on the first reading session for the inter-reader comparison.

AI use statement: Generative AI was used as a support tool during drafting, code refinement and presentation of results. The statistical methodology, simulation design, analyses and conclusions were independently reviewed by the authors, and all reported results can be reproduced from the accompanying source code and data.

References

1. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20(1):37-46. doi:10.1177/001316446002000104.

2. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971;76(5):378-382. doi:10.1037/h0031619.

3. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology. 1990;43(6):543-549. doi:10.1016/0895-4356(90)90158-L.

4. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. Journal of Clinical Epidemiology. 1993;46(5):423-429. doi:10.1016/0895-4356(93)90018-V.

5. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology. 2008;61(1):29-48. doi:10.1348/000711006X126600.

6. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Communication Methods and Measures. 2007;1(1):77-89. doi:10.1080/19312450709336664.