Critical success index or F measure to validate the accuracy of administrative healthcare data identifying epilepsy in deceased adults in Scotland

Citation for published version:
Mbizvo, GK, Simpson, CR, Duncan, SE, Chin, RFM & Larner, AJ 2024, 'Critical success index or F measure to validate the accuracy of administrative healthcare data identifying epilepsy in deceased adults in Scotland', *Epilepsy research*, vol. 199, pp. 107275. https://doi.org/10.1016/j.eplepsyres.2023.107275

Digital Object Identifier (DOI):
10.1016/j.eplepsyres.2023.107275

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Epilepsy research

Epilepsy Research 199 (2024) 107275

Contents lists available at ScienceDirect

Epilepsy Research

journal homepage: www.elsevier.com/locate/epilepsyres

Critical success index or F measure to validate the accuracy of administrative healthcare data identifying epilepsy in deceased adults in Scotland

Gashirai K. Mbizvo[a,b,c,*], Colin R. Simpson[d,e], Susan E. Duncan[f,g], Richard F.M. Chin[f,h], Andrew J. Larner[c]

[a] Pharmacology and Therapeutics, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, United Kingdom
[b] Liverpool Centre of Cardiovascular Science at University of Liverpool, Liverpool John Moores University and Liverpool Heart and Chest Hospital, Liverpool, United Kingdom
[c] Cognitive Function Clinic, Walton Centre NHS Foundation Trust, Liverpool, United Kingdom
[d] School of Health, Wellington Faculty of Health, Victoria University of Wellington, Wellington, New Zealand
[e] The Usher Institute, The University of Edinburgh, Edinburgh, United Kingdom
[f] Muir Maxwell Epilepsy Centre, Centre for Clinical Brain Sciences, The University of Edinburgh, Edinburgh, United Kingdom
[g] Department of Clinical Neurosciences, NHS Lothian, Edinburgh, United Kingdom
[h] Royal Hospital for Children and Young People, Edinburgh, United Kingdom

ARTICLE INFO

Keywords:
Diagnostic test accuracy
Routine data
Threat score
Ratio of verification
Validation studies
Systematic review

ABSTRACT

Background: Methods to undertake diagnostic accuracy studies of administrative epilepsy data are challenged by lack of a way to reliably rank case-ascertainment algorithms in order of their accuracy. This is because it is difficult to know how to prioritise positive predictive value (PPV) and sensitivity (Sens). Large numbers of true negative (TN) instances frequently found in epilepsy studies make it difficult to discriminate algorithm accuracy on the basis of negative predictive value (NPV) and specificity (Spec) as these become inflated (usually >90%). This study demonstrates the complementary value of using weather forecasting or machine learning metrics critical success index (CSI) or F measure, respectively, as unitary metrics combining PPV and sensitivity. We reanalyse data published in a diagnostic accuracy study of administrative epilepsy mortality data in Scotland.

Method: CSI was calculated as 1/[(1/PPV) + (1/Sens) – 1]. F measure was calculated as 2.PPV.Sens/(PPV + Sens). CSI and F values range from 0 to 1, interpreted as 0 = inaccurate prediction and 1 = perfect accuracy. The published algorithms were reanalysed using these and their accuracy re-ranked according to CSI in order to allow comparison to the original rankings.

Results: CSI scores were conservative (range 0.02–0.826), always less than or equal to the lower of the corresponding PPV (range 39–100%) and sensitivity (range 2–93%). F values were less conservative (range 0.039–0.905), sometimes higher than either PPV or sensitivity, but were always higher than CSI. Low CSI and F values occurred when there was a large difference between PPV and sensitivity, e.g. CSI was 0.02 and F was 0.039 in an instance when PPV was 100% and sensitivity was 2%. Algorithms with both high PPV and sensitivity performed best in terms of CSI and F measure, e.g. CSI was 0.826 and F was 0.905 in an instance when PPV was 90% and sensitivity was 91%.

Conclusion: CSI or F measure can combine PPV and sensitivity values into a convenient single metric that is easier to interpret and rank in terms of diagnostic accuracy than trying to rank diagnostic accuracy according to the two measures themselves. CSI or F prioritise instances where both PPV and sensitivity are high over instances where there are large differences between PPV and sensitivity (even if one of these is very high), allowing diagnostic accuracy thresholds based on combined PPV and sensitivity to be determined. Therefore, CSI or F measures may be helpful complementary metrics to report alongside PPV and sensitivity in diagnostic accuracy studies of administrative epilepsy data.

* Correspondence to: Pharmacology and Therapeutics, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7BE, United Kingdom.

E-mail addresses: Gashirai.Mbizvo@liverpool.ac.uk (G.K. Mbizvo), colin.simpson@vuw.ac.nz (C.R. Simpson), susanxduncan@gmail.com (S.E. Duncan), rchin@exseed.ed.ac.uk (R.F.M. Chin), ajlarner241@aol.com (A.J. Larner).

https://doi.org/10.1016/j.eplepsyres.2023.107275
Received 22 September 2023; Received in revised form 20 November 2023; Accepted 5 December 2023
Available online 12 December 2023

1. Introduction

In a recently published diagnostic accuracy study using administrative healthcare data to identify epilepsy in deceased adults in Scotland, four different sources of administrative data were used to develop diagnostic algorithms whose accuracy was then examined (Mbizvo et al., 2020a). Algorithms developed from one, two, or three database coding or antiepileptic drug (AED) strategies, respectively labelled levels 1, 2 and 3, were ranked according to the outcomes of interest. These were positive predictive value (PPV) and sensitivity (Sens), the most commonly used outcomes in diagnostic accuracy studies of administrative epilepsy data (Mbizvo et al., 2020b). Whilst negative predictive values (NPV) and specificity (Spec) can be used as outcome measures for diagnostic accuracy ranking (Chubak et al., 2012), these values are often > 90% because of the large numbers of true negative (TN) instances found in the base data of epilepsy studies, making it difficult to discriminate algorithm accuracy on the basis of these measures (Mbizvo et al., 2020b; Mbizvo et al., 2023). Therefore, algorithms are typically ranked in order of highest to lowest PPV and sensitivity, with priority given to those with higher values in both estimates (Horrocks et al., 2017; Mbizvo et al., 2020b; Mbizvo et al., 2020a; Wilkinson et al., 2018). However, such a method is challenging to apply objectively because there is a trade-off relationship between PPV and sensitivity, where one decreases as the other increases (Wang et al., 2021; Wilkinson et al., 2018). This makes it difficult to know which estimate to prioritise (PPV or sensitivity) when trying to rank the diagnostic algorithms in order of their accuracy.

There is clearly a need to consider novel ways to combine PPV and sensitivity into a single metric to make it easier and more objective to rank diagnostic algorithms by their accuracy. We propose the use of critical success index (CSI) (Schaefer, 1990) or F measure (Powers, 2015) (also known as the Dice co-efficient) (Jolliffe, 2016) for this purpose. CSI is commonly used in meteorology to verify the accuracy of weather forecasts (Doswell et al., 1990; Gerapetritis and Pelissier, 2004; Palmer and Allen, 1949; Schaefer, 1990; Space Weather Prediction Center, 2022; Spyrou et al., 2020; World Meteorological Organization, 2014). In signal detection theory, CSI is defined as ratio of hits to the sum of hits, false alarms, and misses (Larner, 2021, 2024; Space Weather Prediction Center, 2022). Therefore, it includes a measure of both the PPV and sensitivity. F measure (or F_1 score) is a machine learning evaluation metric that measures a model's accuracy as the weighted harmonic mean of precision (PPV) and recall (sensitivity) (Hicks et al., 2022; Powers, 2015). CSI values range from 0 to 1, interpreted as 0 = unable to forecast and 1 = perfect forecast (Spyrou et al., 2020; World Meteorological Organization, 2014). F measure is also bounded 0–1, where 1 represents perfect precision and recall values, and 0 represents absent precision and/or recall (Hicks et al., 2022). In effect, CSI or F measure combine PPV and sensitivity values into a convenient single metric that is more easy to interpret and rank in terms of diagnostic accuracy than trying to do so for the two measures (PPV and sensitivity) alongside one another. However, CSI is seldom used outside weather forecasting, with as yet no studies published using CSI in medicine, to our knowledge, beyond our recent proof-of-concept works (Larner, 2021; Mbizvo and Larner, 2022; Mbizvo et al., 2023). Although use of the F measure is commonplace in artificial intelligence (AI), its applicability and convenience as a non-AI diagnostic accuracy tool is yet to be fully demonstrated (Larner, 2021, 2024).

In the current study, we aim to reanalyse data published in a Scottish diagnostic accuracy study (Mbizvo et al., 2020a) of administrative epilepsy mortality data by calculating CSI scores for each of the diagnostic algorithms. We do this in order to see whether and how this alters the original diagnostic accuracy rankings proposed by the authors, which were done in order of highest to lowest PPV and sensitivity (Mbizvo et al., 2020a). This will help further understanding of the role CSI may play as a medical diagnostic accuracy measure. Corresponding F measures will also be reported for illustrative purposes as there is a monotonic relation between CSI and F (Jolliffe, 2016), meaning the ranking of CSI and F values calculated for any dataset is the same.

2. Methods

Details of the study design, study population, linkage of datasets and approvals accessed, as well as the algorithms used and their rankings may be found in the previous publication (Mbizvo et al., 2020a) (the

Table 1
Results of Level 1 validation of a single database coding or AED strategy.

Data Base (ranked in study's original accuracy order[1])	Coding algorithm	N	TP	FP	FN	PPV	Sens	CSI (95% CI) [ranking by CSI]	F
1 PIS	**≥ 1 AED (NL)**	**22,460**	**560**	**64**	**54**	**90% (87–92%)**	**91% (89–93%)**	**0.826 (0.797–0.854) [1]**	**0.905**
2 NRS	≥ 1 G40	2001	422	47	192	90% (87–93%)	69% (65–72%)	0.638 (0.602–0.675) [4]	0.779
3 NRS	≥ 1 G40–41	2143	446	70	168	86% (84–89%)	73% (69–76%)	0.652 (0.616–0.688) [3]	0.789
4 SMR01	≥ 1 G40–41, R56.8	8239	450	110	164	80% (77–84%)	73% (70–77%)	0.622 (0.586–0.657) [5 =]	0.767
5 PIS	≥ 1 AED (BL)	157,509	576	149	38	79% (77–82%)	94% (92–96%)	0.755 (0.724–0.785) [2]	0.860
6 Primary care	≥ 1 F25	1483	120	35	NA	77% (70–83%)	NA	-	-
7 NRS	≥ 1 G40–41, R56.8	2571	533	243	81	69% (65–72%)	87% (84–90%)	0.622 (0.589–0.654) [5 =]	0.767
8 NRS	≥ 1 G41	179	39	25	575	61% (49–73%)	6.4% (4–8%)	0.061 (0.042–0.080) [8]	0.115
9 NRS	≥ 1 R56.8	609	102	160	512	39% (33–45%)	17% (14–20%)	0.132 (0.108–0.156) [7]	0.233

Bold = most accurate algorithm in original study [1]
Abbreviations: SMR – Scottish Morbidity Record; PIS – Prescription Information Service; NRS – National Records of Scotland; AED – antiepileptic drug; F25 – primary care diagnostic Read Codes for epilepsy; G40–41 – International Classification of Disease 10 (ICD-10) codes for epilepsy and status epilepticus; R56.8 – ICD-10 code for seizures; NL – AEDs on the narrow list (appendix S2b); BL – AEDs on the broad list (appendix S2a); TP – true positive; FP – false positives; FN – false negatives, PPV – positive predictive value; Sn – sensitivity; CI – confidence intervals; NA – not applicable (negative cases unavailable as primary care data were taken from a 10% sample of Scottish GP practices; CSI – critical success index; F – F value.

Table 2
Results of Level 2 validation of algorithms combining two database coding or AED strategies together.

Data Base (ranked in study's original accuracy order[1])	Coding algorithm	N	TP	FP	FN	PPV	Sens	CSI (95% CI) [ranking by CSI]	F
1 Primary care + NRS	≥ 1 F25 + ≥1 R56.8	13	8	0	112	100% (100–100%)	7% (2–11%)	0.067 (0.022–0.111) [22]	0.125
2 NRS + PIS	≥ 1 G40–41 + ≥1 AED (NL)	1781	419	28	195	94% (92–96%)	68% (65–72%)	0.653 (0.616–0.689) [5]	0.790
3 NRS + PIS	≥ 1 G40 + ≥1 AED (NL)	1732	402	26	212	94% (92–96%)	66% (62–69%)	0.628 (0.591–0.666) [11]	0.772
4 SMR01 + PIS	≥ 1 G40–41, R56.8 + ≥1 AED (NL)	6501	418	34	196	93% (90–95%)	68% (64–72%)	0.645 (0.608–0.682) [7]	0.784
5 NRS + PIS	≥ 1 G40 + ≥1 AED (BL)	1798	407	31	207	93% (91–95%)	66% (63–70%)	0.631 (0.594–0.668) [9]	0.774
6 SMR01 + NRS	SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 G40	1167	325	23	289	93% (91–96%)	53% (49–57%)	0.510 (0.471–0.549) [15]	0.676
7 NRS + PIS	**≥ 1 G40–41, R56.8 + ≥1 AED (NL)**	**1921**	**494**	**47**	**120**	**91% (89–94%)**	**81% (77–84%)**	**0.747 (0.714–0.780) [3]**	**0.855**
8 NRS +PIS	≥ 1 G40–41 + ≥1 AED (BL)	1883	426	42	188	91% (88–94%)	69% (66–73%)	0.649 (0.613–0.686) [6]	0.787
9 SMR01 + PIS	≥ 1 G40–41, R56.8 + ≥1 AED (BL)	7227	424	48	190	90% (87–93%)	69% (65–73%)	0.640 (0.604–0.677) [8]	0.781
10 Primary care + NRS	≥ 1 F25 + ≥1 G40–41, R56.8	199	68	8	52	90% (83–96%)	57% (48–66%)	0.531 (0.445–0.618) [13]	0.694
11 SMR01 + NRS	SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 G40–41	1273	347	42	267	89% (86–92%)	57% (53–60%)	0.529 (0.491–0.567) [14]	0.692
12 Primary care + NRS	≥ 1 F25 + ≥1 G40–41	192	63	8	57	89% (81–96%)	53% (44–61%)	0.492 (0.406–0.579) [17]	0.660
13 Primary care + NRS	≥ 1 F25 + ≥1 G41	8	X	X	X	89% (81–96%)	53% (44–61%)	0.497 (NA) [16]	0.664
14 Primary care + NRS	≥ 1 F25 + ≥1 G40	188	62	8	58	89% (81–96%)	52% (43–61%)	0.484 (0.398–0.571) [18]	0.653
15 NRS + PIS	≥ 1 G41 + ≥1 AED (NL)	81	X	X	X	88% (77–99%)	5% (3–7%)	0.050 (NA) [25]	0.095
16 Primary care + PIS	**≥ 1 F25 + ≥1 AED (NL)**	**1089**	**111**	**18**	**9**	**86% (80–92%)**	**93% (88–97%)**	**0.804 (0.738–0.871) [1]**	**0.892**
17 Primary care + SMR01	≥ 1 F25 + ≥1 G40–41, R56.8	613	85	15	35	85% (78–92%)	71% (63–79%)	0.630 (0.548–0.711) [10]	0.773
18 NRS + PIS	≥ 1 G40–41, R56.8 + ≥1 AED (BL)	2107	507	98	107	84% (81–87%)	83% (80–86%)	0.712 (0.679–0.745) [4]	0.832
19 SMR01 + NRS	≥ 1 G40–41, R56.8	1390	386	76	228	84% (80–87%)	63% (59–67%)	0.559 (0.522–0.596) [12]	0.717
20 NRS + PIS	≥ 1 R56.8 + ≥1 AED (NL)	281	89	17	525	84% (77–91%)	15% (12–17%)	0.141 (0.114–0.168) [20]	0.247
21 Primary care + PIS	≥ 1 F25 + ≥1 AED (BL)	1255	112	27	8	81% (74–87%)	93% (89–98%)	0.762 (0.693–0.831) [2]	0.865
22 NRS + PIS	≥ 1 G41 + ≥1 AED (BL)	117	32	13	582	71% (58–84%)	5.2% (4–7%)	0.051 (0.034–0.068) [24]	0.097
23 SMR01 + NRS	SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 R56.8	220	58	30	556	66% (56–76%)	9% (7–12%)	0.090 (0.068–0.112) [21]	0.165
24 NRS + PIS	≥ 1 R56.8 + ≥1 AED (BL)	403	95	52	519	65% (57–72%)	16% (13–18%)	0.143 (0.116–0.169) [19]	0.250
25 SMR01 + NRS	SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 G41	139	36	21	578	63% (51–76%)	6% (4–8%)	0.057 (0.039–0.075) [23]	0.107

Bold = most accurate algorithm in original study [1]
Abbreviations: SMR – Scottish Morbidity Record; PIS – Prescription Information Service; NRS – National Records of Scotland; AED – antiepileptic drug; F25 – primary care diagnostic Read Codes for epilepsy; G40–41 – International Classification of Disease 10 (ICD-10) codes for epilepsy and status epilepticus; R56.8 – ICD-10 code for seizures; NL – AEDs on the narrow list; BL – AEDs on the broad list; TP – true positive; FP – false positives; FN – false negatives, PPV – positive predictive value; Sn – sensitivity; CI – confidence intervals. CSI – critical success index; F – F value.

rankings are reproduced here in Tables 1, 2 and 3). In brief, the following administrative healthcare databases were used to create algorithms: National Records of Scotland (NRS) death records; Scottish Morbidity Record 01 (SMR01) hospital admissions; Prescribing Information System (PIS) Scottish prescribing data; and GP primary care data. PIS was used to screen for deceased adults prescribed one or more AEDs, using both broad (36 AEDs) and narrow (21 AEDs) filters (Mbizvo et al., 2020a). The study was designed to conform to the Standards for Reporting of Diagnostic Accuracy studies (STARD 2015 iteration) (Bossuyt et al., 2015).

2.1. *Statistical analysis*

From the numbers of true positive (TP), false positive (FP), and false negative (FN) cases, CSI values were calculated for each algorithm according to the formula (Larner, in press):

CSI=TP/(TP + FN + FP)

This eschews true negatives (TN) (Mbizvo et al., 2023). CSI may also be expressed in terms of PPV and sensitivity:

CSI=1/[(1/PPV) + (1/Sens) – 1]·

This relation of CSI to PPV means that CSI is affected by prevalence,

Table 3
Results of Level 3 validation of algorithms combining three database coding or AED strategies together.

Data Base (ranked in study's original accuracy order[1])	Coding algorithm	N	TP	FP	FN	PPV	Sens	CSI (95% CI) [ranking by CSI]	F
1 Primary care + NRS + PIS	≥ 1 F25 + ≥1 R56.8 + ≥1 AED (BL)	11	7	0	113	100% (100–100%)	6% (2–10%)	0.058 (0.016–0.100) [20 =]	0.110
2 Primary care + NRS + PIS	≥ 1 F25 + ≥1 R56.8 + ≥1 AED (NL)	11	7	0	113	100% (100–100%)	6% (2–10%)	0.058 (0.016–0.100) [20 =]	0.110
3 Primary care + SMR01 + NRS	≥ 1 F25 + SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 R56.8	11	7	0	113	100% (100–100%)	6% (2–10%)	0.058 (0.016–0.100) [20 =]	0.110
4 Primary care + NRS + PIS	≥ 1 F25 + ≥1 G41 + ≥1 AED (BL)	8	X	X	X	100% (100–100%)	3% (0–5%)	0.03 (NA) [25 =]	0.058
5 Primary care + NRS + PIS	≥ 1 F25 + ≥1 G41 + ≥1 AED (NL)	8	X	X	X	100% (100–100%)	3% (0–5%)	0.03 (NA) [25 =]	0.058
6 Primary care + SMR01 + NRS	≥ 1 F25 + SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 G41	7	X	X	X	100% (100–100%)	2% (0–4%)	0.02 (NA) [27]	0.039
7 SMR01 + NRS + PIS	SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 G40–41 + ≥1 AED (NL)	1119	328	19	286	95% (92–97%)	53% (50–57%)	0.518 (0.479–0.557) [5]	0.683
8 SMR01 + NRS + PIS	SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 G40 + ≥1 AED (BL)	1101	314	17	300	95% (93–97%)	51% (47–55%)	0.498 (0.459–0.537) [9 =]	0.665
9 SMR01 + NRS + PIS	SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 G40 + ≥1 AED (NL)	1075	314	17	300	95% (93–97%)	51% (47–55%)	0.498 (0.459–0.537) [9 =]	0.665
10 SMR01 + NRS + PIS	≥ 1 G40–41, R56.8 + ≥1 AED (NL)	1185	366	26	248	93% (91–96%)	60% (56–64%)	0.572 (0.534–0.610) [3]	0.728
11 SMR01 + NRS + PIS	SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 G40–41 + ≥1 AED (BL)	1168	331	26	283	93% (90–95%)	54% (50–58%)	0.517 (0.478–0.556) [6]	0.682
12 SMR01 + NRS + PIS	≥ 1 G40–41, R56.8 + ≥1 AED (BL)	1247	370	38	244	91% (88–94%)	60% (56–64%)	0.567 (0.529–0.606) [4]	0.724
13 Primary care + SMR01 + PIS	≥ 1 F25 + ≥1 G40–41, R56.8 + ≥1 AED (NL)	555	79	9	41	90% (83–96%)	66% (57–74%)	0.612 (0.528–0.696) [1]	0.760
14 SMR01 + NRS + PIS	SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 R56.8 + ≥1 AED (NL)	156	56	6	558	90% (83–98%)	9% (7–11%)	0.090 (0.068–0.113) [19]	0.166
15 Primary care + NRS + PIS	≥ 1 F25 + ≥1 G40–41, R56.8 + ≥1 AED (BL)	195	66	8	54	89% (82–96%)	55% (46–64%)	0.516 (0.429–0.602) [7 =]	0.680
16 Primary care + NRS + PIS	≥ 1 F25 + ≥1 G40–41, R56.8 + ≥1 AED (NL)	194	66	8	54	89% (82–96%)	55% (46–64%)	0.516 (0.429–0.602) [7 =]	0.680
17 Primary care + SMR01 + NRS	≥ 1 F25 + ≥1 G40–41, R56.8	115	48	6	72	89% (81–97%)	40% (31–49%)	0.381 (0.296–0.466) [15]	0.552
18 Primary care + NRS + PIS	≥ 1 F25 + ≥1 G40–41 + ≥1 AED (BL)	188	61	8	59	88% (81–96%)	51% (42–60%)	0.477 (0.390–0.563) [11 =]	0.646
19 Primary care + NRS + PIS	≥ 1 F25 + ≥1 G40–41 + ≥1 AED (NL)	187	61	8	59	88% (81–96%)	51% (42–60%)	0.477 (0.390–0.563) [11 =]	0.646
20 Primary care + NRS + PIS	≥ 1 F25 + ≥1 G40 + ≥1 AED (BL)	184	60	8	60	88% (81–96%)	50% (41–59%)	0.469 (0.382–0.555) [13 =]	0.638
21 Primary care + NRS + PIS	≥ 1 F25 + ≥1 G40 + ≥1 AED (NL)	183	60	8	60	88% (81–96%)	50% (41–59%)	0.469 (0.382–0.555) [13 =]	0.638
22 Primary care + SMR01 + NRS	≥ 1 F25 + SMR01 ≥ 1 G40–41, R56.8 + NRS G40–41	111	45	6	75	88% (79–97%)	38% (29–46%)	0.357 (0.273–0.441) [16]	0.526
23 Primary care + SMR01 +NRS	≥ 1 F25 + SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 G40	107	44	6	76	88% (79–97%)	37% (28–45%)	0.349 (0.266–0.432) [17]	0.518
24 SMR01 + NRS + PIS	SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 G41 + ≥1 AED (NL)	73	X	X	X	88% (76–99%)	5% (3–6%)	0.050 (NA) [23]	0.095

(*continued on next page*)

Table 3 (*continued*)

Data Base (ranked in study's original accuracy order[1])	Coding algorithm	N	TP	FP	FN	PPV	Sens	CSI (95% CI) [ranking by CSI]	F
25 Primary care +SMR01 + PIS	≥ 1 F25 + ≥1 G40–41, R56.8 + ≥1 AED (BL)	577	80	11	40	85% (78–92%)	71% (63–79%)	0.611 (0.527–0.694) [2]	0.758
26 SMR01 + NRS + PIS	SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 R56.8 + ≥1 AED (BL)	175	57	11	557	84% (75–93%)	9.3% (7–12%)	0.091 (0.069–0.114) [18]	0.167
27 SMR01 + NRS + PIS	SMR01 ≥ 1 G40–41, R56.8 + NRS ≥ 1 G41 + ≥1 AED (BL)	96	29	11	585	73% (59–86%)	5% (3–6%)	0.046 (0.030–0.063) [24]	0.089

Abbreviations: SMR – Scottish Morbidity Record; PIS – Prescription Information Service; NRS – National Records of Scotland; AED – antiepileptic drug; F25 – primary care diagnostic Read Codes for epilepsy; G40–41 – International Classification of Disease 10 (ICD-10) codes for epilepsy and status epilepticus; R56.8 – ICD-10 code for seizures; NL – AEDs on the narrow list; BL – AEDs on the broad list; TP – true positive; FP – false positives; FN – false negatives, PPV – positive predictive value; Sn – sensitivity; CI – confidence intervals; CSI – critical success index; F – F value.

Key: X – categories with five or less events hidden to protect patient identity.

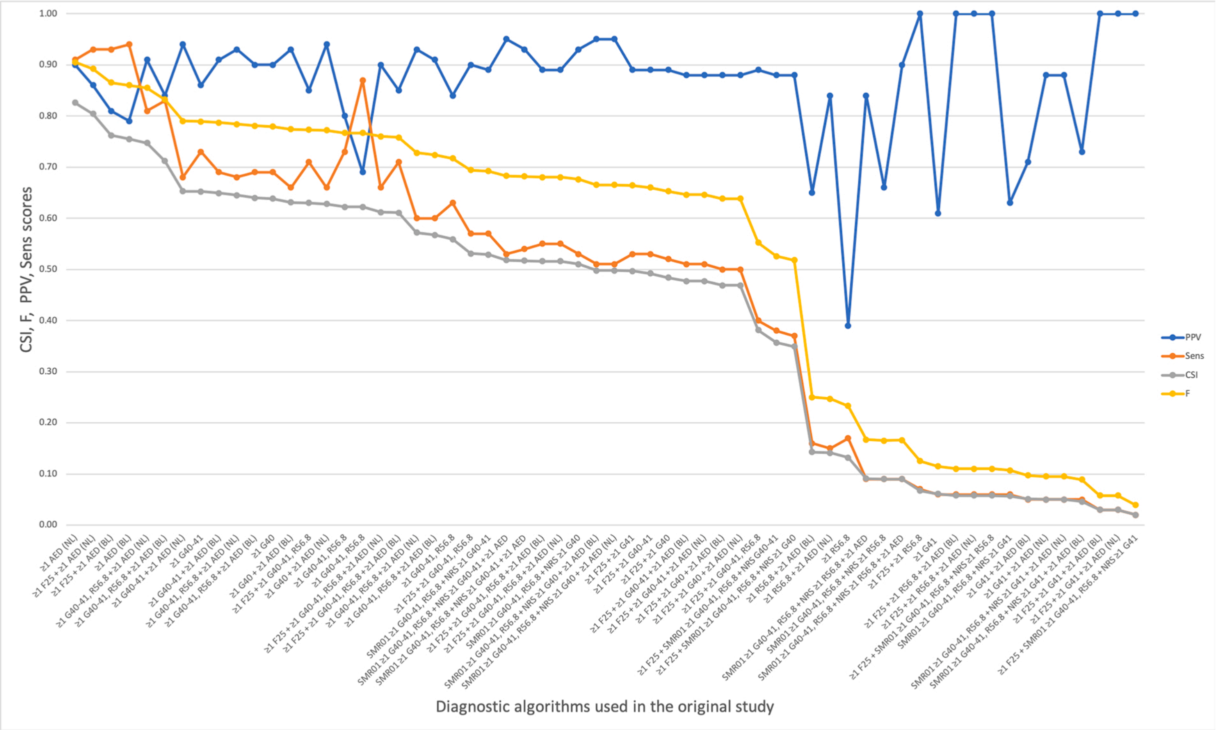

Fig. 1. **:** Dotted line plot of CSI, PPV and sensitivity estimates across the diagnostic study algorithms **Abbreviations:** CSI = Critical success index; PPV = Positive predictive value; Sens = sensitivity; F – F value; SMR – Scottish Morbidity Record; NRS – National Records of Scotland; AED – antiepileptic drug; F25 – primary care diagnostic Read Codes for epilepsy; G40–41 – International Classification of Disease 10 (ICD-10) codes for epilepsy and status epilepticus; R56.8 – ICD-10 code for seizures; NL – AEDs on the narrow list; BL – AEDs on the broad list.

the probability of a positive diagnosis (P), as well as by test threshold, the probability of a positive test (Q):

$$\mathrm{CSI}=1/[(P+Q)/\mathrm{Sens}\cdot P]-1$$

$$=1/[(P+Q)/\mathrm{PPV}\cdot Q]-1$$

F values are also based on the same base data (Larner, in press):

$$F=2\cdot TP/(2\cdot TP+FP+FN)$$

Or from PPV and sensitivity values:

$$F=2\cdot \mathrm{PPV}\cdot \mathrm{Sens}/(\mathrm{PPV}+\mathrm{Sens})$$

This relation of F to PPV means that F is affected by P, as well as Q:

$$F=2\cdot \mathrm{Sens}\cdot P/(P+Q)$$

$$=2\cdot \mathrm{PPV}\cdot Q/(P+Q)$$

The monotonic relation between CSI and F is such that:

$$F=2\mathrm{CSI}/(1+\mathrm{CSI})\cdot$$

3. Results

CSI and F values for each of the 60 algorithms for which calculations could be made were plotted along with the corresponding PPV and sensitivity values (Fig. 1). The plot shows CSI scores were conservative (range 0.02–0.826), always less than or equal to the lower of the corresponding PPV (range 39–100%) and sensitivity (range 2–93%). Unlike CSI, F values were less conservative (range 0.039–0.905), sometimes higher than either PPV or sensitivity, but were always higher than CSI. Low CSI and F values occurred when there was a large difference between PPV and sensitivity, e.g. CSI was 0.02 and F was 0.039 in an instance when PPV was 100% and sensitivity was 2%. Algorithms with both high PPV and sensitivity performed best in terms of CSI and F measure, e.g. CSI was 0.826 and F was 0.905 in an instance when PPV was 90% and sensitivity was 91%.

Results were further considered according to the three levels of validation published in the original study (Mbizvo et al., 2020a).

3.1. Level 1 validation results assessing nine algorithms (Table 1)

CSI and F values could not be calculated for one of these algorithms because of the absence of information on FN instances and hence sensitivity. Amongst the remaining eight level 1 algorithms, the optimal coding strategy by PPV was the narrow list of AEDs from the PIS dataset, and this remained the optimal algorithm according to CSI values (= 0.826). The algorithm using the broad AED list rose from fifth ranking by PPV to second by CSI, displacing the NRS death records using either G40 (epilepsy) alone or G40–41 (epilepsy and/or status epilepticus) codes, whose 2nd and 3rd overall ranking by PPV swapped places to 4th and 3rd respectively by CSI.

3.2. Level 2 validation results assessing 25 algorithms (Table 2)

CSI and F values could be calculated for all 25 algorithms in level 2 validation. The optimal coding strategy by PPV, combining *F25* epilepsy Read codes in primary care with R56.8 seizures codes within the NRS causes of death, dropped to 22nd of 25 on the ranking by CSI values (= 0.067) as a consequence of its very low sensitivity (7%).

The algorithm with the (joint) highest sensitivity (93%), combining *F25* epilepsy Read codes from primary care with AEDs in the narrow list, became the new optimal coding strategy according to CSI (= 0.804) values, rising from 16th of 25 based on PPV (86%).

3.3. Level 3 validation results assessing 27 algorithms (Table 3)

CSI and F values could be calculated for all 27 algorithms in level 3 validation. The top six ranked algorithms all achieved maximal PPV but had very low sensitivity (range 2–6%), and hence all dropped in the ranking according to CSI values (= 0.02–0.058) to no better than ≥ 20th. The new optimal algorithm was previously 13th of 27 (PPV = 90%), but achieved a CSI of only 0.612.

4. Discussion

The key finding in our study was of substantial changes in the accuracy ranking of the diagnostic algorithms compared to the original rankings in the published study (Mbizvo et al., 2020a). These changes can perhaps be considered objective improvements in the rankings because the original rankings were based on a narrative comparison of PPV and sensitivity magnitudes against one another, as is frequently done in diagnostic accuracy studies of administrative data (Horrocks et al., 2017; Kee et al., 2012; Mbizvo et al., 2020b; Wilkinson et al., 2018). Algorithms with high PPV but low sensitivity were ranked lower using CSI and F as outcome measures than algorithms with both high PPV and sensitivity. This is because CSI or F prioritise instances where both PPV and sensitivity are high over instances where there are large differences between PPV and sensitivity (even if one of these is very high). This may potentially allow diagnostic accuracy thresholds based on combined PPV and sensitivity to be determined in future, e.g. CSI ≥ 0.80 (Mbizvo et al., 2023).

The findings of our study showed that the choice of “optimal” algorithm is influenced by the outcome measure used. Reasons for selecting or privileging PPV, sensitivity (or indeed NPV or specificity) as outcome measures have been discussed (Chubak et al., 2012). We understand that investigators may wish to prioritise PPV or sensitivity depending on the use case and the relative cost of false positives and false negatives for various applications. However, using either CSI or F values to complement these measures presents the opportunity to select the best balance of both PPV and sensitivity, and in a manner that can be standardised between studies. As there is a monotonic relation between CSI and F their rankings will always be the same, so there is no a priori reason to choose one measure over the other. However, our preference is for CSI since it is easy to calculate from the base data, is easily understood in terms of signal detection theory, and is a more conservative measure than F. Moreover, various shortcomings of the F measure have been described (Powers, 2015).

To our knowledge, this is the first study to use CSI scores and F measures to assist in establishing the diagnostic accuracy of administrative epilepsy data. These measures were proposed because they combine information from both PPV and sensitivity, the most commonly reported measures of diagnostic accuracy in studies validating administrative epilepsy data (Mbizvo et al., 2020b), thereby standardising their interpretation. We illustrate how it is easier to rank diagnostic algorithms in order of their relative accuracy when using unitary CSI or F measures. Like PPV and sensitivity, CSI and F values avoid the risk of very high values of NPV and specificity consequent upon large numbers of TN instances by eschewing this number. Like PPV, CSI and F values are inherently affected by disease prevalence. An alternative would be to use the Gilbert Skill score (GS) (Gilbert, 1884), another metric used in weather forecasting thought to be less biased because it is less affected by rare events (Space Weather Prediction Center, 2022). However, we have shown there is a monotonic relation between CSI and GS (Larner, 2021) that, in practice, leads in little overall difference in conclusions between the two metrics in epilepsy literature (Mbizvo et al., 2023). Area under the Receiver Operating Characteristic curve (AUC) would be another unitary metric to consider (Metz, 1978). This measures the ability of a test to discriminate whether a specific condition is present or not present. However, its calculation is dependent on having data available on specificity measured from TN instances. TNs and specificity are seldom reported in diagnostic accuracy studies of administrative epilepsy data (Mbizvo et al., 2020b). This is likely due to challenges researchers face in gathering all of the clinical information needed to confirm that cases without a coded diagnosis in the available administrative dataset are truly negative everywhere else in the health record (e.g. in primary care or specialist records that may be difficult to access). Therefore, AUC is rarely used in diagnostic accuracy studies of administrative epilepsy data (Mbizvo et al., 2020b). The wide availability of PPV and sensitivity in such studies leaves room to consider using CSI or F instead.

We suggest that CSI and/or F metrics should be further explored in diagnostic accuracy studies of administrative epilepsy data, as well more broadly in any screening or test accuracy studies involving people with epilepsy, e.g. those assessing the accuracy of electroencephalography. For example, ranking by CSI or F value may have been helpful in the largest systematic review of diagnostic accuracy studies of administrative epilepsy data (Mbizvo et al., 2020b), as the researchers identified the optimal diagnostic algorithms by ranking them in order of PPV and sensitivity and making a judgement on which had the best balance of both high PPV and sensitivity, selecting an arbitrary threshold of > 80% to represent accuracy. It was also difficult to identify the optimal diagnostic algorithms by NPV and specificity in that systematic review as these were nearly 100% across most studies due to very high numbers of TN, often far outnumbering TP, FP, and FN. Another systematic review of diagnostic accuracy studies of administrative epilepsy data took a similar approach of narrative comparison of PPV and sensitivity balances (Kee et al., 2012), as did a systematic review of administrative dementia data (Wilkinson et al., 2018), and a systematic review of administrative motor neurone disease data (Horrocks et al., 2017).

Ranking by CSI or F values may have been helpful in these reviews. Additional ways to explore CSI and F measure in future might be to consider whether both track predominantly with sensitivity, much more so than PPV, and whether this relationship changes with prevalence. To examine this question, researchers could use the equations expressing CSI and F in terms of P and substitute in different values of P, or perhaps calculate rescaled PPVs for different values of P using Bayes equation and thence calculate rescaled CSI and F values. We have not examined these possibilities here as they fall beyond the scope of the current study.

Funding

GKM is supported by an NIHR Clinical Lectureship (CL-2022–07–002). The funders played no role in the design or conduct of this work.

Declarations of interest

None of the authors have any conflict of interests to disclose.

Acknowledgement

We are grateful to Professor Ian Buchan and Dr Glen Martin for providing strategic advice on this work.

Approvals

This study is a secondary analysis of publicly available data from a published study (Mbizvo et al., 2020a), and the data used are available from the link provided in the availability of data and materials, which has been shared in the published study using a CCBY license. The published study generating these data (Mbizvo et al., 2020a) was approved by South East Scotland Research Ethics Committee (REC) 2 (IRAS 181131, 15/SS/0165), and Scottish Public Benefit and Privacy Panel for Health and Social (PBPP). All methods were carried out in accordance with relevant ethical guidelines and regulations and all experimental protocols were approved by South East Scotland REC2 and PBPP. Informed consent was not applicable as it was a study of routinely-collected healthcare datasets that were centrally anonymised and made available to approved researchers by Information Services Division (ISD) Scotland through The Farr Institute.

References

Bossuyt, P.M., et al., 2015. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clin. Chem. 61 (12), 1446–1452.

Chubak, J., Pocobelli, G., Weiss, N.S., 2012. Tradeoffs between accuracy measures for electronic health care data algorithms. J. Clin. Epidemiol. 65 (3), 343–349 e2.

Doswell, Charles A., Davies-Jones, Robert, Keller, David L., 1990. On summary measures of skill in rare event forecasting based on contingency tables. Weather Forecast. 5 (4), 576–585.

Gerapetritis, Harry and Pelissier, J.M. (2004), 'The critical success index and warning strategy', *17th Conference on Probablity and Statistics in the Atmospheric Sciences, Seattle*.

Gilbert, G.K., 1884. Finley's tornado predictions. Am. Meteor. J. 1, 166–172.

Hicks, S.A., et al., 2022. On evaluation metrics for medical applications of artificial intelligence. Sci. Rep. 12 (1), 5979.

Horrocks, S., et al., 2017. Accuracy of routinely-collected healthcare data for identifying motor neurone disease cases: a systematic review. PLoS One 12 (2), e0172639.

Jolliffe, Ian T., 2016. The Dice co-efficient: a neglected verification performance measure for deterministic forecasts of binary events. Meteorol. Appl. 23.

Kee, V.R., et al., 2012. A systematic review of validated methods for identifying seizures, convulsions, or epilepsy using administrative and claims data. Pharmacoepidemiol. Drug Saf. 21, 183–193.

Larner, A.J., 2021. Assessing cognitive screeners with the critical success index. Prog. Neurol. Psychiatry 25 (3), 33–37.

Larner, A.J., 2023. The 2×2 Matrix: Contingency, Confusion And The Metrics of Binary Classification, Second edn. Springer, Cham, Switzerland ([in press]).

Mbizvo, G.K., et al., 2020a. Validating the accuracy of administrative healthcare data identifying epilepsy in deceased adults: a Scottish data linkage study. Epilepsy Res 167, 106462.

Mbizvo, G.K., et al., 2020b. The accuracy of using administrative healthcare data to identify epilepsy cases: a systematic review of validation studies. Epilepsia 61 (7), 1319–1335.

Mbizvo, G.K., et al., 2023. 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. Epilepsia.

Mbizvo, G.K., Larner, A.J., 2022. Isolated headache is not a reliable indicator for brain cancer. Clin. Med (Lond.) 22 (1), 92–93.

Metz, C.E., 1978. Basic principles of ROC analysis. Semin Nucl. Med 8 (4), 283–298.

Palmer, W.C., Allen, R.A., 1949. Note on the accuracy of forecasts concerning the rain problem. U.S. Weather Bureau manuscript, Washington, DC.

Powers, David M.W. (2015), 'What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes', arXiv:1503.06410. ⟨https://ui.adsabs.harvard.edu/abs/2015arXiv150306410P⟩, accessed March 01, 2015.

Schaefer, J.T., 1990. The critical success index as an indicator of warning skill. Weather Forecast 5, 570–575.

Space Weather Prediction Center (2022) *Forecast Verification Glossary* [online text], National Oceanic and Atmospheric Administration ⟨https://bit.ly/3A7BchD⟩.

Spyrou, C., et al., 2020. Implementation of a nowcasting hydrometeorological system for studying flash flood events: the case of Mandra, Greece. Remote Sens. 12 (17).

Wang, H., et al., 2021. Relations among sensitivity, specificity and predictive values of medical tests based on biomarkers. Gen. Psychiatr. 34 (2), e100453.

Wilkinson, T., et al., 2018. Identifying dementia cases with routinely collected health data: a systematic review. Alzheimers Dement 14 (8), 1038–1051.

World Meteorological Organization (2014) *Forecast Verification for the African Severe Weather Forecasting Demonstration Projects; No. 1132* [online text], World Meteorological Organization ⟨https://library.wmo.int/doc_num.php?explnum_id=7868⟩.