EXTRA Dementia and Geriatric Cognitive Disorders

Dement Geriatr Cogn Disord Extra 2019;9:389–396
DOI: 10.1159/000503026
Received: August 27, 2019
Accepted: August 29, 2019
Published online: December 20, 2019

Research Article

Applying Kraemer's Q (Positive Sign Rate): Some Implications for Diagnostic Test Accuracy Study Results

Andrew J. Larner

Cognitive Function Clinic, Walton Centre for Neurology and Neurosurgery, Liverpool, UK

Keywords

Dementia · Diagnosis · Kraemer's Q · Level of test · Mini-Addenbrooke's Cognitive Examination · Mild cognitive impairment · Screening · Sensitivity and specificity

Abstract

Background/Aims: Sensitivity and specificity (Sens, Spec) are not invariant properties of diagnostic and screening tests, but vary in different patient samples. Kraemer [Evaluating medical tests. Objective and quantitative guidelines. 1992] used the level of test, Q, also known as "positive sign rate" (sum of true and false positives divided by sample size), to calculate quality sensitivity and specificity (QSN, QSP). These scaled indices may be more comparable across different patient samples, but have been little studied hitherto. ***Methods:*** The dataset of a pragmatic test accuracy study of the Mini-Addenbrooke's Cognitive Examination (MACE) was re-interrogated to calculate values of QSN and QSP and other paired and unitary test outcome measures based on them, and comparison was made with outcomes previously calculated by standard methods. ***Results:*** QSN and QSP values in this cohort (n = 755; overall prevalence of dementia and mild cognitive impairment [MCI] 0.15 and 0.29, respectively) were inferior to Sens and Spec, as were all other outcome measures for MACE for the diagnosis of both dementia and MCI. QSN was relatively preserved, indicating the sensitivity of MACE. ***Conclusion:*** Indices of test outcome scaled according to Kraemer's Q, the positive sign rate, are less impressive than outcomes calculated by standard methods. These discrepancies may have implications for test evaluation.

A.J. Larner
Cognitive Function Clinic
Walton Centre for Neurology and Neurosurgery, Lower Lane
Fazakerley, Liverpool L9 7LJ (UK)
E-Mail a.larner@thewaltoncentre.nhs.uk

Dement Geriatr Cogn Disord Extra 2019;9:389–396
DOI: 10.1159/000503026

Introduction

Sensitivity and specificity (Sens and Spec) are standard outcome measures in diagnostic or screening test accuracy studies. Yerushalmy [1] introduced these terms to denote respectively the inherent ability of a test to detect correctly a condition when it is present and to rule it out correctly when it is absent, and hence to promote understanding of the utility of diagnostic tests. Test accuracy studies generally present results in a 2 × 2 table cross-classifying all patients (*N*) by test outcome (dichotomised, sometimes using a cut-off value) and by reference standard (disease present or absent) into four categories: true positive (TP), false positive (FP), false negative (FN) and true negative (TN), such that:

Sens = TP / (TP + FN)
Spec = TN / (FP + TN).

Sens is sometimes known as hit rate, TP rate, or recall. Spec is sometimes known as TN rate. Guidelines for papers reporting diagnostic test accuracy studies recommend that Sens and Spec be amongst the keywords, both generally [2] and in the context of dementia [3].

Sens and Spec were once thought to be invariant intrinsic test properties, independent of study sample and location. However, it is now recognised that heterogeneity of clinical populations imposes potentially serious limitations on the utility of Sens and Spec measures, since very different values may be found, for example in different patient subgroups within the sampled population [4]. Moreover, Sens and Spec are "uncalibrated measures of test quality... with a variable zero-point and scale" [5, p. 65].

To try to address this issue, Kraemer [5] developed a metric, Q, the level of a test, given by the sum of true and false positives divided by sample size, or Q = (TP + FP) / *N*. This is also known as the "positive sign rate" or the probability of a positive test in the patient population, with Q' representing 1 – level (= 1 – Q) or the probability of a negative test in the patient population. From these values, quality sensitivity and specificity (QSN, QSP, respectively) values may be calculated, which give the increment in each parameter beyond the level, such that:

QSN = (Sens – Q) / Q'
QSP = (Spec –Q') / Q.

It has been suggested that these calibrated, rescaled, or standardised indices of test parameters are more comparable across different samples [6]. However, to the author's knowledge, QSN and QSP have seldom, if ever, been used explicitly in dementia diagnostic or screening test accuracy studies, despite their potential implications for understanding test utility.

The purpose of this study was to examine QSN and QSP in comparison to Sens and Spec in a large dataset from a screening test accuracy study, and also to examine other metrics based on Sens and Spec, both standard paired (likelihood ratios, clinical utility indexes) and unitary (Youden index, diagnostic odds ratio, number needed to diagnose) measures [7], and also some recently described unitary metrics (likelihood to diagnose or misdiagnose [LDM], the summary utility index [SUI] and the number needed for screening utility [NNSU] [8, 9]).

Methods

The dataset of a previously reported pragmatic screening test accuracy study [9] examining the Mini-Addenbrooke's Cognitive Examination (MACE) [10] for screening of dementia and mild cognitive impairment (MCI) was re-interrogated. This single-centre study recruited

Dement Geriatr Cogn Disord Extra 2019;9:389–396
DOI: 10.1159/000503026

consecutive patients (*n* = 755) over a period of 4.5 years (June 2014 – December 2018). Reference standard was criterion diagnosis of dementia or MCI based on the judgment of an experienced clinician using standard diagnostic criteria (DSM-IV; Petersen; respectively). MACE scores were not used in making criterion diagnoses to avoid review bias. Subjects (or their guardians) gave written informed consent and the study protocol was approved by the institute's committee on human research.

From the 2 × 2 table at various MACE cut-off values, hence at different values of Q, values of QSN and QSP were calculated as per the method of Kraemer [5] for the diagnosis of dementia and MCI, respectively, and compared to values of Sens and Spec. Values of QSN and QSP were checked using the equivalences reported by Kraemer [5, pp. 40–41], namely:

$$(Sens - Q)/Q' = (NPV - P') / P,$$

where NPV is the negative predictive value, NPV = TN / (FN + TN); P = the prevalence of disease (P = TP + FN / *N*); and P' = (1 – P); and

$$(Spec - Q') / Q = (PPV - P) / P'$$

where PPV is the positive predictive value, PPV = TP / (TP + FP).

At the MACE cut-off values previously reported to maximise Youden index (Y = Sens + Spec – 1) for the diagnosis of dementia and MCI (respectively ≤20/30 and ≤24/30 [9]), and at the observed value of P in this cohort, the following parameters based on QSN and QSP were derived:

Positive likelihood ratio (LR_Q+) = QSN / (1 – QSP)
Negative likelihood ratio (LR_Q-) = (1 – QSN) / QSP
Positive clinical utility index (CUI_Q+) = QSN × PPV
Negative clinical utility index (CUI_Q-) = QSP × NPV

Also, the following unitary parameters:

Youden index (Y_Q) = QSN + QSP – 1
Diagnostic odds ratio (DOR_Q) = LR_Q+ / LR_Q-
Number needed to diagnose (NND_Q) = 1 / Y_Q.

Also, the following recently described unitary metrics:

Likelihood to diagnose or misdiagnose (LDM_Q) = NNM / NND_Q,

where NNM is the number needed to misdiagnose = 1 / (1 – Acc) [11], where Acc is accuracy, Acc = (TP + TN) / *N*. LDM is ideally >>1.

Summary utility index (SUI_Q) = CUI_Q+ + CUI_Q-
Number needed for screening utility ($NNSU_Q$) = 1 / SUI_Q.

All these values were compared to the standard calculations already performed for this dataset [9]. Likelihood ratios and clinical utility indexes were classified qualitatively using standard classifications [12, 13] and SUI and NNSU using the classification derived previously [8, 9].

Dement Geriatr Cogn Disord Extra 2019;9:389–396
DOI: 10.1159/000503026

Table 1. Diagnosis of dementia: paired measures of discrimination at various MACE cut-offs

Cut-off	Q; Q'	Sensitivity (= recall)	QSN	Specificity	QSP
≤26/30	0.812; 0.188	0.991	0.953	0.219	0.039
≤25/30	0.731; 0.269	0.991	0.968	0.315	0.063
≤24/30	0.650; 0.350	0.982	0.951	0.409	0.091
≤23/30	0.585; 0.415	0.982	0.959	0.485	0.121
≤22/30	0.518; 0.482	0.974	0.945	0.563	0.156
≤21/30	0.448; 0.552	0.947	0.904	0.641	0.198
≤20/30	0.387; 0.613	0.912	0.857	0.707	0.242
≤19/30	0.338; 0.662	0.860	0.788	0.755	0.275
≤18/30	0.289; 0.711	0.798	0.716	0.802	0.314
≤17/30	0.264; 0.736	0.737	0.642	0.821	0.319

To further illustrate the changes in test parameters according to the value of Q, QSN and QSP were calculated at various arbitrarily predetermined values of Q (namely 0.25, 0.5, 0.75). This also permits the calculation of predictive values, PPV_Q and NPV_Q, at the chosen value of Q. Since, from Kraemer's equations:

$$QSN = (Sens - Q) / Q' = (NPV - P') / P,$$

and

$$QSP = (Spec - Q') / Q = (PPV - P) / P',$$

rearranging it follows that:

$$NPV_Q = (QSN \times P) + P'$$

and

$$PPV_Q = (QSP \times P') + P.$$

Results

The study cohort (n = 755; F:M = 352:403, 47% female; median age 60 years) comprised 114 patients who received a criterion diagnosis of dementia (prevalence = 0.15) and 222 with MCI (overall prevalence = 0.29; prevalence in non-dementia cases = 0.35); the remainder (n = 419) were diagnosed with subjective memory complaints.

For the diagnosis of dementia, between the MACE cut-offs of ≤26/30 and ≤17/30, Q ranged from about 0.8 to about 0.25 (Table 1, column 2). Over this range, there was a decline in both Sens (from 0.991 to 0.737) and QSN (from about 0.95 to 0.64), whilst there was an increase in both Spec (from 0.219 to 0.821) and QSP (from <0.1 to nearly 0.32). All values for both QSN and QSP were inferior to those of Sens and Spec at the same MACE cut-off, although QSN approximated Sens more closely (difference <0.1) than QSP approximated Spec. Indeed, Spec was consistently greater than QSP by 0.3–0.5 (Table 1). The MACE cut-off giving the maximal Y_Q for dementia diagnosis was ≤21/30.

Dement Geriatr Cogn Disord Extra 2019;9:389–396
DOI: 10.1159/000503026

Table 2. Diagnosis of dementia: comparison of various test metrics at MACE cut-off ≤20/30 (maximal Youden index [9])

	Sens, Spec methods	Kraemer Q-based methods
Sens, Spec (QSN, QSP)	0.912, 0.707	0.857, 0.242
[PPV, NPV]	[0.356, 0.978]	–, –
LR+, LR–	3.11 (moderate), 0.12 (large)	1.13 (slight), 0.59 (slight)
CUI+, CUI–	0.32 (very poor), 0.69 (good)	0.31 (very poor), 0.24 (very poor)
[Acc]	[0.738]	–
Y	0.619	0.099
DOR	25.1	1.91
NND	1.61 (2)	10.1 (11)
[NNM]	[3.82 (4)]	–
LDM = NNM/NND	2.37 (3)	0.38 (1)
SUI	1.01 (adequate)	0.55 (very poor)
NNSU	0.99 (1; adequate)	1.54 (2; poor)

NND, NNM and NNSU, patient number metrics, were rounded to the next highest integer.

Table 3. MACE QSN, QSP, PPVQ and NPVQ values at differing predetermined values of Q (level of test or "positive sign rate") for the diagnosis of dementia

	Q			
	0.25	0.387 (observed)	0.50	0.75
QSN	0.883	0.857	0.825	0.649
QSP	–0.173	0.242	0.413	0.609
PPV_Q	0.004	0.356	0.502	0.668
NPV_Q	0.982	0.978	0.974	0.947

Table 4. Diagnosis of MCI: paired measures of discrimination at various MACE cut-offs

Cut-off	Q; Q'	Sensitivity (= recall)	QSN	Specificity	QSP
≤26/30	0.780; 0.220	0.977	0.898	0.325	0.134
≤25/30	0.685; 0.315	0.955	0.857	0.458	0.209
≤24/30	0.591; 0.409	0.901	0.758	0.573	0.278
≤23/30	0.515; 0.485	0.815	0.619	0.644	0.309
≤22/30	0.437; 0.563	0.734	0.528	0.721	0.361
≤21/30	0.359; 0.641	0.635	0.431	0.788	0.408
≤20/30	0.293; 0.707	0.541	0.350	0.838	0.447

Using the previously determined optimal MACE cut-off for dementia diagnosis (≤20/30, from maximal Youden index [9]), all calculated metrics based on QSN and QSP were worse than those based on Sens and Spec (Table 2), often considerably so (note worsening classification of LR+, LR–, CUI–, Y, DOR, NND, LDM, SUI, and NNSU, with relative preservation of only QSN and CUI+).

At predetermined values of Q (0.25, 0.5, 0.75), calculated values of PPV_Q and NPV_Q for MACE for the diagnosis of dementia showed that NPV_Q was preserved at all Q values (Table 3).

For the diagnosis of MCI, between the MACE cut-offs of ≤26/30 and ≤20/30, Q ranged from about 0.8 to about 0.3 (Table 4, column 2). Over this range, there was a decline in both

Table 5. Diagnosis of MCI: comparison of various test metrics at MACE cut-off ≤24/30 (maximal Youden index [9])

	Sens, Spec methods	Kraemer Q-based methods
Sens, Spec (QSN, QSP)	0.901, 0.573	0.758, 0.278
[PPV, NPV]	[0.528, 0.916]	–, –
LR+, LR–	2.11 (moderate), 0.17 (large)	1.05 (slight), 0.87 (slight)
CUI+, CUI–	0.48 (poor), 0.52 (adequate)	0.40 (poor), 0.25 (very poor)
[Acc]	[0.686]	–
Y	0.474	0.036
DOR	12.2	1.21
NND	2.11 (3)	27.8 (28)
[NNM]	[3.19 (4)]	–
LDM = NNM/NND	1.51 (2)	0.11 (1)
SUI	1.00 (adequate)	0.66 (poor)
NNSU	1.00 (1; adequate)	1.54 (2; poor)

NND, NNM and NNSU, patient number metrics, were rounded to the next highest integer.

Table 6. MACE QSN, QSP, PPV_Q and NPV_Q values at differing predetermined values of Q (level of test or "positive sign rate") for the diagnosis of MCI

	Q			
	0.25	0.50	0.591 (observed)	0.75
QSN	0.868	0.802	0.758	0.604
QSP	–0.709	0.146	0.278	0.430
PPV_Q	–0.117	0.442	0.528	0.627
NPV_Q	0.954	0.871	0.842	0.741

Sens (from 0.977 to 0.541) and QSN (from about 0.9 to 0.35), whilst there was an increase in both Spec (from 0.325 to 0.838) and QSP (from about 0.1 to nearly 0.5). All values for both QSN and QSP were inferior to those of Sens and Spec at the same MACE cut-off, although QSN generally approximated Sens (difference <0.25) more closely than QSP approximated Spec (generally difference >0.25). The MACE cut-off giving the maximal Y_Q for MCI diagnosis was ≤25/30.

Using the previously determined optimal MACE cut-off for MCI diagnosis (≤24/30, from maximal Youden index [9]), all other paired and unitary metrics calculated on the basis of QSN and QSP were inferior to those calculated on the basis of Sens and Spec (Table 5), sometimes markedly so, with only QSN and CUI+ approximating the standard measures.

At predetermined values of Q (0.25, 0.5, 0.75), calculated values of PPV_Q and NPV_Q for MACE for the diagnosis of MCI showed that NPV_Q was relatively preserved at all Q values (Table 6).

Discussion

This study shows how the use of Kraemer's Q to derive QSN and QSP values may alter the outcomes of a screening test accuracy study. Here, QSN and NPV_Q of MACE for the diagnosis of dementia and MCI were relatively preserved, reflecting the fact that MACE is a very sensitive test for cognitive impairment, but QSP and other paired and unitary parameters based on QSN

Dement Geriatr Cogn Disord Extra 2019;9:389–396
DOI: 10.1159/000503026

and QSP were inferior, sometimes markedly so. This suggests that knowledge of Q, the level of test or the "positive sign rate," is pertinent to test result interpretation.

In this respect it may be worthwhile to consider relationships between Q and P in test accuracy studies. In any individual study, P will be fixed, according to the setting in which the study is performed, be that community, primary, or secondary care. Exact values of accuracy, PPV and NPV will differ in different settings according to the value of P. Based on study values of Sens and Spec, there are simple equations which permit the calculation of accuracy, PPV and NPV for different values of P. This possibility for rescaling is sometimes exploited by researchers to give an indication of test performance in settings other than that of their own study [9, 14, 15].

Q may also be fixed in an individual test accuracy study. For example, if the test being studied has a cut-off value established in a prior index test accuracy study, investigators may adhere to this single cut-off point in the evaluation (and reporting) of their study. Indeed, there are objections to changing test cut-offs: a study may be classified as being at "higher risk of bias if the authors define the optimal cut-off post hoc based on their own study data" [16]. However, individual study cohorts may differ significantly in case mix from those in the index study (e.g., absence of a control population of healthy individuals and greater clinical heterogeneity in phase III studies versus phase I or II studies). A case for revision of cut-offs established in index studies in order to optimise test performance in pragmatic studies which more closely resemble clinical practice has been outlined [17]. Diagnostic test accuracy studies seldom present results for all potential test cut-offs to permit readers to see the effects of different Q values. If this is the case, then the use of Kraemer's simple equations will permit calculation of QSN and QSP at different set values of Q, as well as values of PPV_Q and NPV_Q.

By constructing a table of QSN, QSP, PPV_Q and NPV_Q (and any other parameters derived from them) at predetermined values of Q (as in Tables 3 and 6), test performance at different levels or positive sign rates may be illustrated. These data may be useful in indicating to clinicians appropriate or acceptable levels of Q, and hence test cut-offs, for their clinical purpose, be it to rule in (case finding; high QSN) or rule out (screening; high QSP) cases, or to strike the optimal balance between QSN and QSP by maximising Y_Q. This would be analogous to the process, already familiar to clinicians, of constructing tables of PPV and NPV at predetermined values of P to indicate test performance at different levels of disease prevalence.

Unscaled values of Sens and Spec may potentially overestimate test quality and mislead clinicians [5]. Use of rescaled, calibrated, or standardised test parameters, using Kraemer's methodology [5], may render the outcomes of different studies in different samples more comparable [6]. This approach might also have implications for the optimisation of test cut-offs, and for the inclusion and evaluation of studies in systematic reviews and meta-analyses.

Statement of Ethics

Subjects (or their guardians) gave written informed consent and the study protocol was approved by the institute's committee on human research.

Disclosure Statement

The author has no conflicts of interest to declare.

Funding Sources

Not funded.

Author Contributions

A.J. Larner conceived and planned the study, collected and analysed the data, and drafted the manuscript.

References

1 Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Public Health Rep. 1947 Oct;62(40):1432–49.
2 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al.; Standards for Reporting of Diagnostic Accuracy. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem. 2003 Jan;49(1):7–18.
3 Noel-Storr AH, McCleery JM, Richard E, Ritchie CW, Flicker L, Cullum SJ, et al. Reporting standards for studies of diagnostic test accuracy in dementia: The STARDdem Initiative. Neurology. 2014 Jul;83(4):364–73.
4 Hlatky MA, Mark DB, Harrell FE Jr, Lee KL, Califf RM, Pryor DB. Rethinking sensitivity and specificity. Am J Cardiol. 1987 May;59(12):1195–8.
5 Kraemer HC. Evaluating medical tests. Objective and quantitative guidelines. Newbery Park (California): Sage; 1992.
6 Larrabee GJ, Barry DT. Diagnostic classification statistics and diagnostic validity of malingering assessment. In: Larrabee GJ, editor. Assessment of malingered neuropsychological deficits. Oxford: Oxford University Press; 2007. pp. 14–26.
7 Larner AJ. Diagnostic test accuracy studies in dementia. A pragmatic approach. 2nd ed. London: Springer; 2019. https://doi.org/10.1007/978-3-030-17562-7.
8 Larner AJ. New unitary metrics for dementia test accuracy studies. Prog Neurol Psychiatry. 2019;23(3):21–5.
9 Larner AJ. MACE for diagnosis of dementia and MCI: examining cut-offs and predictive values. Diagnostics (Basel). 2019 May;9(2):E51.
10 Hsieh S, McGrory S, Leslie F, Dawson K, Ahmed S, Butler CR, et al. The Mini-Addenbrooke's Cognitive Examination: a new assessment tool for dementia. Dement Geriatr Cogn Disord. 2015;39(1-2):1–11.
11 Habibzadeh F, Yadollahie M. Number needed to misdiagnose: a measure of diagnostic test effectiveness. Epidemiology. 2013 Jan;24(1):170.
12 Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA. 1994 Mar 2;271(9):703–7.
13 Mitchell AJ. Sensitivity × PPV is a recognized test called the clinical utility index (CUI+). Eur J Epidemiol. 2011 Mar;26(3):251–2.
14 Mathuranath PS, Nestor PJ, Berrios GE, Rakowicz W, Hodges JR. A brief cognitive test battery to differentiate Alzheimer's disease and frontotemporal dementia. Neurology. 2000 Dec;55(11):1613–20.
15 Mioshi E, Dawson K, Mitchell J, Arnold R, Hodges JR. The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening. Int J Geriatr Psychiatry. 2006 Nov;21(11):1078–85.
16 Davis DH, Creavin ST, Noel-Storr A, Quinn TJ, Smailagic N, Hyde C, et al. Neuropsychological tests for the diagnosis of Alzheimer's disease dementia and other dementias: a generic protocol for cross-sectional and delayed-verification studies. Cochrane Database Syst Rev. 2013 Mar;3(3):CD010460.
17 Larner AJ. Optimising the cutoffs of cognitive screening instruments in pragmatic diagnostic accuracy studies: maximising accuracy or the Youden index? Dement Geriatr Cogn Disord. 2015;39(3-4):167–75.