

## Article

# Communicating Risk: Developing an “Efficiency Index” for Dementia Screening Tests

Andrew J. Larner

Cognitive Function Clinic, Walton Centre for Neurology and Neurosurgery, Lower Lane, Fazakerley, Liverpool L9 7LJ, UK; a.larner@thewaltoncentre.nhs.uk

**Abstract:** Diagnostic and screening tests may have risks such as misdiagnosis, as well as the potential benefits of correct diagnosis. Effective communication of this risk to both clinicians and patients can be problematic. The purpose of this study was to develop a metric called the “efficiency index” (EI), defined as the ratio of test accuracy and inaccuracy, to evaluate screening tests for dementia. This measure was compared with a previously described “likelihood to be diagnosed or misdiagnosed” (LDM), also based on “numbers needed” metrics. Datasets from prospective pragmatic test accuracy studies examining four brief cognitive screening instruments (Mini-Mental State Examination; Montreal Cognitive Assessment; Mini-Addenbrooke’s Cognitive Examination (MACE); and Free-Cog) were analysed to calculate values for EI and LDM, and to examine their variation with test cut-off for MACE and dementia prevalence. EI values were also calculated using a modification of McGee’s heuristic for the simplification of likelihood ratios to estimate percentage change in diagnostic probability. The findings indicate that EI is easier to calculate than LDM and, unlike LDM, may be classified either qualitatively or quantitatively in a manner similar to likelihood ratios. EI shows the utility or inutility of diagnostic and screening tests, illustrating the inevitable trade-off between diagnosis and misdiagnosis. It may be a useful metric to communicate risk in a way that is easily intelligible for both clinicians and patients.



**Citation:** Larner, A.J.

Communicating Risk: Developing an “Efficiency Index” for Dementia Screening Tests. *Brain Sci.* **2021**, *11*, 1473. <https://doi.org/10.3390/brainsci11111473>

Academic Editor: Laura Zamarian

Received: 6 October 2021

Accepted: 3 November 2021

Published: 6 November 2021

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** dementia; diagnosis; efficiency index; risk communication; screening test

## 1. Introduction

No medical treatment or test is without potential harms as well as benefits, and hence associated with risk. Communicating such risk to patients for the purpose of shared decision making has attracted much attention and research in recent times, for example into the most appropriate methods by which to achieve such communication effectively. Guidance on both verbal and numerical qualifiers of risk has appeared [1,2]. As regards numerical qualifiers, options often used in the context of therapeutic interventions include absolute risk (AR), relative risk (RR), and the number needed to treat (NNT) [3], but no consensus on the optimum method has been established.

Considering diagnostic or screening tests, performance is typically described by comparison with a reference standard, such as a criterion diagnosis or a reference test, by constructing a  $2 \times 2$  contingency table, such that all (N) index test results may be cross-tabulated as true positive (TP), false positive (FP), false negative (FN), or true negative (TN). From this standard  $2 \times 2$  contingency table (Figure 1), various parameters of test discrimination may be calculated, many of which are familiar to clinicians as descriptors of test performance, such as sensitivity (Sens; or true positive rate) and specificity (Spec; or true negative rate), positive and negative predictive values, and positive and negative likelihood ratios (LR+, LR−) [4].

		Reference Standard	
		Condition present	Condition absent
Index Test Outcome	Positive	True positive [TP]	False positive [FP]
	Negative	False negative [FN]	True negative [TN]

**Figure 1.** Standard  $2 \times 2$  contingency table for diagnostic or screening test accuracy studies and formulae for paired measures.

$$\text{Sensitivity (Sens)} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Specificity (Spec)} = \text{TN}/(\text{FP} + \text{TN})$$

$$\text{Positive predictive value (PPV)} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Negative predictive value (NPV)} = \text{TN}/(\text{FN} + \text{TN})$$

$$\text{Positive likelihood ratio (LR+)} = \text{TP}/(\text{TP} + \text{FN})/\text{FP}/(\text{FP} + \text{TN})$$

$$\text{Negative likelihood ratio (LR-)} = \text{FN}/(\text{TP} + \text{FN})/\text{TN}/(\text{FP} + \text{TN})$$

$$\text{Accuracy (Acc)} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

$$\text{Inaccuracy (Inacc)} = (\text{FP} + \text{FN})/(\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

One may thus distinguish from the  $2 \times 2$  contingency table two conditions or relations between the index test and the reference standard: consistency, or matching, of outcomes (+/+ or TP, and -/- or TN); and contradiction, or mismatching (+/- or FP, and -/+ or FN). From these two conditions, the paired complementary parameters of accuracy ( $\text{Acc} = \text{TP} + \text{TN}/\text{N}$ ) and inaccuracy or error rate ( $\text{Inacc} = \text{FP} + \text{FN}/\text{N}$ ) may be derived. As negations, these may also be described using the Boolean NOT operator, since  $\text{Acc} = 1 - \text{Inacc}$  and  $\text{Inacc} = 1 - \text{Acc}$ .

How might these various test measures be effectively communicated to patients who are unfamiliar with the principles and nomenclature of binary classicism, but worried that they might have a dementia disorder because of memory symptoms? A metric called the “likelihood to be diagnosed or misdiagnosed” (LDM) has been developed [5,6] which may be useful for the purpose of communicating risk, specifically the risk of testing leading to misdiagnosis as opposed to correct diagnosis.

LDM was conceptualised as analogous to the “likelihood to be helped or harmed” (LHH) metric which was developed to communicate the results of therapeutic (randomised controlled) trials. LHH is based on “number needed to” metrics, specifically the number needed to treat (NNT) for a specified treatment outcome (e.g., cure, remission, 50% reduction in symptoms) [7] and the number needed to harm (NNH) [8]. LHH is the ratio of NNH to NNT, which is desirably as large as possible (high NNH, low NNT), thus summarising treatment benefits and risks [9].

LDM for diagnostic and screening tests is the ratio of number needed to misdiagnose (NNM) to number needed to diagnose (NND), where NNM is  $1/\text{Inacc}$  (as defined by Habibzadeh and Yadollahie [10]) and NND is  $1/(\text{Sens} + \text{Spec} - 1)$  or  $1/\text{Youden index (Y)}$  (as defined by Linn and Grunau [11]). LDM is desirably as large as possible (high NNM, low NND) [5,6], thus summarising testing benefits and risks.

The LDM metric has proved serviceable in evaluating a wide range of neurological signs and cognitive screening instruments (CSIs) used in the evaluation of disorders of cognition [5,6,12]. Nevertheless, LDM has some limitations and shortcomings. Consistent with its ad hoc development, based on existing metrics, LDM combined rates with different denominators which are not easily reconciled. Calculation of several parameters from the  $2 \times 2$  table is required to reach LDM (Sens, Spec, Y, NND, Inacc, NNM), although ad hoc

calculators exist [13]. Furthermore, the “number needed to diagnose” based on the Youden index incorporates considerations not only of diagnosis but also of misdiagnosis, since  $\text{Sens} = 1 - \text{false negative rate}$ , and  $\text{Spec} = 1 - \text{false positive rate}$ . The resulting LDM has boundary values of  $-1$  (useless test:  $\text{Sens} = \text{Spec} = 0$ ,  $\text{NND} = -1$ ;  $\text{Inacc} = 1$ ,  $\text{NNM} = 1$ ) and  $\infty$  (perfect test:  $\text{Sens} = \text{Spec} = 1$ ,  $\text{NND} = 1$ ;  $\text{Inacc} = 0$ ,  $\text{NNM} = \infty$ ), and so the LDM values cannot be perfectly equated with the qualitative classification scheme developed for likelihood ratios (LR) [14] which has been used for making recommendations on tests suitable for dementia by the UK National Institute for Health and Care Excellence [15]. Unlike LDM, LR has boundary values of 0 and  $\infty$ , although LDM shares with LR an inflection point at 1 (LDM  $< 1$  favours misdiagnosis, LDM  $> 1$  favours diagnosis) [5,6].

A simple method to overcome these shortcomings of LDM may be proposed. The NND may be redefined, as  $\text{NND}^*$ , using Acc, rather than Sens and Spec, such that  $\text{NND}^* = 1/\text{Acc}$ . This formulation is analogous to the previous definition of NNM [10], where  $\text{NNM} = 1/\text{Inacc}$ . Both these measures now share the same denominator from the  $2 \times 2$  contingency table, N, and calculation is thus simplified, such that:

$$\text{NNM}/\text{NND}^* = (1/\text{Inacc})/(1/\text{Acc})$$

Whilst this ratio might justifiably be termed a “likelihood to be diagnosed or misdiagnosed”, an alternative name would be preferable to avoid confusion with the previously defined LDM. Kraemer denoted TP + TN as “efficiency” [16], so FP + FN might be termed “inefficiency”, and hence the ratio of efficiency/inefficiency may be denoted as the “efficiency index” (EI). Hence:

$$\text{EI} = \text{Acc}/\text{Inacc} \quad (1)$$

$$= (\text{TP} + \text{TN})/(\text{FP} + \text{FN}) \quad (2)$$

The boundary values of EI are 0 (useless test:  $\text{Acc} = 0$ ;  $\text{Inacc} = 1$ ) and  $\infty$  (perfect test:  $\text{Acc} = 1$ ,  $\text{Inacc} = 0$ ), as for likelihood ratios.

The primary aim of this study was to examine the utility of EI and compare it to the previously defined LDM parameter when applied to test accuracy studies of several brief CSIs, namely Mini-Mental State Examination (MMSE) [17], Montreal Cognitive Assessment (MoCA) [18], Mini-Addenbrooke’s Cognitive Examination (MACE) [19], and Free-Cog [20]. Secondary aims were: to examine other methods to calculate EI and to compare performance with LRs, particularly with McGee’s method of simplifying LR values as percentage changes in diagnostic probability [21]; and to compare EI with a previously described measure based on Acc and Inacc, the identification index (II) [22], defined as  $\text{II} = (\text{Acc} - \text{Inacc}) = 2 \text{Acc} - 1$ .

## 2. Methods

Data from pragmatic prospective screening test accuracy studies using a standardised methodology were re-analysed. The studies were undertaken in a dedicated cognitive disorders clinic located in a secondary care setting (regional neuroscience centre) and examined four brief cognitive screening instruments (administration time ca. 5–10 min), all with a denominator of 30 points: Mini-Mental State Examination (MMSE) [23,24], Montreal Cognitive Assessment (MoCA) [25], Mini-Addenbrooke’s Cognitive Examination (MACE) [26], and Free-Cog [27].

In each study, criterion diagnosis of dementia followed standard diagnostic criteria (DSM-IV) and was made independent of scores on index CSIs to avoid review bias. For each study, prevalence of dementia was calculated as the sum of TP and FN divided by the total number of patients (N) assessed. All studies followed either the STAndards for the Reporting of Diagnostic accuracy studies (STARD) [28] or the derived guidelines specific for dementia studies, STARDdem [29], dependent on the exact date at which each study was undertaken. In all studies, subjects gave informed consent and study protocols were approved by the institute’s committee on human research (Walton Centre for Neurology and Neurosurgery Approval: N 310).

For each CSI, the following parameters were calculated: Acc, Inacc, Y, LDM, II; and EI by using the values of Acc and Inacc (Equation (1)).

The variation of EI with test cut-off was examined using data from the test accuracy study of MACE [26] and compared with the variation in LDM. The variation of EI with prevalence of dementia (P) was also examined and compared to values for LDM [12].

EI was also calculated from values of test Sens and Spec in the MACE study. Since:

$$\text{Acc} = \text{Sens} \cdot P + \text{Spec} \cdot (1 - P)$$

and

$$\text{Inacc} = (1 - \text{Sens}) P + (1 - \text{Spec}) \cdot (1 - P)$$

Hence:

$$\text{EI} = \frac{\text{Sens} \cdot P + \text{Spec} \cdot (1 - P)}{(1 - \text{Sens}) P + (1 - \text{Spec}) (1 - P)} \quad (3)$$

The performance of EI was also compared to that of LRs which may be used to calculate difference in pre- and post-test odds, since post-test odds = pre-test odds  $\times$  LR. McGee showed that LR+ values of 2, 5, and 10 increased the probability of diagnosis by approximately 15%, 30%, and 45%, respectively, whereas LR– values of 0.5, 0.2, and 0.1 decreased the probability of diagnosis by approximately 15%, 30% and 45%, respectively. These figures derive from the almost linear relationship of probability and the natural logarithm of odds over the range 0.1–0.9, such that the percentage change in probability may be calculated independent of pre-test probability as:

$$\text{Change in probability} = 0.19 \times \log_e(\text{LR})$$

This simple heuristic obviates calculations between pre- and post-test odds and probabilities [21]. As the boundary values of EI (0,  $\infty$ ) correspond to those of LRs, calculations were undertaken to assess whether or not the heuristic described for LR values also holds for EI values. These calculations used data from the studies of MACE [26] and MoCA [25]. As both of these studies had a similar (low) pre-test probability of dementia, the issue was further examined using data from a test accuracy study of the Test Your Memory (TYM) test [30], this being the study with the highest pre-test probability of dementia reported from this clinic [31] (because an informant is generally required to assist with TYM, and many patients without dementia attend the clinic alone [32]).

### 3. Results

A summary of the studies of MMSE, MoCA, MACE, and Free-Cog (Table 1) showed broadly similar prevalence of dementia, median age, and gender ratio in each patient cohort.

**Table 1.** Study demographics and test thresholds for dementia.

CSI	N	P = Prevalence of Dementia = (TP + FN)/N	Age, Median (years)	Gender (F:M; %F)	Test Threshold for Dementia	Ref(s)
MMSE	244	0.18	60	117:127; 48	<26/30	[23,24]
MoCA	260	0.17	59	118:142; 45	<26/30	[25]
MACE	755	0.15	60	352:403; 47	$\leq$ 20/30	[26]
Free-Cog	141	0.11	62	61:80; 43	$\leq$ 22/30	[27]

Abbreviations: CSI = cognitive screening instrument; TP = true positive; FN = false negative; MMSE = Mini-Mental State Examination; MoCA = Montreal Cognitive Assessment; MACE = Mini-Addenbrooke's Cognitive Examination.

Comparing the various metrics for the diagnosis of dementia versus no dementia for each of these CSIs (Table 2) showed a similar ranking (best to worst) for Acc, Y, LDM, II, and EI.

**Table 2.** Comparing metrics for diagnosis of dementia vs. no dementia by CSI (using cut-offs in Table 1).

CSI	Acc	Inacc	Y (=Sens + Spec – 1)	LDM (=NNM/NND)	II (=2.Acc – 1)	EI (=Acc/Inacc)
MMSE	0.676	0.324	0.497	1.536	0.352	2.089
MoCA	0.427	0.573	0.313	0.547	−0.146	0.745
MACE	0.738	0.262	0.619	2.360	0.475	2.817
Free-Cog	0.709	0.291	0.670	2.320	0.418	2.439

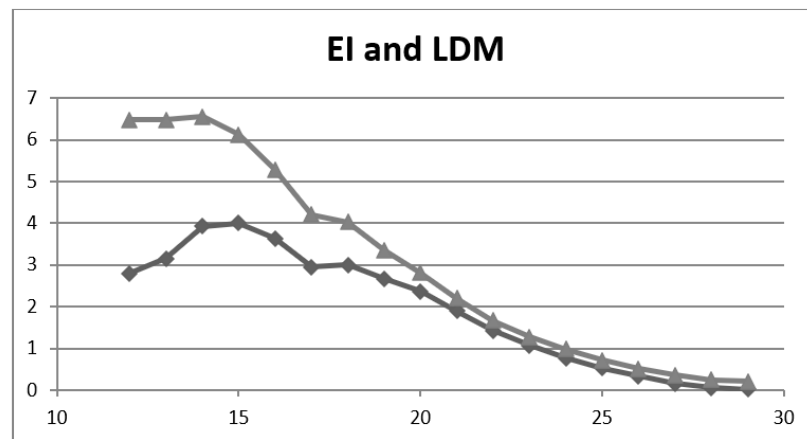
Abbreviations: CSI = cognitive screening instrument; Acc = correct classification accuracy; Inacc = inaccuracy; Y = Youden index; LDM = likelihood to be diagnosed or misdiagnosed; NNM = number needed to misdiagnosis; NND = number needed to diagnose; II = identification index; EI = efficiency index; MMSE = Mini-Mental State Examination; MoCA = Montreal Cognitive Assessment; MACE = Mini-Addenbrooke's Cognitive Examination.

Comparing the EI and LDM metrics across a range of MACE cut-offs (Table 3, Figure 2) showed that, as for the other CSIs (Table 2), EI was a more optimistic score than LDM and that, unlike II, EI nowhere had a negative value. The maxima for EI and LDM almost coincided ( $LDM \leq 15/30$ ;  $EI \leq 14/30$ ). Values for EI and LDM were approximately equal at higher test cut-offs but diverged at lower cut-offs in this dataset, which may be a reflection of high sensitivity and low specificity of MACE for the diagnosis of dementia [26].

**Table 3.** Diagnosis of dementia: comparing metrics at various MACE cut-offs.

Cut-Off	Acc	Inacc	Y	LDM	II	EI
$\leq 29/30$	0.170	0.830	0.02	0.02	−0.66	0.204
$\leq 28/30$	0.197	0.803	0.05	0.06	−0.61	0.246
$\leq 27/30$	0.262	0.738	0.12	0.16	−0.48	0.355
$\leq 26/30$	0.336	0.664	0.21	0.33	−0.33	0.507
$\leq 25/30$	0.417	0.583	0.31	0.53	−0.17	0.716
$\leq 24/30$	0.495	0.505	0.39	0.76	−0.01	0.982
$\leq 23/30$	0.560	0.440	0.47	1.07	0.12	1.27
$\leq 22/30$	0.625	0.375	0.53	1.43	0.25	1.67
$\leq 21/30$	0.687	0.313	0.59	1.90	0.37	2.20
$\leq 20/30$	0.738	0.262	0.62	2.36	0.48	2.82
$\leq 19/30$	0.771	0.229	0.61	2.67	0.54	3.36
$\leq 18/30$	0.801	0.199	0.60	3.00	0.60	4.03
$\leq 17/30$	0.808	0.192	0.56	2.95	0.62	4.21
$\leq 16/30$	0.841	0.159	0.58	3.63	0.68	5.29
$\leq 15/30$	0.860	0.140	0.56	4.00	0.72	6.12
$\leq 14/30$	0.868	0.132	0.51	3.92	0.74	6.55
$\leq 13/30$	0.866	0.134	0.41	3.15	0.73	6.48
$\leq 12/30$	0.866	0.134	0.37	2.79	0.73	6.48

Abbreviations: Acc = correct classification accuracy; Inacc = inaccuracy; Y = Youden index; LDM = likelihood to be diagnosed or misdiagnosed; II = identification index; EI = efficiency index.

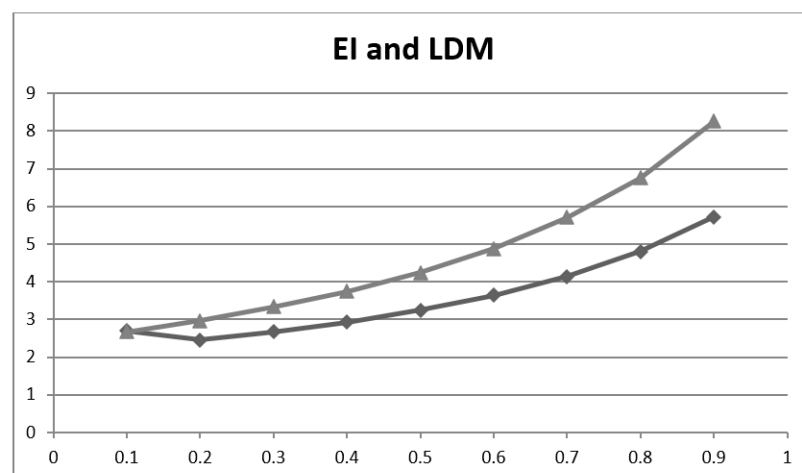


**Figure 2.** Plot of efficiency index (EI; upper line, triangles) and of likelihood to be diagnosed or misdiagnosed (LDM; lower line, diamonds) values (y axis) vs. MACE cut-off score (x-axis).

Comparing the EI and LDM metrics across a range of P (Table 4, Figure 3) showed that values increasingly diverge at higher prevalence in this dataset.

**Table 4.** EI and LDM values of MACE for dementia diagnosis at various prevalence levels at fixed test cut-off ( $\leq 20/30$ ).

P, P'	Acc	Inacc	LDM (=NNM/NND =Y/Inacc)	EI (=NNM/NND* =Acc/Inacc)
0.1, 0.9	0.728	0.272	2.70	2.68
0.2, 0.8	0.748	0.252	2.45	2.97
0.3, 0.7	0.768	0.232	2.67	3.33
0.4, 0.6	0.789	0.211	2.93	3.74
0.5, 0.5	0.809	0.191	3.25	4.24
0.6, 0.4	0.830	0.170	3.64	4.88
0.7, 0.3	0.851	0.149	4.14	5.71
0.8, 0.2	0.871	0.129	4.80	6.75
0.9, 0.1	0.892	0.108	5.72	8.26



**Figure 3.** Plot of efficiency index (EI; upper line, triangles) and of likelihood to be diagnosed or misdiagnosed (LDM; lower line, diamonds) values (y axis) vs. prevalence of dementia (x-axis).

Using Equation (1), the value of EI for MACE was 2.817 (Table 2). This value was checked by using Equation (3), substituting the values for dementia prevalence ( $P = 0.151$ ) and Sens and Spec at the test threshold for MACE ( $\leq 20/30$ ; Table 1), respectively 0.912 and 0.707. Hence,

$$\begin{aligned} EI &= (0.912 \times 0.151) + (0.707 \times 0.849) / (0.088 \times 0.151) + (0.293 \times 0.849) \\ &= 0.738 / 0.262 \\ &= 2.817 \end{aligned}$$

The same value of EI was thus obtained using two different methods.

To examine whether or not McGee's simple rules obviating calculations between pre- and post-test odds and probabilities for LRs [21] are also applicable to EIs, data from the MACE study were used [26], wherein:

$$\text{Dementia prevalence} = 114/755 = 0.151 = \text{pre-test probability}$$

$$\text{Pre-test odds} = \text{pre-test probability} / (1 - \text{pre-test probability}) = 0.1778$$

It is known that:

$$\text{Post-test odds} = \text{pre-test odds} \times \text{LR}$$

Substituting EI for LR, let:

$$\text{Post-test odds} = \text{pre-test odds} \times \text{EI}$$

In the MACE study,  $\text{EI} = 2.817$ , favouring correct diagnosis. Hence:

$$\text{Post-test odds} = 0.1778 \times 2.817 = 0.500$$

$$\text{Post-test probability} = \text{post-test odds} / (1 + \text{post-test odds}) = 0.33$$

So, using calculations based on the observed pre-test probability, MACE increased diagnostic probability of dementia in this patient cohort from approximately 15% to approximately 33%, an 18% increase.

Using the equation derived by McGee to calculate change in diagnostic probability independent of pre-test probability [21], and substituting LR with EI:

$$\begin{aligned} \text{Change in probability} &= 0.19 \times \log_e(\text{EI}) \\ &= 0.19 \times \log_e(2.817) \\ &= 0.197 \end{aligned}$$

Thus similar values for percentage change in probability were obtained using two different methods (18% vs. 19.7%).

Calculations of change in diagnostic probability were performed assuming EI values of 2.0, 5.0, 10.0, 0.5, 0.2, and 0.1 using both the methods, i.e., dependent and independent of observed pre-test probability. Similar calculations were also performed using the MoCA test accuracy study data [25], in which pre-test probability was similar to MACE but EI was 0.745 (i.e., favouring misdiagnosis). The results (Table 5) show that for EI values  $> 1$ , favouring correct diagnosis, the percentage changes in diagnostic probability were similar when calculated independent of pre-test probability (column 1) and when calculated using observed pre-test probabilities (columns 2 and 3), approximating McGee's 15, 30, and 45% increases for EI = 2, 5 and 10. However, for EI values  $< 1$ , favouring misdiagnosis, McGee's 15, 30, and 45% decreases for EI 0.5, 0.2 and 0.1 were not observed, presumably because of the low pre-test probabilities (ca. 15%) in these patient cohorts. In the S-shaped curve which describes the relationship between probability and  $\log_e$  odds [21], this may correspond to the part of the plot away from the nearly linear portion which runs from approximately 0.1 to approximately 0.9.

**Table 5.** EI values calculated independent of pre-test probability and for selected cognitive screening instruments based on pre-test probability observed in test accuracy studies.

EI	% Change in Diagnostic Probability Calculated Independent of Pre-Test Probability as $0.19 \times \log_e(\text{EI})$	MACE: % Change in Diagnostic Probability Based on Pre-Test Probability (0.15) [26]	MoCA: % Change in Diagnostic Probability Based on Pre-Test Probability (0.17) [25]	TYM: % Change in Diagnostic Probability Based on Pre-Test Probability (0.35) [31]
10.0	+43.7	+49	+54	+49
5.0	+30.5	+32	+34	+38
4.882 (TYM)	+30.1	-	-	+37
2.817 (MACE)	+19.7	+18	-	-
2.0	+13.2	+11	+12	+17
1.0	0	0	0	0
0.745 (MoCA)	-5.6	-	-4	-
0.5	-13.2	-7	-8	-14
0.2	-30.5	-12	-13	-25
0.1	-43.7	-13	-15	-30

To further examine this point, data from a test accuracy study of the Test Your Memory (TYM) test [31] were reanalysed, wherein the pre-test probability of dementia was 0.35, the highest reported in studies from this clinic. The change in probability based on the observed pre-test probability (Table 5, column 4) approximated the values calculated independent of pre-test probability (column 1) more closely than for MACE and MoCA for EI values < 1.

#### 4. Discussion

This paper defines a new parameter, EI, which may be of use not only for the evaluation of diagnostic and screening tests but also for the communication of risk to both clinicians and patients. The examples presented illustrate how EI may be used in a clinical setting. EI may be applied in any test accuracy study which permits construction of a  $2 \times 2$  contingency table.

EI may be conceptualised, like a previously defined LDM parameter, as a ratio of test harms (misdiagnosis) and benefits (diagnosis), and hence a measure of what has previously been termed the “fragility” of screening and diagnostic tests [33]. Indeed, it might have been named the “fragility index”, understood as a propensity to break or fail. However, “efficiency index” emphasises its relation to efficiency, understood as the ability to do things well, and conceptualised as a ratio of useful output to total input, or product per cost. However, EI differs from efficiency in that efficiency always has a value < 1, whereas the upper bound of the EI, for a perfect diagnostic or screening test, is  $\infty$ .

Comparing EI to LDM, EI has kinship with, but advantages over, LDM. It is more easily explicable (and, hence, more elegant) than the makeshift (although not arbitrary) derivation of LDM. EI is easier to calculate than LDM, requiring at its simplest only the four values from the cells of the  $2 \times 2$  contingency table (Equation (2)), whilst retaining the inflection point at the value of 1 (EI > 1 indicates greater likelihood of correct diagnosis; EI < 1 indicates greater likelihood of misdiagnosis).

EI and LDM share the same denominator, Inacc, but differ in numerator, Acc and Y, respectively, these numerators being the multiplicative inverse of NND\* vs. NND, respectively. Interpreting the results (Table 3, Figure 2), the differences in EI and LDM values thus relate to the higher values of Acc compared to Y (lower values of NND\* compared to NND), especially at lower cut-off values.

Comparing EI to LRs, whilst both can be calculated from the values of Sens and Spec, and share boundary values (0,  $\infty$ ), EI is dependent on prevalence (Equation (3)) whereas LRs are not (at least algebraically, although in practice there may be variation [34]). Based on the



calculations (Table 5) it seems appropriate to use the same qualitative classification scheme for EI values as proposed by Jaeschke et al. for LRs [14] (Table 6, column 1). Furthermore, McGee's simplification of LRs, as percentage change in diagnostic probability [21], also appears to be applicable to EI values (Table 5), and hence this numerical classification might also be used (Table 6, column 2). Interpreting the EI results is therefore straightforward for clinicians evaluating diagnostic or screening tests, requiring no new classificatory system.

**Table 6.** Suggested classification of EI values.

EI Value	Qualitative Classification of Change in Probability of Diagnosis (after Jaeschke et al. [14])	Approximate % Change in Probability of Diagnosis (after McGee [21])
≤0.1	Very large decrease	-
0.1	Large decrease	−45
0.2	Large decrease	−30
0.5	Moderate decrease	−15
1.0		0
2.0	Moderate increase	+15
5.0	Moderate increase	+30
10.0	Large increase	+45
≥10.0	Very large increase	-

EI may be compared to other unitary measures which have been used to summarise diagnostic or screening test performance, for example II [22]. Unlike this simple subtraction of Inacc from Acc, EI does not produce negative values which have been previously noted to occur with II (if Inacc > Acc; Table 3) and whose meaning is difficult to comprehend (indeed may be meaningless) [26]. Part of the reason for this may be that, unlike II, EI is developed from “number needed to” metrics (NND\*, NNM), whereas II was used as the basis for a “number needed to” metric, the “number needed to screen” [22]. EI therefore has advantages over II.

Comparing EI to Y, EI is dependent on disease prevalence (Table 4, Figure 3), since Acc and Inacc are calculated using values from both columns of the 2 × 2 contingency table, whereas Y is independent of P since Sens and Spec are strict columnar ratios (although these values may of course vary with the heterogeneity of clinical populations, or spectrum bias [34]). The maximal value of Y arbitrarily assumes disease prevalence to be 50%, which is not often the case in clinical practice. Both EI and Y treat FN and FP as equally undesirable, an assumption which is often not the case in clinical practice where FN may be considered more costly. Y can be negative (boundary values −1, +1), with negative values occurring if the test result is negatively associated with the correct diagnosis (although Y can be normalised, as the balanced accuracy). This is unlike EI (boundary values 0, ∞), which makes risk of misdiagnosis more explicit (values < 1), although this does not directly indicate whether FP or FN is the principal cause of inaccuracy and hence misdiagnosis.

Comparing EI to the diagnostic odds ratio, DOR (=TP × TN/FP × FN), both treat FN and FP as equally undesirable. DOR is independent of P, at least notionally, since it may be expressed solely in terms of Sens and Spec (=Sens × Spec)/[(1 − Sens) × (1 − Spec)]. Both DOR and EI give optimistic results, DOR by choosing the best quality of a test and ignoring its weaknesses, particularly in populations with very high or very low risk. Ratios of DOR become unstable and inflated as the denominator approaches zero, which is also true of EI, although because the classes from the 2 × 2 contingency table are treated additively in EI rather than multiplicatively as in DOR the chance of denominator being zero is less. Hence, EI may have advantages over DOR.

EI may also be compared to other unitary measures, including the critical success index, F measure, area under the receiver operating characteristic curve (AUC ROC), and

Matthews' correlation coefficient (MCC) [4,33,35]. Critical success index and F measure ignore TN values, unlike EI. AUC ROC combines test accuracy over a range of thresholds which may be both clinically relevant and clinically nonsensical, and hence gives a very optimistic measure of test accuracy. MCC takes into account the size of all four classes in the  $2 \times 2$  contingency table and is widely regarded as being a very informative score for establishing the quality of a binary classifier, but the calculation (geometric mean of Y and the predictive summary index) is less straightforward than for EI, and values can be negative (boundary values  $-1, +1$ ). Hence, EI may have advantages over these measures, none of which readily conveys risk of misdiagnosis.

For the communication of risk to patients, use of qualitative information is generally discouraged because of the potential ambiguity of such terms [1,2]. Hence, the suggested qualitative classification of EI values (Table 6, column 1) may not be of use in this situation. However, unlike other quantitative measures, EI involves no fractions, frequencies, or percentages, which may be advantageous when discussing risk with those with low numeracy skills [1,2]. EI is based on "number needed" metrics which were originally deemed more intuitive for patients as well as clinicians [7], but empirical studies have suggested that NNT is more difficult for patients to understand than AR and RR [3,36].

EI is a dimensionless number, like RR and DOR, and a value of  $>1$  suggests diagnostic value, just as a value of  $>1$  suggests association for RR and better than random classification for DOR. As a consequence of the coronavirus pandemic, there may be a general awareness of another metric which has an inflection point at 1, namely  $R_0$ , the reproduction number, used to denote the spread of infectious disease in a population, where infection is spreading if  $R_0 > 1$ , but not so if  $<1$ .

In summary, the proposed EI may prove an acceptable unitary measure of diagnostic and screening test utility for clinicians as it is easy to calculate and interpret. It may also be useful for communicating risk of diagnosis and misdiagnosis, in the specific example of dementia, to patients, but further empirical studies will be required specifically to address this question.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of the Walton Centre for Neurology and Neurosurgery, Approval: N 310, approved 7 September 2020.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** Data available from author on reasonable request.

**Conflicts of Interest:** The author declares no conflict of interest.

## References

1. Naik, G.; Ahmed, H.; Edwards, A.G. Communicating risk to patients and the public. *Br. J. Gen. Pract.* **2012**, *62*, 213–216. [[CrossRef](#)] [[PubMed](#)]
2. Schrager, S.B. Five Ways to Communicate Risks So That Patients Understand. *Fam. Pract. Manag.* **2018**, *25*, 28–31.
3. Gigerenzer, G.; Gaissmaier, W.; Kurz-Milcke, E.; Schwartz, L.M.; Woloshin, S. Helping Doctors and Patients Make Sense of Health Statistics. *Psychol. Sci. Public Interest* **2007**, *8*, 53–96. [[CrossRef](#)] [[PubMed](#)]
4. Lerner, A.J. *The 2x2 Matrix: Contingency, Confusion and the Metrics of Binary Classification*; Springer: London, UK, 2021; in press. [[CrossRef](#)]
5. Lerner, A. Number Needed to Diagnose, Predict, or Misdiagnose: Useful Metrics for Non-Canonical Signs of Cognitive Status? *Dement. Geriatr. Cogn. Disord. Extra* **2018**, *8*, 321–327. [[CrossRef](#)]
6. Lerner, A.J. Evaluating cognitive screening instruments with the "likelihood to be diagnosed or misdiagnosed" measure. *Int. J. Clin. Pract.* **2019**, *73*, e13265. [[CrossRef](#)] [[PubMed](#)]
7. Cook, R.J.; Sackett, D.L. The number needed to treat: A clinically useful measure of treatment effect. *BMJ* **1995**, *310*, 452–454. [[CrossRef](#)] [[PubMed](#)]
8. Zermansky, A. Number needed to harm should be measured for treatments. *BMJ* **1998**, *317*, 1014. [[CrossRef](#)] [[PubMed](#)]
9. Citrome, L.; Ketter, T.A. When does a difference make a difference? Interpretation of number needed to treat, number needed to harm, and likelihood to be helped or harmed. *Int. J. Clin. Pract.* **2013**, *67*, 407–411. [[CrossRef](#)]

10. Habibzadeh, F.; Yadollahie, M. Number Needed to Misdiagnose: A measure of diagnostic test effectiveness. *Epidemiology* **2013**, *24*, 170. [[CrossRef](#)] [[PubMed](#)]
11. Linn, S.; Grunau, P.D. New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests. *Epidemiol. Perspect. Innov.* **2006**, *3*, 11. [[CrossRef](#)] [[PubMed](#)]
12. Larner, A.J. Manual of Screeners for Dementia. In *Pragmatic Test Accuracy Studies*; Springer: London, UK, 2020. [[CrossRef](#)]
13. Williamson, J.C.; Larner, A.J. 'Likelihood to be diagnosed or misdiagnosed': Application to meta-analytic data for cognitive screening instruments. *Neurodegener. Dis. Manag.* **2019**, *9*, 91–95. [[CrossRef](#)] [[PubMed](#)]
14. Jaeschke, R.; Guyatt, G.; Sackett, D.L. 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? *JAMA* **1994**, *271*, 703–707. [[CrossRef](#)] [[PubMed](#)]
15. National Institute for Health and Care Excellence. *Dementia: Assessment, Management and Support for People Living with Dementia and Their Carers (NICE Guideline 97)*; Methods, evidence and recommendations; NICE: London, UK, 2018.
16. Kraemer, H.C. Evaluating medical tests. In *Objective and Quantitative Guidelines*; Sage: Newbery Park, CA, USA, 1992; pp. 27, 34, 115.
17. Folstein, M.F.; Folstein, S.E.; McHugh, P.R. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **1975**, *12*, 189–198. [[CrossRef](#)]
18. Nasreddine, Z.S.; Phillips, N.A.; Bédirian, V.; Charbonneau, S.; Whitehead, V.; Collin, I.; Cummings, J.L.; Chertkow, H. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool for Mild Cognitive Impairment. *J. Am. Geriatr. Soc.* **2005**, *53*, 695–699. [[CrossRef](#)] [[PubMed](#)]
19. Hsieh, S.; McGrory, S.; Leslie, F.; Dawson, K.; Ahmed, S.; Butler, C.; Rowe, J.; Mioshi, E.; Hodges, J.R. The Mini-Addenbrooke's Cognitive Examination: A New Assessment Tool for Dementia. *Dement. Geriatr. Cogn. Disord.* **2015**, *39*, 1–11. [[CrossRef](#)] [[PubMed](#)]
20. Burns, A.; Harrison, J.R.; Symonds, C.; Morris, J. A novel hybrid scale for the assessment of cognitive and executive function: The Free-Cog. *Int. J. Geriatr. Psychiatry* **2021**, *36*, 566–572. [[CrossRef](#)]
21. McGee, S. Simplifying likelihood ratios. *J. Gen. Intern. Med.* **2002**, *17*, 647–650. [[CrossRef](#)]
22. Mitchell, A.J. Index test. In *Encyclopedia of Medical Decision Making*; Kattan, M.W., Ed.; Sage: Los Angeles, CA, USA, 2009; pp. 613–617.
23. Larner, A.J. Mini-Addenbrooke's Cognitive Examination: A pragmatic diagnostic accuracy study. *Int. J. Geriatr. Psychiatry* **2015**, *30*, 547–548. [[CrossRef](#)] [[PubMed](#)]
24. Larner, A.J. Mini-Addenbrooke's Cognitive Examination diagnostic accuracy for dementia: Reproducibility study. *Int. J. Geriatr. Psychiatry* **2015**, *30*, 1103–1104. [[CrossRef](#)] [[PubMed](#)]
25. Larner, A. MACE versus MoCA: Equivalence or superiority? Pragmatic diagnostic test accuracy study. *Int. Psychogeriatrics* **2017**, *29*, 931–937. [[CrossRef](#)] [[PubMed](#)]
26. Larner, A.J. MACE for Diagnosis of Dementia and MCI: Examining Cut-Offs and Predictive Values. *Diagnostics* **2019**, *9*, 51. [[CrossRef](#)] [[PubMed](#)]
27. Larner, A.J. Free-Cog: Pragmatic Test Accuracy Study and Comparison with Mini-Addenbrooke's Cognitive Examination. *Dement. Geriatr. Cogn. Disord.* **2019**, *47*, 254–263. [[CrossRef](#)] [[PubMed](#)]
28. Bossuyt, P.M.; Reitsma, J.B.; Bruns, D.E.; Gatsonis, C.A.; Glasziou, P.; Irwig, L.M.; Moher, D.; Rennie, D.; De Vet, H.C.W.; Lijmer, J.G. The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration. *Clin. Chem.* **2003**, *49*, 7–18. [[CrossRef](#)] [[PubMed](#)]
29. Noel-Storr, A.H.; McCleery, J.M.; Richard, E.; Ritchie, C.W.; Flicker, L.; Cullum, S.J.; Davis, D.; Quinn, T.J.; Hyde, C.; Rutjes, A.W.; et al. Reporting standards for studies of diagnostic test accuracy in dementia: The STARDdem Initiative. *Neurology* **2014**, *83*, 364–373. [[CrossRef](#)]
30. Brown, J.; Pengas, G.; Dawson, K.; Brown, L.A.; Clatworthy, P. Self administered cognitive screening test (TYM) for detection of Alzheimer's disease: Cross sectional study. *BMJ* **2009**, *338*, b2030. [[CrossRef](#)]
31. Hancock, P.; Larner, A.J. Test Your Memory test: Diagnostic utility in a memory clinic population. *Int. J. Geriatr. Psychiatry* **2011**, *26*, 976–980. [[CrossRef](#)] [[PubMed](#)]
32. Larner, A. The 'attended alone' and 'attended with' signs in the assessment of cognitive impairment: A revalidation. *Postgrad. Med.* **2020**, *132*, 595–600. [[CrossRef](#)]
33. Larner, A.J. New unitary metrics for dementia test accuracy studies. *Prog. Neurol. Psychiatry* **2019**, *23*, 21–25. [[CrossRef](#)]
34. Brenner, H.; Gefeller, O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat. Med.* **1997**, *16*, 981–991. [[CrossRef](#)]
35. Larner, A.J. Cognitive screening instruments for dementia: Comparing metrics of test limitation. *Dement. Neuropsychol.* **2021**, *15*, in press.
36. Fagerlin, A.; Peters, E. Quantitative information. In *Communicating Risks and Benefits: An Evidence-Based User's Guide*; Fischhoff, B., Brewer, N.T., Downs, J.S., Eds.; Department of Health and Human Services, Food and Drug Administration: Silver Spring, MD, USA, 2011; pp. 53–64.