Mbizvo Gashirai K (Orcid ID: 0000-0002-9588-2944)
Bennett Kyle Hemingway (Orcid ID: 0000-0002-3922-7056)

# Using Critical Success Index or Gilbert Skill score as composite measures of positive predictive value and sensitivity in diagnostic accuracy studies: weather forecasting informing epilepsy research

**Authors:** Gashirai K Mbizvo[1,2], Kyle H Bennett[2], Colin R Simpson[3,4], Susan E Duncan[2,5], Richard FM Chin[2, 6], Andrew J Larner[7]

[1] Pharmacology and Therapeutics, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, United Kingdom

[2] Muir Maxwell Epilepsy Centre, Centre for Clinical Brain Sciences, The University of Edinburgh, Edinburgh, United Kingdom

[3] The Usher Institute, The University of Edinburgh, Edinburgh, United Kingdom

[4] School of Health, Wellington Faculty of Health, Victoria University of Wellington, Wellington, New Zealand

[5] Department of Clinical Neurosciences, NHS Lothian, Edinburgh, United Kingdom

[6] Royal Hospital for Children and Young People, Edinburgh, United Kingdom

[7] Cognitive Function Clinic, Walton Centre NHS Foundation Trust, Liverpool, United Kingdom

**Corresponding Author**

Dr Gashirai Mbizvo, Pharmacology and Therapeutics, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Biosciences Building, Crown Street, Liverpool, L69 7BE

*Tel:* +44 (0)151 795 4400, Fax: N/A

*Email:* Gashirai.Mbizvo@liverpool.ac.uk

**KEYWORDS:** Diagnostic test accuracy, routine data, threat score, ratio of verification, validation studies, systematic review

**Counts:** Number of text pages: 4; words: 721; references: 13; figures: 1.

**ORCID:** *Gashirai Mbizvo* - https://orcid.org/0000-0002-9588-2944, *Colin Simpson* - https://orcid.org/0000-0002-7256-3027, *Richard Chin* - https://orcid.org/0000-0002-7256-3027

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/epi.17537

We recently published a systematic review ascertaining the accuracy of using administrative data to identify epilepsy cases.[1] This showed that the commonest reported diagnostic accuracy measures across studies validating administrative epilepsy data (n=30) are positive predictive value (PPV) and sensitivity (Sens): 28 studies reported PPV, of which 14 also reported Sens and two reported Sens alone. In contrast, negative predictive value (NPV) and specificity (Spec) were only reported in 14 studies, and one study reported Spec alone. It was difficult to identify the optimal diagnostic algorithms by NPV and Spec as these were nearly 100% across most studies due to very high numbers of true negatives (TN), often far outnumbering true positives (TP), false positives (FP), and false negatives (FN). Instead, we identified the optimal diagnostic algorithms by ranking them in order of PPV and Sens and making a judgement on which had the best balance of both high PPV and Sens, selecting a threshold of >80% to represent accuracy.[1]

In markedly imbalanced datasets like these, where mostly PPVs and Sens have been reported, it may be logical to apply a single diagnostic accuracy measure which encompasses both PPV and Sens to aid interpretation and ranking. We are not aware of any such single measures currently used in medical literature,[2] and we propose the use of the Critical Success Index (CSI)[3] or Gilbert Skill score (GS)[4] for this purpose. Both are used in weather forecasting.[5]

CSI (also known as threat score[5,6] or ratio of verification[4]) eschews TNs, such that:

$$CSI = TP/(TP + FP + FN).^{7}$$

In signal detection theory, CSI is defined as ratio of hits to the sum of hits, false alarms, and misses.[8,9] CSI may also be expressed in terms of Sens and PPV:

$$CSI = 1/[(1/PPV) + (1/Sens) - 1].^{7}$$

CSI values range from 0–1 , interpreted 0 = unable to forecast and 1 = perfect forecast.[5,10]

CSI is not unbiased, since CSI = TP/(sample size N – TN), giving lower scores for rarer events.[9] In such circumstances, GS (also known as equitable threat score (ETS)[5,11]) may be preferred as it takes into account the number of hits due to chance (CH),[8] where:

$$GS = TP - CH/(TP + FP + FN - CH)$$

and:

$$CH = (TP + FP) \times (TP + FN)/N$$

or:

$$CH = (\text{total forecasts of event}) \times (\text{total observations of event})/\text{sample size}.^{9}$$

GS values range from −1/3 to 1.[5] As we previously demonstrated,[9] there is a monotonic relation between CSI and GS, where GS values are lower than CSI.

We reanalysed data from the systematic review[1] to calculate CSI and GS scores for the 91 algorithms from ten studies where base data (TP, FP, FN, TN) were reported. CSI and GS scores were plotted alongside PPV and Sens for each algorithm, as percentages (Figure 1).

The plot shows CSI and GS scores are conservative, always less than or equal to the lower of the corresponding PPV and Sens. For CSI scores ≥0.8, both PPV and Sens were ≥0.8, whereas low CSI scores occurred when there was a large difference between PPV and Sens, even when one of these values was high (~0.9). The monotonic relationship between CSI and GS was preserved, with GS scores remaining generally more conservative than CSI. However, epilepsy provides a real-world example of how any differences between CSI and GS become negligible in studies were large sample sizes (N) are driven by very high numbers of TN far outnumbering TP, FP, and FN to the extent of markedly lowering CH (all studies to the right of Holden 2005 in Figure 1, raw data available here: www.bit.ly/3WzcsqP).

Although CSI and GS are established prediction metrics in meteorological literature,[3-6, 8, 10-12] few texts have translated them into medical literature.[9, 13] We provide here the first translation of CSI and GS into epilepsy literature. We suggest CSI may be an appropriate measure to complement Sens, Spec, PPV and NPV, particularly as it allows combined interpretation of PPV and Sens whilst also avoiding the inflation of NPV and Spec when there are many TNs. GS may be a better metric when there are fewer TN and more CH. Based on the current findings, we suggest a CSI of ≥0.8 would be a reasonable threshold score for achieving diagnostic accuracy. Optimal diagnostic thresholds for GS remain to be elucidated.

**Acknowledgements**

GKM is supported by an NIHR Clinical Lectureship (CL-2022-07-002).

**Conflicts of Interest**

None of the authors has any conflict of interest to disclose.

**Ethical Publication Statement**

We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

**Data Availability Statement**

Base data are available from the published systematic review.[1] Tabulated CSI and GS scores are available here: www.bit.ly/3WzcsqP.

**References**

1. Mbizvo GK, Bennett KH, Schnier C, et al. The accuracy of using administrative healthcare data to identify epilepsy cases: A systematic review of validation studies Epilepsia. 2020 Jul;61(7):1319-1335.

2. Larner AJ. Diagnostic Test Accuracy Studies in Dementia: A Pragmatic Approach. 2nd ed. 2019 ed. Cham: Springer International Publishing.

3. Schaefer JT. The critical success index as an indicator of warning skill Weather Forecast 1990;5:570-575.

4. Gilbert GK. Finley's tornado predictions Am Meteorol J 1884;1:166-172.

5. World Meteorological Organization. Forecast Verification for the African Severe Weather Forecasting Demonstration Projects; No. 1132. Geneva, Switzerland: World Meteorological Organization; 2014. Available from: https://library.wmo.int/doc_num.php?explnum_id=7868.

6. Palmer WC, Allen RA. Note on the accuracy of forecasts concerning the rain problem. Washington, DC: U.S. Weather Bureau manuscript; 1949.

7. Larner AJ. The 2x2 matrix: contingency, confusion and the metrics of binary classification. Cham, Switzerland: Springer; 2021.

8. Space Weather Prediction Center. Forecast Verification Glossary: National Oceanic and Atmospheric Administration; 2022. Available from: https://bit.ly/3A7BchD.

9. Larner AJ. Assessing cognitive screeners with the critical success index Progress in Neurology and Psychiatry. 2021 Jul;25(3):33-37.

10. Spyrou C, Varlas G, Pappa A, et al. Implementation of a Nowcasting Hydrometeorological System for Studying Flash Flood Events: The Case of Mandra, Greece Remote Sens-Basel. 2020 Sep;12(17).

11. Doswell CA, Davies-Jones R, Keller DL. On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables Weather and forecasting. 1990;5(4):576-585.

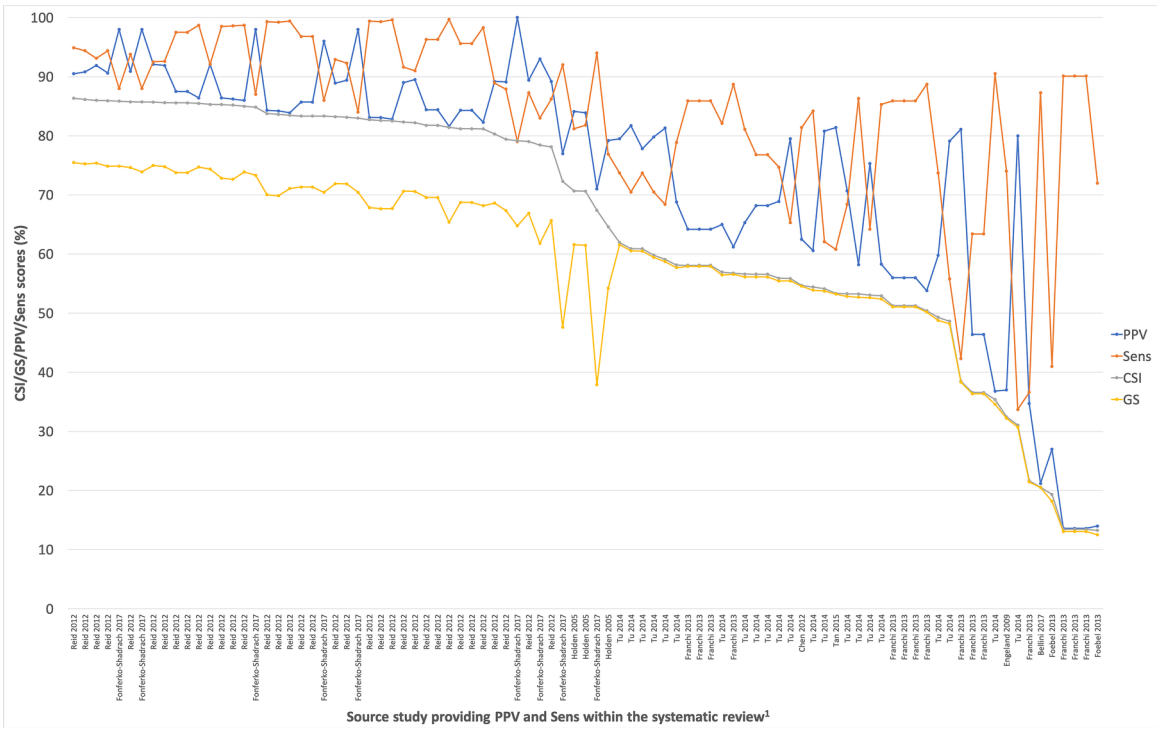
12. Gerapetritis H, Pelissier J, editors. The critical success index and warning strategy. 17th Conference on Probablity and Statistics in the Atmospheric Sciences, Seattle; 2004.

13. Mbizvo GK, Larner AJ. Isolated headache is not a reliable indicator for brain cancer Clin Med (Lond). 2022 Jan;22(1):92-93.

**Figure legend**

**Figure 1:** Dotted line plot of CSI, GS, PPV and Sens estimates across the systematic review algorithms

**Abbreviations:** CSI = Critical Success Index; GS = Gilbert Skill Score; PPV = Positive predictive value; Sens = Sensitivity




EPI_17537_2)_Plot_of_CSI_GS_PPV_and_Sens.tif