Hellenic Complex Systems Laboratory

Uncertainty Estimation of Diagnostic Accuracy Measures under Parametric Distributions

Technical Report XXIX

Rallou A. Chatzimichail, Theodora Chatzimichail, and
Aristides T, Hatjimihail

2025

Uncertainty Estimation of Diagnostic Accuracy Measures under Parametric Distributions

Rallou A. Chatzimichail, M.Eng., M.Sc., Ph.D. [a], *Theodora Chatzimichail, MRCS* [a],
Aristides T. Hatjimihail, MD, PhD [a]

[a] *Hellenic Complex Systems Laboratory*

Abstract

Background: Diagnostic accuracy measures (DAMs) are widely used in evaluating diagnostic tests, yet their uncertainty is often underreported or inconsistently quantified, which can bias threshold-based decisions in clinical practice.

Methods: We developed a computational framework to estimate the measurement, sampling, and combined uncertainty of sixteen DAMs for threshold-based screening or diagnostic tests under three measurand distributional models: normal, lognormal, and gamma. Measurement uncertainty is modelled with linear and nonlinear heteroscedastic functions. Uncertainty is propagated using a first-order Taylor-series expansion. Optimality conditions are derived numerically where applicable. The framework has been implemented in the freely available program *DiagAccU,* in Wolfram Language, allowing parameter specification and estimation and plotting of the DAMs, their uncertainties and confidence intervals (CIs).

Results: Using fasting plasma glucose for diabetes diagnosis as a case study, we demonstrated that across operating thresholds, specificity (Sp), negative predictive value (NPV), and prevalence-adjusted bias-adjusted kappa ($PABAK$) showed narrower CIs. At extreme thresholds, ratio-type measures showed widened CIs. Agreement and concordance based indices exhibited stable behaviour. Confirmatory measures optimised for higher thresholds maximizing specificity, whereas exclusionary measures optimised for lower thresholds maximizing sensitivity.

Conclusions: This framework offers a novel integrated, diagnostic threshold–based approach for estimating uncertainty across a broad spectrum of DAMs, promoting reproducibility and uptake in laboratory medicine and diagnostic research, and directly supporting clinical decision-making for confirmatory diagnosis, diagnosis for exclusion, and triage.

Keywords:

Diagnostic accuracy measures; Measurement uncertainty; Sampling uncertainty; Diagnostic threshold; Normal distribution; Lognormal distribution; Gamma distribution; Heteroscedasticity; GUM; Clinical decision-making

1. Introduction

Medical diagnosis is the process of identifying a disease by analysing its distinctive characteristics through abduction, deduction, and induction [1]. The term diagnosis, from the Greek διάγνωσις (discernment) [2], reflects the central role of distinguishing between healthy and diseased states in individuals. In probabilistic terms, diagnosis can be defined as the stochastic mapping of symptoms, signs, and laboratory or imaging findings onto a specific disease state, informed by established medical knowledge.

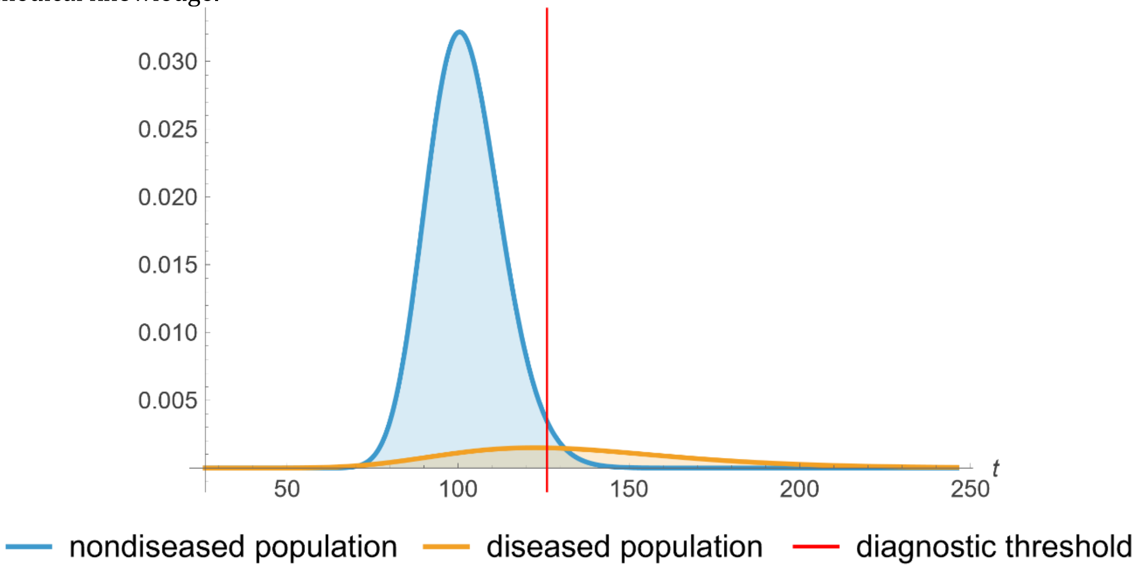


Figure 1. Probability density functions plots of fasting plasma glucose (FPG) in a diabetic (diseased) and nondiabetic (nondiseased) population.

In laboratory medicine, this process is often formalised through in vitro diagnostic tests that act as binary classifiers. For a given measurand, values are compared to a diagnostic threshold t, which dichotomises results into positive (T) and negative outcomes ($\bar{T}$) [3] (refer to Figure 1). While operationally simple, this approach introduces uncertainty because the distributions of measurand values in diseased (D) and nondiseased populations ($\bar{D}$) usually overlap, leading to inevitable misclassifications. Nevertheless, dichotomization has transformed clinical decision-making by mapping continuous biological evidence into actionable categories, such as whether to initiate or withhold treatment [4].

1.1. Diagnostic accuracy measures (DAMs)

The correctness of this threshold-based classification (refer to Table 1) is evaluated using DAMs. Although many DAMs exist [5], a smaller subset is most widely used in clinical research and practice.

Sixteen DAMs were considered for this study, spanning probability-based, predictive, odds and likelihood, concordance, and agreement indices (refer to Section *2.2. DAMs*) . Together, they capture complementary aspects of diagnostic decision-making.

Table 1: 2x2 contingency table

		populations	
		nondiseased ($\bar{D}$)	diseased (D)
test results	negative ($\bar{T}$)	true negative (TN)	false negative (FN)
	positive (T)	false positive (FP)	true positive (TP)

Within this framework, two thresholding strategies are especially relevant to clinical practice. *Confirmatory diagnosis* refers to applying a higher threshold that gives precedence to specificity, ensuring that positive results are rarely false and strengthening the confidence with which disease is confirmed. In contrast, *exclusionary diagnosis* or *diagnosis for exclusion* refers to applying a lower threshold that gives precedence to sensitivity, ensuring that negative results are rarely false and strengthening the confidence with which disease is ruled out. These complementary approaches illustrate how different DAMs support distinct clinical objectives.

Beyond point estimates, however, DAMs are inherently uncertain [6]. From a metrological perspective, uncertainty represents the range of values that could reasonably be attributed to a measure given finite data and imperfect measurement. Two sources are particularly relevant: *sampling uncertainty*, which arises from limited sample sizes, and *measurement uncertainty*, which reflects analytical imprecision of the test [7–9]. Both contribute to the dispersion of DAMs and may alter clinical interpretation[10]. Yet in applied studies, uncertainty is often incompletely reported, limited to standard errors of Se and Sp, or entirely absent for derived measures such as LR^+, DOR, or MCC.

1.2. Measurements distributions

The distributional assumptions underlying measurand values further influence DAM estimation. Although normal models are common, biological variables often exhibit skewness or heavy tails, making *lognormal* or *gamma* distributions more realistic alternatives [11, 12]. Incorporating these distributions allows more faithful representation of measurand behaviour and facilitates uncertainty estimation across a wider range of diagnostic contexts.

The aim of this work is to present a unified, diagnostic threshold–based computational framework that integrates estimation, uncertainty quantification, and optimization across sixteen DAMs under normal, lognormal, and gamma distributional assumptions. By explicitly modelling both sampling variability and measurement imprecision, we provide an approach that supports clinically relevant threshold selection for confirmatory diagnosis, diagnosis for exclusion, and triage.

2. Methods

2.1. Overview

We developed a computational framework extending prior work on uncertainty estimation for DAMs of threshold-based tests [11]. This framework incorporates three parametric models for measurand distributions commonly encountered in laboratory medicine—normal, lognormal, and gamma—and accommodates both homoscedastic and linear and nonlinear heteroscedastic measurement uncertainty functions.

The framework is implemented in the program *DiagAccU*, which estimates sampling, measurement, and combined uncertainties for sixteen DAMs and their confidence intervals (CIs). Uncertainty propagation follows the first-order Taylor-series approach specified in the *Guide to the Expression of Uncertainty in Measurement* (GUM) [13] and its adaptation for laboratory medicine [7]

2.2. DAMs

For this study, the following 16 DAMs were considered: sensitivity (Se), specificity (Sp), overall diagnostic accuracy (ODA), positive predictive value (PPV), negative predictive value (NPV), diagnostic odds ratio (DOR), likelihood ratio for a positive result $(LR^+),$ likelihood ratio for a negative result (LR^-), Youden's index (JS), Euclidean distance (ED), CZ (CZ), Fowlkes–Mallows index (FMI), Cohen's kappa coefficient ($C\kappa$), prevalence-adjusted bias-adjusted kappa ($PABAK$), $F1$ score ($F1S$) , and Matthews correlation coefficient (MCC) [14] (refer to Appendix A.3 and Supplemental file II: *DiagAccUCalculations.nb*.).

2.3. Measurand distributions

For diseased (D) and nondiseased ($\overline{D}$) populations, the measurand distribution are modelled as normal, lognormal, or gamma. Parameters (μ, σ, n) were either estimated from empirical data or specified by the user. This parametric approach ensures coherent propagation of uncertainty across measures and thresholds.

2.4. Measurement and sampling uncertainty

Measurement uncertainty $u_m(t)$ was represented using either a constant contribution or a linear or nonlinear magnitude-dependent function, reflecting homoscedastic or heteroscedastic assay profiles [13]. Sampling uncertainty for means and variances was derived from the central limit theorem and chi-squared distribution [15–17], with prevalence uncertainty approximated by Agresti–Coull intervals [18]. Combined uncertainty was propagated through first-order Taylor expansions [13], with effective degrees of freedom estimated by the Welch–Satterthwaite formula [19, 20]. Expanded uncertainties and CIs were derived accordingly. Full derivations are provided in Appendix A.5–A.6 and in Supplemental file II: *DiagAccUCalculations.nb*.

2.5. Optimization

Optimization was performed numerically for $ODA, JS, ED, CZ, FMI, C\kappa, PABAK, F1S, and\ MCC$.

2.6. Software implementation

The program *DiagAccU* was implemented in Wolfram Language (Wolfram Mathematica® v14.3). It is distributed as a Wolfram Language notebook (.nb), executable in Wolfram Player® or Mathematica®. *DiagAccU* is available in Supplemental File I (*DiagAccU.nb*)

Users can specify sample sizes, population parameters, uncertainty models, and prevalence. The program outputs tables and plots of DAMs with their sampling, measurement, and combined uncertainties, as well as CIs and optimal thresholds.

Figure 2 illustrates a simplified interface flowchart. More detailed interface documentation is available in Supplemental File III (*DiagAccUInterface.pdf*), while software specifications are detailed in Appendix A.6.

3. Results

3.1. Illustrative case study

We applied the framework to fasting plasma glucose (FPG) for diagnosing diabetes mellitus, using oral glucose tolerance testing (OGTT) as the reference standard [21], in adults aged 65–68 years from the National Health and Nutrition Examination Survey (NHANES) 2005–2016 (n = 414) [22]. Diabetes prevalence in this subgroup was 12.6% (52/414) [23]. FPG values for diabetic and nondiabetic participants were well modelled by lognormal distributions, using the maximum likelihood estimation method [27] (Table 2, Figure 3).

Measurement uncertainty was estimated by nonlinear least squares regression [28, 29] from 1,350 QC samples and modelled with a nonlinear heteroscedastic function (constant term b_0 and proportionality term b_1) (refer to Appendix A.5.1).

Results of the application of the program on the illustrative case study dataset are presented in Figures 4-12. Unless otherwise noted, all figures use the settings in Table 3. The selected diagnostic threshold $t = 126$ mg/dl of Figures 6 and 9 is the American Diabetes Association (ADA) diagnostic threshold of FPG for diabetes (refer to Figure 1).

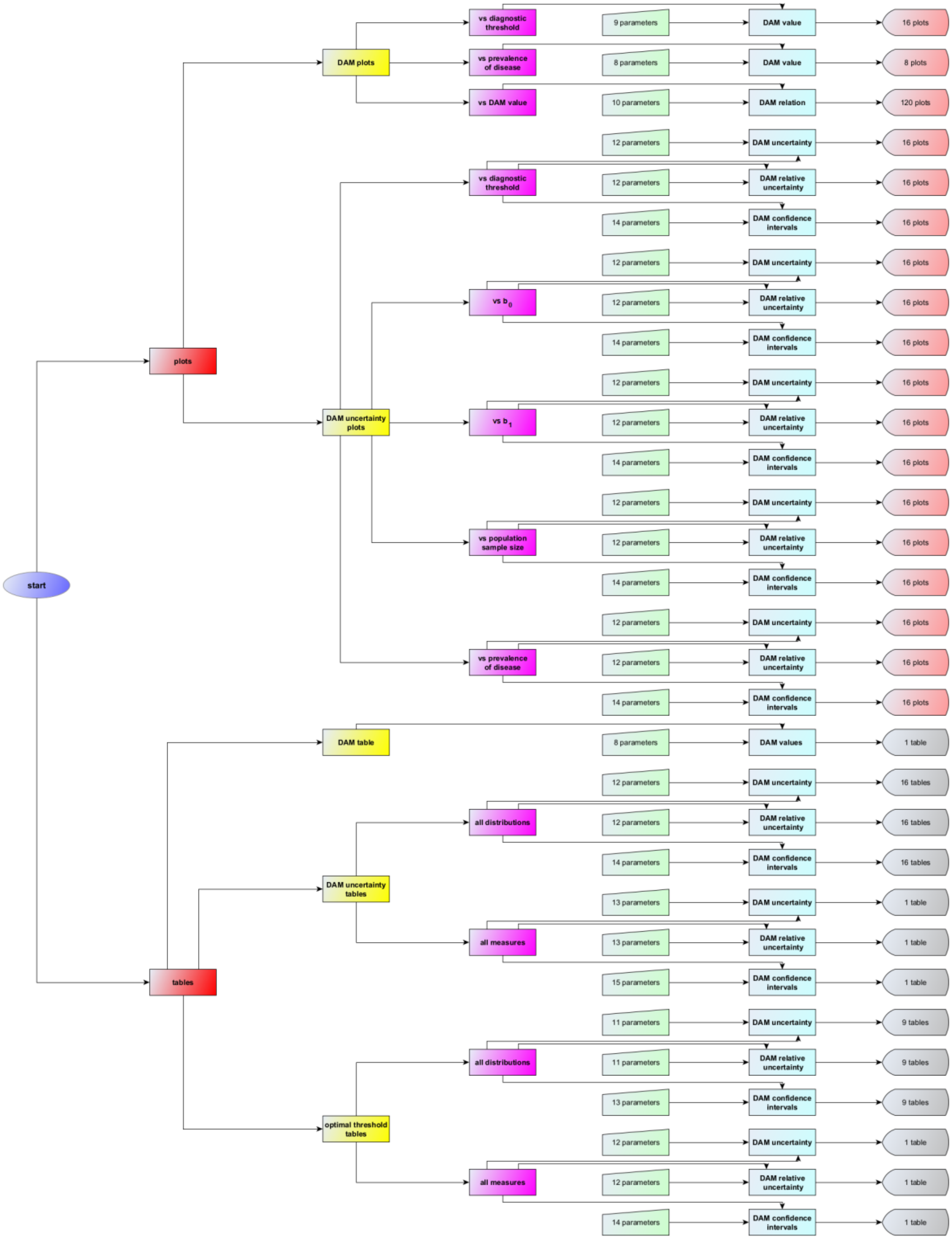


Figure 2. A simplified interface flowchart of the program *DiagAccU*.

Table 2. Descriptive statistics of the datasets and the estimated lognormal distributions of the diabetic and nondiabetic populations.

	Diabetic Participants			**Nondiabetic Participants**		
	Dataset	L_D	l_D	Dataset	$L_{\bar{D}}$	$l_{\bar{D}}$
n	52	-	-	362	-	-
Mean (mg/dL)	136.6	136.0	136.0	102.6	102.2	102.2
Median (mg/dL)	123.5	131.3	131.3	102.0	101.6	101.6
Standard Deviation (mg/dL)	44.7	36.7	36.6	10.9	11.1	11.0
Mean Uncertainty (mg/dL)	1.863	1.863	0	1.469	1.469	0
Skewness	2.168	0.829	0.827	0.521	0.328	0.325
Kurtosis	7.762	4.245	4.242	3.435	3.192	3.189
p-value (Cramér–von Mises test)	-	0.156	0.156	-	0.542	0.509

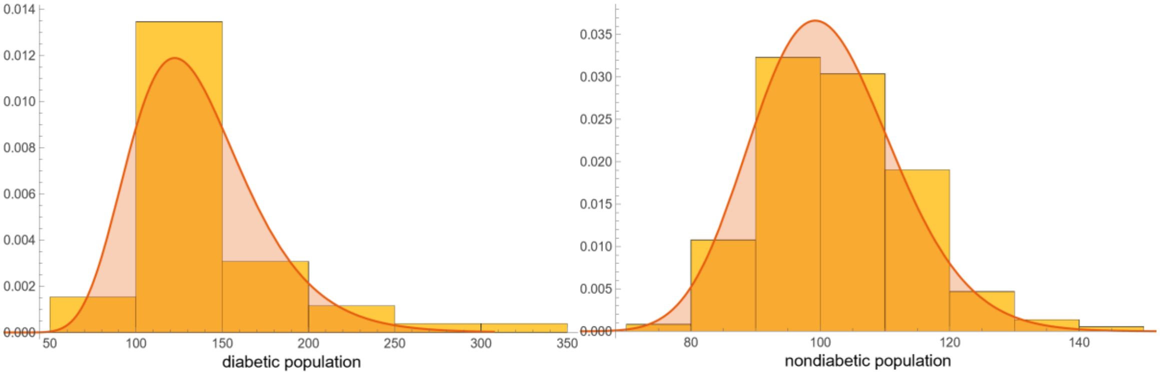


Figure 3. The estimated PDF of the FPG (mg/dL) in diabetic and nondiabetic participants.

1.1. Threshold-dependent profiles

Across operating diagnostic thresholds (refer to Figure 4), Se decreased monotonically with diagnostic threshold t, while Sp increased to a high plateau. PPV increased with t, whereas NPV was high at low t and declined thereafter. ODA rose rapidly from low t, plateaued and showed a slight late downturn. Ratio-type indices exhibited tail amplification: DOR and LR^{+} increased sharply at high t, whereas LR^{-} attained a shallow interior minimum and then increased. Association and agreement indices show interior optima: $JS, CZ, C\kappa, FMI, F1S$, and MCC displayed interior maxima, while ED showed a distinct interior minimum. $PABAK$ transitioned from negative values at the lowest t to a broad positive plateau at higher t.

1.2. Prevalence-dependent profiles

At a fixed diagnostic threshold, predictive values were strongly prevalence-dependent (refer to Figure 5): PPV increased steeply at low v and approached an upper plateau, whereas NPV decreased monotonically toward zero. ODA declined approximately linearly with increasing v, reflecting increasing class imbalance. FMI and $F1S$ rose steeply at low v and then levelled. In contrast, $C\kappa$ and the MCC were unimodal, each attaining an interior maximum at moderate v and decreasing at high v. $PABAK$ decreased roughly linearly across the range of v.

These patterns illustrate the strong prevalence dependence of predictive values and F-type measures, the near-linear attenuation of ODA with increasing class imbalance, and the interior-optimum behaviour of association metrics such as $C\kappa$ and MCC.

Table 3. The settings of the program *DiagAccU* for the figures 4-11

	Units	**Fig 4**	**Fig 5**	**Fig 6**	**Fig 7-8**	**Fig 9**	**Fig 10**	**Fig 11**	**Fig 12**
t	mg/dL	91.0–173.0	126	-	91.0–173.0	91.0–173.0	126	126	-
μ_D	mg/dL	136.0	136.0	136.0	136.0	136.0	136.0	136.0	136.0
σ_D	mg/dL	36.6	17.7	17.7	17.7	17.7	17.7	17.7	17.7
$\mu_{\overline{D}}$	mg/dL	102.2	102.2	102.2	102.2	102.2	102.2	102.2	102.2
$\sigma_{\overline{D}}$	mg/dL	11.0	11.0	11.0	11.0	11.0	11.0	11.0	11.0
v		0.126	0.001-0.999	0.126	0.126	0.126	0.126	0.126	0.126
n_U						1350		1350	1350
b_0		-	-	-	0.8124	0.8124	-	0.8124	0.8124
b_1		-	-	-	0.0119	0.0119	-	0.0119	0.0119
p		-	-	-	-	0.95	-	0.95	0.95
d_D	lognormal						normal lognormal gamma	lognormal	lognormal
$d_{\overline{D}}$	lognormal						normal lognormal gamma	lognormal	lognormal

1.1. DAM relations

Pairwise relations between measures (refer to Figure 6) were frequently non-bijective across t. Nearly monotone relations were observed for $F1S$ versus FMI (increasing) and ED versus CZ (decreasing). Narrow loops—indicating multi-valued mappings as t crossed low-to-high regions—were evident for and $C\kappa$ versus JS, ODA versus JS, $C\kappa$ versus MCC, and $F1S$ versus CZ. ED versus ODA showed a U-shaped relation with a clear interior minimum in ED, and ODA versus JS showed a late downturn of ODA at high JS. DOR versus MCC was highly non-linear, with a precipitous rise in DOR in the high-specificity tail for modest changes in MCC, underscoring the instability of ratio-type measures near boundary regions.

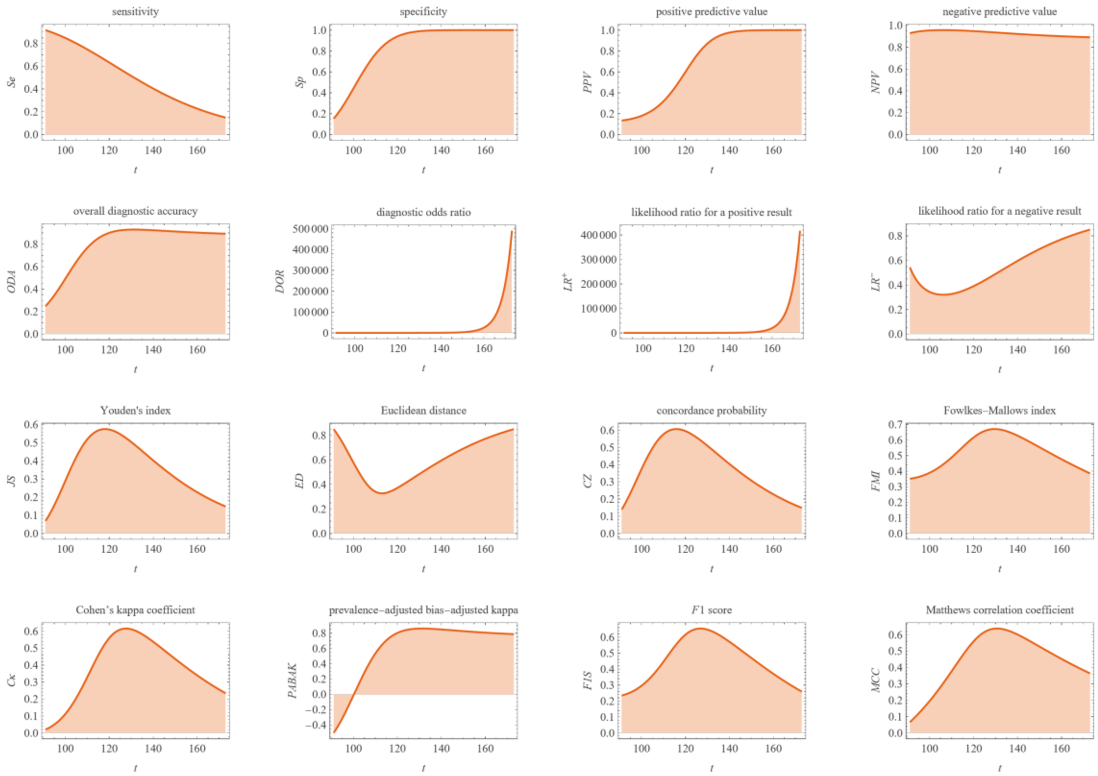


Figure 4. The 16 DAMs versus threshold t.

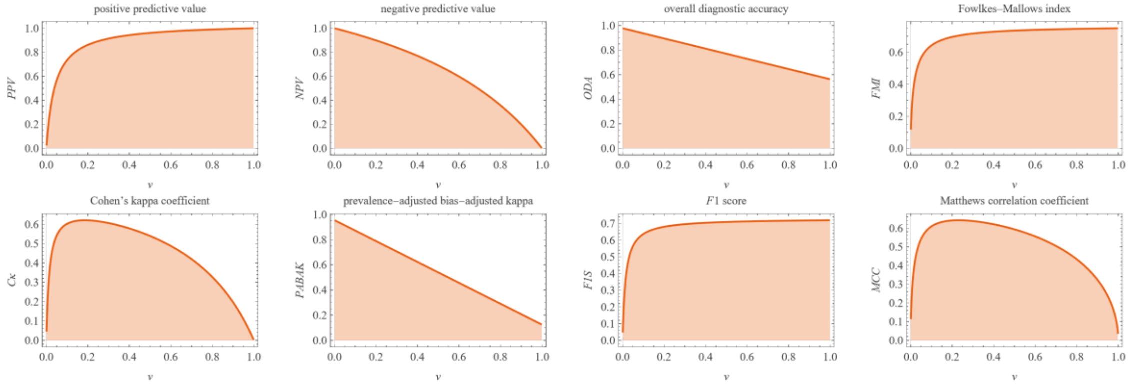


Figure 5. Positive predictive value (PPV), negative predictive value (NPV), overall diagnostic accuracy (ODA), Fowlkes–Mallows index (FMI), Cohen's kappa coefficient ($C\kappa$), prevalence-adjusted bias-adjusted kappa ($PABAK$), $F1$ score ($F1S$), and Matthews correlation coefficient (MCC) versus prevalence of diabetes v.

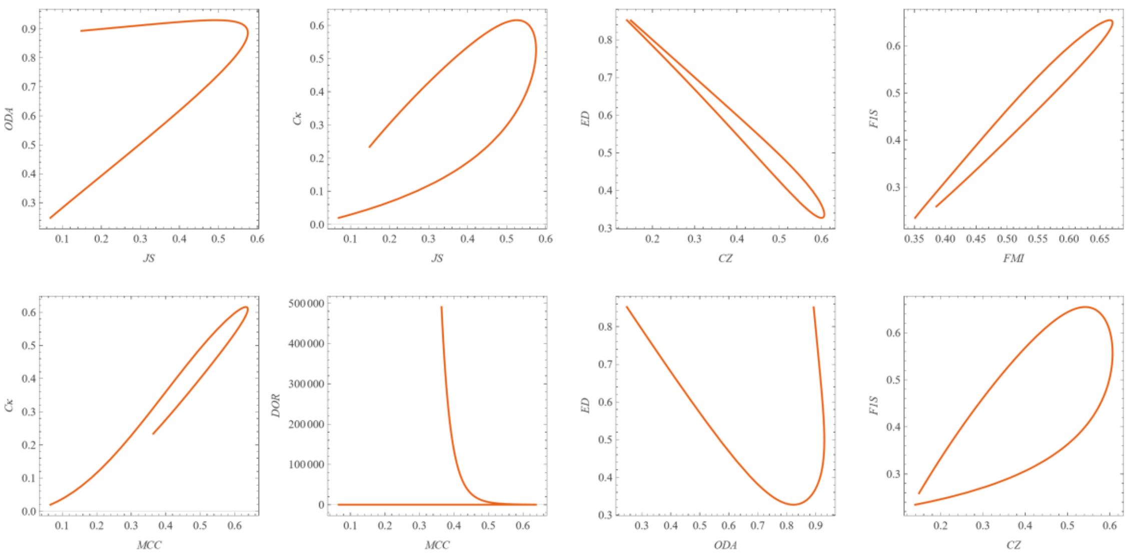


Figure 6. Plots of overall diagnostic accuracy (ODA) and Cohen's kappa coefficient ($C\kappa$) versus Youden's index (JS), Euclidean distance (ED) versus concordance probability (CZ), $F1$ score ($F1S$) versus Fowlkes–Mallows index (FMI), Cohen's kappa coefficient ($C\kappa$) and diagnostic odds ratio (DOR) versus Matthews correlation coefficient (MCC), Euclidean distance (ED) versus overall diagnostic accuracy (ODA), and $F1$ score ($F1S$) versus concordance probability (CZ).

1.2. Uncertainty

The standard combined uncertainty $u_c(t)$ exhibited three patterns:

a) Interior humps: $u_c(t)$ increased from low t, peaked mid-range, and declined for several measures ($Se, ODA, JS, ED, CZ, FMI, C\kappa, PABAK, F1S, MCC$).

b) Monotone decrease: $u_c(t)$ declined with t for Sp, NPV, and the LR^{-}.

c) Right-tail escalation: $u_c(t)$ rose steeply at high t for LR^{+} and DOR.

Component-wise, sampling uncertainty $u_s(t)$ dominated at low t where diseased positives were scarce (notably NPV, and the LR^{-}), whereas measurement uncertainty $u_m(t)$ predominated in the upper tail (particularly LR^{+} and DOR). For PPV, $u_m(t)$ displayed a distinct interior maximum.

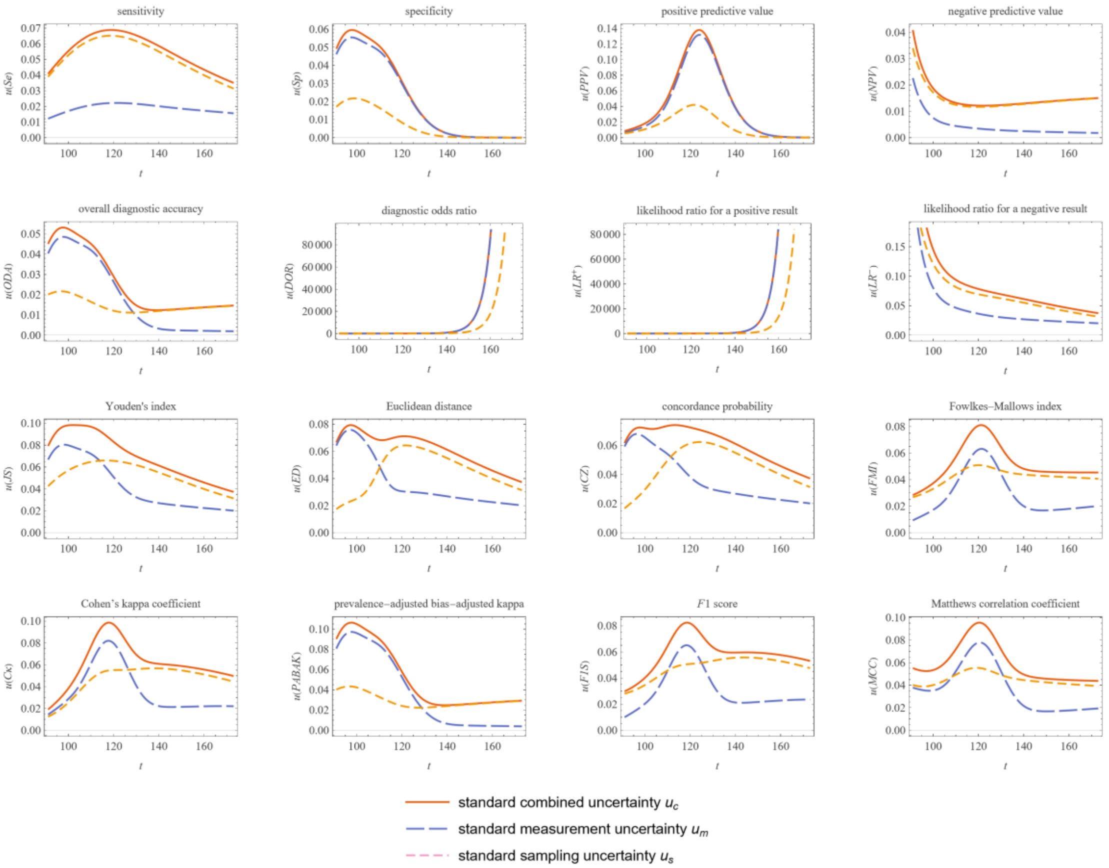


Figure 7. Standard sampling, measurement, and combined uncertainty of the sixteen DAMs versus threshold t.

1.3. Relative uncertainty

The relative standard combined uncertainty (refer to Figure 8) generally decreased from low to interior t and increased again in the right tail. For sensitivity, specificity, NPV, ODA, JS, CZ, $C\kappa$, $PABAK$, and MCC, the combined curve closely tracked $u_m(t)$ indicating predominance of measurement error across most thresholds (with monotone declines for Sp, NPV, and ODA related profiles and an increase for Se). Ratio-type measures showed pronounced tail effects: for LR^-, $u_s(t)$ dominated at low t and then fell rapidly; for LR^+ and DOR, relative uncertainties escalated sharply at high t and were largely measurement-driven. Euclidean distance showed an interior maximum, and FMI and $F1S$ exhibited interior humps, consistent with larger relative uncertainty near their extrema. Overall, relative uncertainty was minimised at interior operating points for most non-ratio measures, whereas extremes accentuated uncertainty via sampling (low t) or measurement imprecision (high t).

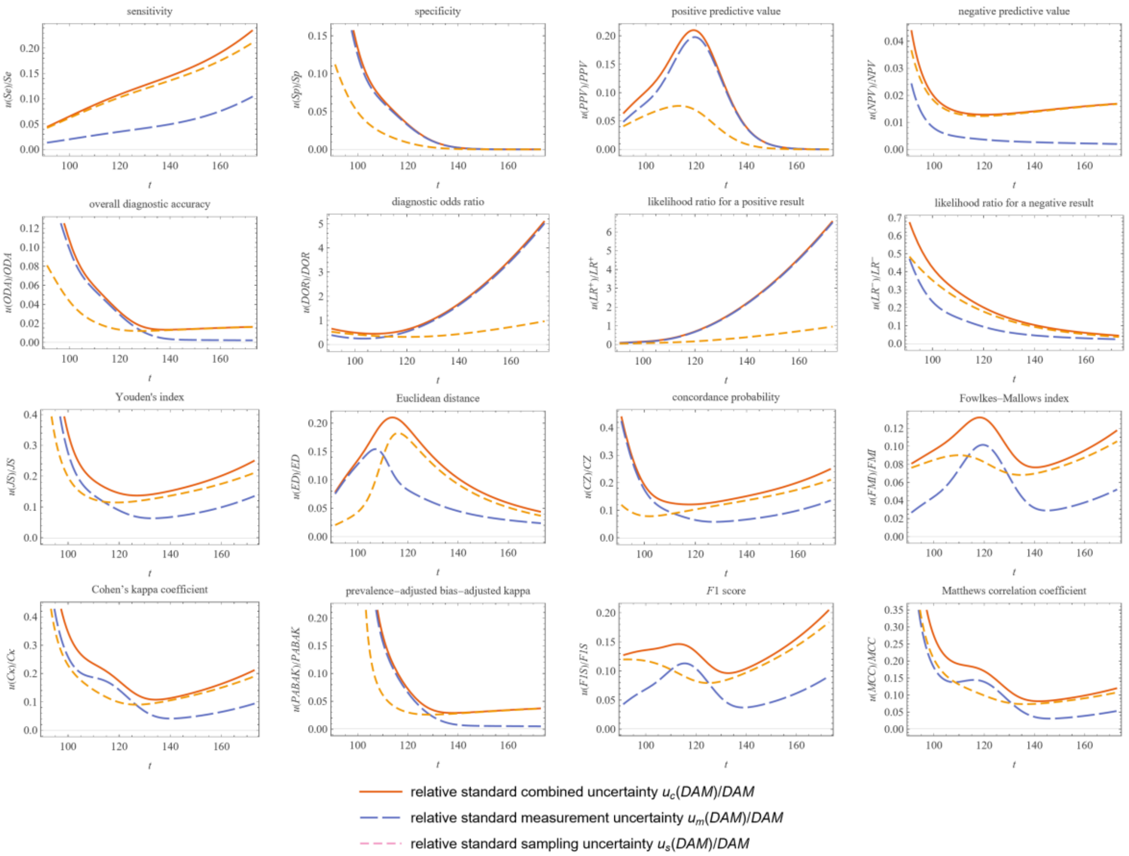


Figure 8. Relative standard sampling, measurement, and combined uncertainty of the sixteen DAMs versus threshold t.

1.4. CIs

Figure 9 presents 95% CIs across operating diagnostic thresholds. CIs were narrower for Sp, NPV, ODA and $PABAK$. Intervals widened for ratio-type measures; at lower t for LR^{-} and at higher t for LR^{+} and DOR, consistent with denominator instability as specificity or sensitivity approaches 1 or 0. By contrast, the CIs of prevalence-invariant association and agreement indices are wider near their maxima (JS, CZ, FMI, $C\kappa$, $PABAK$, $F1S$ and MCC) or near its minimum (ED).

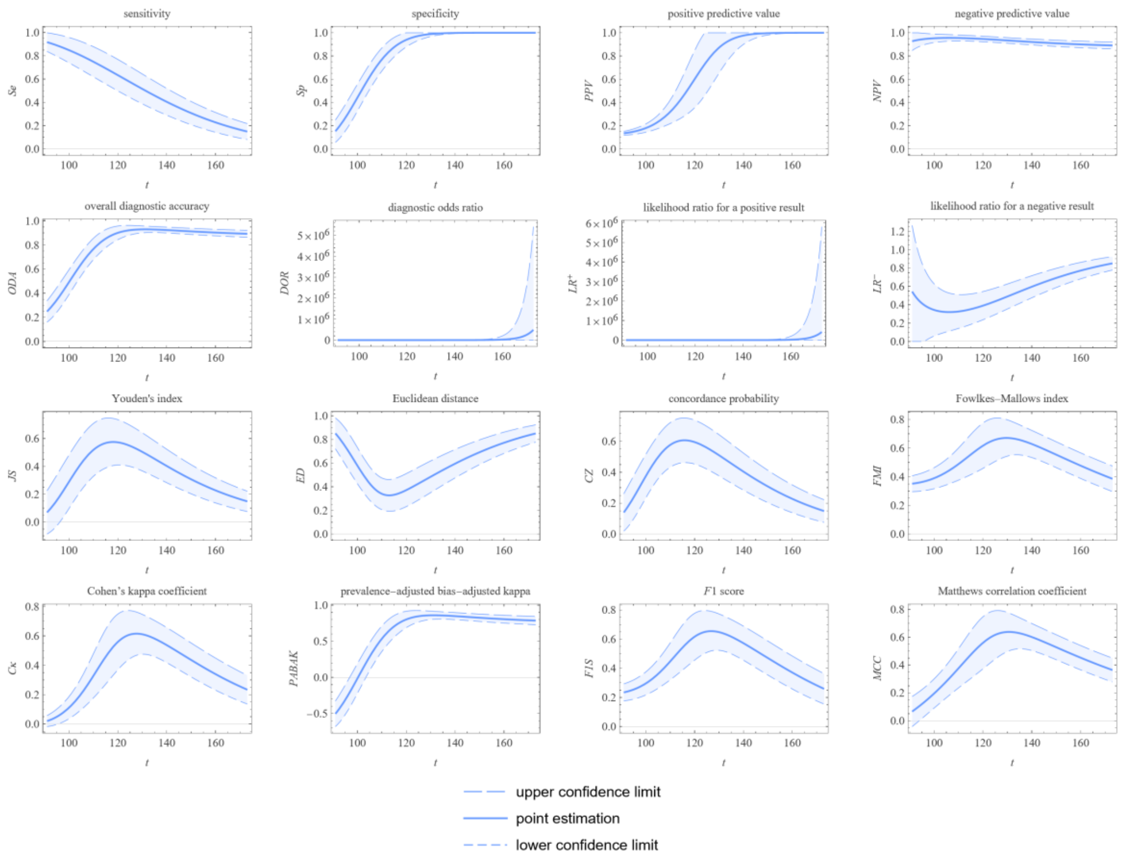
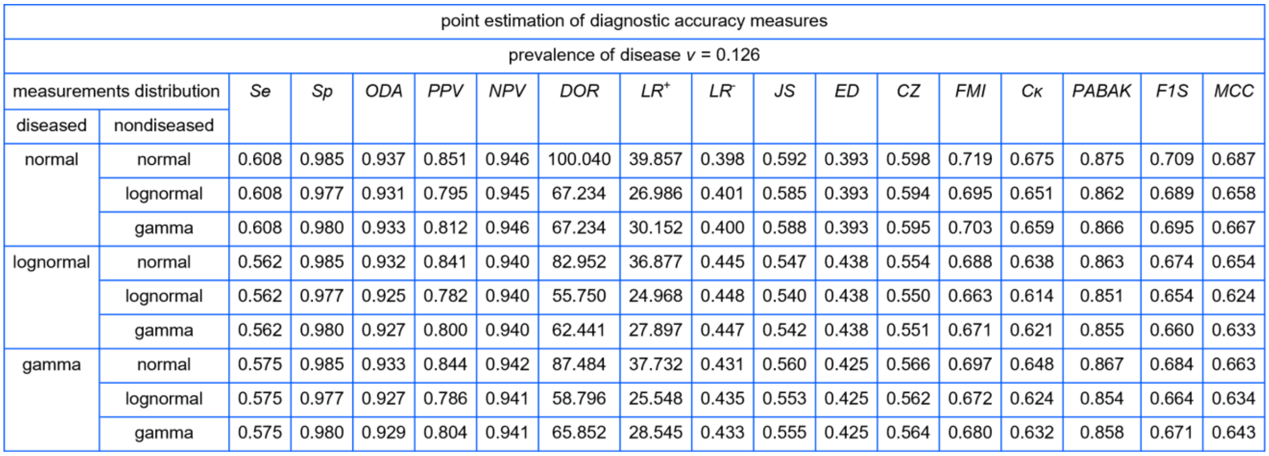


Figure 9. CIs of the 16 DAMs versus threshold t.

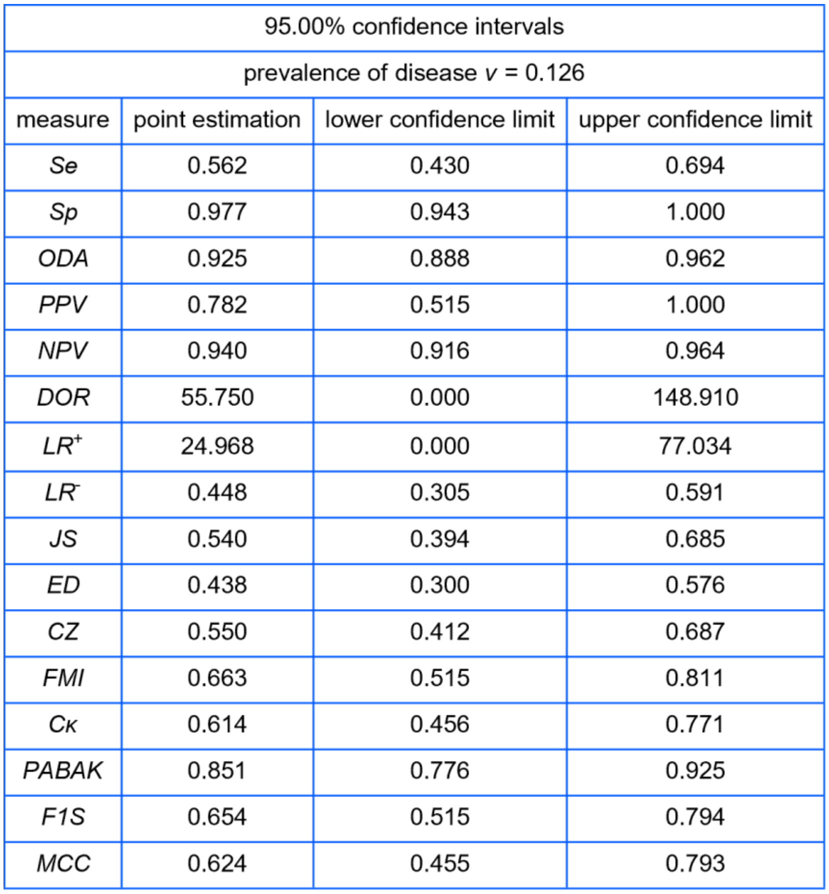
1.5. Point estimates and CIs at a clinical threshold

At the ADA screening cut-off of 126 mg/dL, point estimates and 95% CIs for all 16 DAMs are shown in Figures 10-11. PPV was modest, reflecting the low prevalence of diabetes in this cohort, whereas NPV remained high. Agreement and concordance indices fell in the mid-range, indicating partial but clinically useful discrimination.



point estimation of diagnostic accuracy measures																	
prevalence of disease v = 0.126																	
measurements distribution		Se	Sp	ODA	PPV	NPV	DOR	LR+	LR-	JS	ED	CZ	FMI	Cκ	PABAK	F1S	MCC
diseased	nondiseased																
normal	normal	0.608	0.985	0.937	0.851	0.946	100.040	39.857	0.398	0.592	0.393	0.598	0.719	0.675	0.875	0.709	0.687
	lognormal	0.608	0.977	0.931	0.795	0.945	67.234	26.986	0.401	0.585	0.393	0.594	0.695	0.651	0.862	0.689	0.658
	gamma	0.608	0.980	0.933	0.812	0.946	67.234	30.152	0.400	0.588	0.393	0.595	0.703	0.659	0.866	0.695	0.667
lognormal	normal	0.562	0.985	0.932	0.841	0.940	82.952	36.877	0.445	0.547	0.438	0.554	0.688	0.638	0.863	0.674	0.654
	lognormal	0.562	0.977	0.925	0.782	0.940	55.750	24.968	0.448	0.540	0.438	0.550	0.663	0.614	0.851	0.654	0.624
	gamma	0.562	0.980	0.927	0.800	0.940	62.441	27.897	0.447	0.542	0.438	0.551	0.671	0.621	0.855	0.660	0.633
gamma	normal	0.575	0.985	0.933	0.844	0.942	87.484	37.732	0.431	0.560	0.425	0.566	0.697	0.648	0.867	0.684	0.663
	lognormal	0.575	0.977	0.927	0.786	0.941	58.796	25.548	0.435	0.553	0.425	0.562	0.672	0.624	0.854	0.664	0.634
	gamma	0.575	0.980	0.929	0.804	0.941	65.852	28.545	0.433	0.555	0.425	0.564	0.680	0.632	0.858	0.671	0.643

Figure 10. Table of the point estimations of the DAMs for an FPG value $t = 126$ mg/dL at prevalence $v \cong 0.126$, with the other settings of the program in Table 5.



95.00% confidence intervals			
prevalence of disease *v* = 0.126			
measure	point estimation	lower confidence limit	upper confidence limit
Se	0.562	0.430	0.694
Sp	0.977	0.943	1.000
ODA	0.925	0.888	0.962
PPV	0.782	0.515	1.000
NPV	0.940	0.916	0.964
DOR	55.750	0.000	148.910
LR^+	24.968	0.000	77.034
LR^-	0.448	0.305	0.591
JS	0.540	0.394	0.685
ED	0.438	0.300	0.576
CZ	0.550	0.412	0.687
FMI	0.663	0.515	0.811
Cκ	0.614	0.456	0.771
PABAK	0.851	0.776	0.925
F1S	0.654	0.515	0.794
MCC	0.624	0.455	0.793

Figure 11. Table of the point estimations and the 95% CIs of the DAMs for an FPG value $t = 126$ mg/dL at prevalence $v \cong 0.126$, with the other settings of the program in Table 5. Both distributions of the diseased and nondiseased are assumed lognormal.

1.6. Optimal diagnostic thresholds

Optimal thresholds for each DAM are summarised in Figure 12. Se and Sp based optima clustered near the crossing of the class-conditional densities, while JS and ED identified nearby but distinct interior cut-offs. Confirmatory measures (LR^+, DOR) were characterised by higher thresholds giving precedence to specificity; exclusionary measures (LR^-) were characterised by lower thresholds giving precedence to sensitivity. Agreement and concordance indices selected balanced interior thresholds. These results confirm that no single threshold is universally optimal; choice should reflect the diagnostic objective.

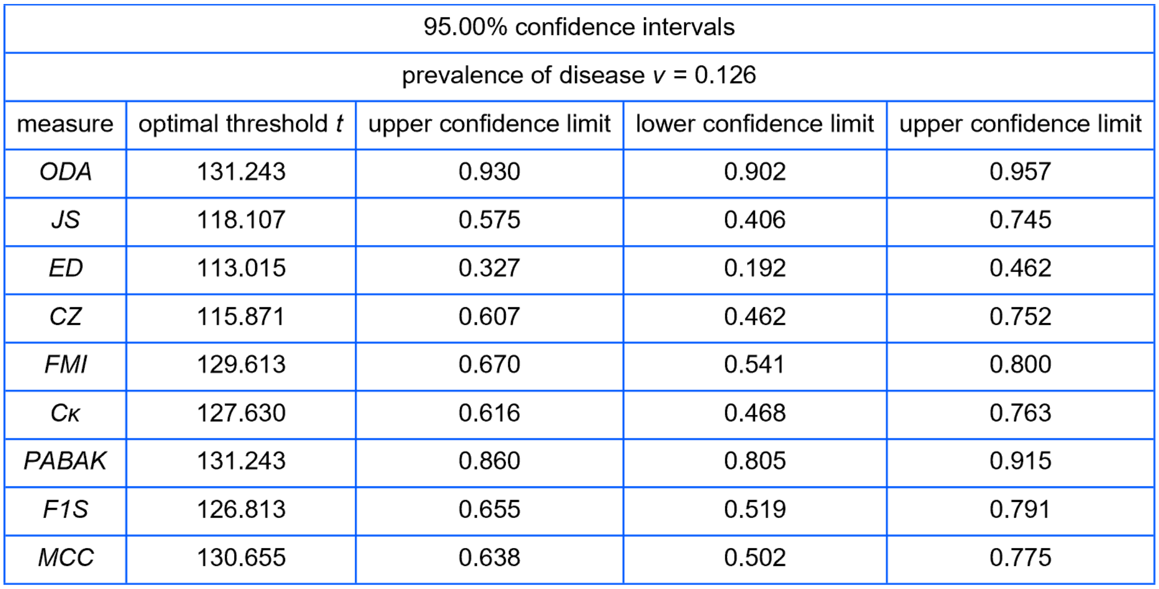
2. Discussion

2.1. Principal findings

We synthesised sixteen DAMs within a unified, threshold-dependent framework and examined their behaviour under parametric modelling of measurand distributions in diseased and non-diseased populations. Consistent with diagnostic principles, confirmatory measures such as $LR^+ and\ DOR$ achieved their optima at higher thresholds, giving precedence to specificity, whereas measures used for diagnosis for exclusion (LR^-) were characterised by lower thresholds, giving precedence to sensitivity. Prevalence-dependent measures (PPV, NPV, ODA, $F1S$, FMI) followed the anticipated monotonic responses to disease prevalence v. Importantly, optimal diagnostic thresholds t^* differed systematically across measures, underscoring that no single cut-off can simultaneously optimise all clinically relevant objectives.

2.2. Relations between measures

Our relational analyses highlighted complementary insights. Prevalence-invariant indices captured the intrinsic separation between diseased and non-diseased populations, while prevalence-dependent indices quantified post-test certainty. Reporting both families prevents misinterpretation of diagnostic accuracy as predictive value. Ratio measures (LR^+, LR^-, DOR) and κ-type indices were unstable near boundaries, supporting the use of bounded measures such as MCC as anchors. The contrast between ODA and JS illustrated that threshold optimization differs when guided by Bayes risk [24], which incorporates prevalence and costs, versus when assessed ignoring prevalence. Pairings such as ED-CZ and $F1S$-FMI provided complementary views of concordance, particularly under asymmetric Se and Sp gains.



95.00% confidence intervals				
prevalence of disease *v* = 0.126				
measure	optimal threshold *t*	upper confidence limit	lower confidence limit	upper confidence limit
ODA	131.243	0.930	0.902	0.957
JS	118.107	0.575	0.406	0.745
ED	113.015	0.327	0.192	0.462
CZ	115.871	0.607	0.462	0.752
FMI	129.613	0.670	0.541	0.800
Cκ	127.630	0.616	0.468	0.763
PABAK	131.243	0.860	0.805	0.915
F1S	126.813	0.655	0.519	0.791
MCC	130.655	0.638	0.502	0.775

Figure 12. Table of the optimal diagnostic thresholds (in mg/dl), the respective point estimations and the 95% CIs of the DAMs at prevalence $v \cong 0.126$, with the other settings of the program in Table 5.

2.3. Appraisal of the uncertainty estimation approach

A key innovation of this work is the explicit partitioning of combined uncertainty into sampling and measurement components; the latter modelled via a heteroscedastic function $u_m(t)$. By partitioning overall uncertainty into sampling and measurement components, the framework clarifies when analytical imprecision dominates (e.g., at high thresholds affecting LR^+ and DOR) and when small sized samples (e.g., at low thresholds with few diseased positives).

Quantifying diagnostic uncertainty is imperative in laboratory medicine to define analytical performance specifications [25], manage quality and risk [26, 27], and design and implement test accuracy studies. Enhancing assay precision and standardization can, in turn, yield more reliable diagnosis and support more effective patient care.

2.4. Potential sources of error and bias

Several methodological caveats warrant consideration:

a) *Reference standard bias:* Reliance on OGTT as a gold standard may introduce misclassification [28–36].

b) *Distributional misspecification:* Normal or lognormal assumptions are parsimonious but can be violated by latent mixtures [37], skewness differences, or heavy tails. Both unimodal and multimodal forms remain plausible [38–40]; model checking and sensitivity analyses are advisable.

c) *Prevalence handling:* Ignoring variance in prevalence underestimates uncertainty. Spectrum effects between source and target populations further complicate transportability.

d) *Boundary behaviour:* Ratio measures exhibit unstable CIs at extreme thresholds, limiting interpretability; log-scale or Fieller-type intervals provide more reliable inference [41, 42].

e) *Selection uncertainty:* Reporting CIs at data-selected optimal thresholds understates uncertainty [43, 44]; bootstrap or cross-validated threshold re-estimation is recommended.

f) *Measurement uncertainty transferability:* The fitted $u_m(t)$ may vary with matrix, lot, or subgroup; periodic validation is necessary.

g) *Dependence among estimators:* Treating sample means and variances as independent can misstate uncertainty; likelihood-based or bootstrap methods better respect dependence [45–47].

h) *Truncation and orientation:* Forcing DAMs into parametric bounds or mis-specifying decision orientation can distort results; analyses should be carefully validated.

2.5. Strengths of the framework

This work provides several contributions:

a) *Comparability across DAMs*: Expressing all sixteen measures as deterministic functions of $\{t, n_D, m_D, s_D, n_{\bar{D}}, m_{\bar{D}}, s_{\bar{D}}\}$ permits their principled comparison and facilitates the derivation of thresholds

b) *Explicit uncertainty decomposition:* By modelling and propagating measurement uncertainty alongside sampling variability, the framework clarifies when analytical imprecision dominates.

c) *Analytic structure:* While threshold optimization is performed numerically, closed-form expressions enable analytic propagation of uncertainty, improving computational efficiency and interpretability.

d) *Reproducibility:* Implementation in the Wolfram Language ensures transparent, auditable analyses under alternative assumptions or updated data.

2.6. Originality and positioning

To our knowledge, few applied studies on DAMs present an integrated framework that simultaneously:

a) Models diseased and non-diseased distributions parametrically,

b) Propagates heteroscedastic measurement uncertainty,

c) Partitions combined uncertainty into sampling and measurement components, and

d) Optimises thresholds across a wide spectrum of DAMs.

While individual elements have precedents, their combined implementation and breadth appear uncommon. The accompanying software (*DiagAccU*) provides a wide range of plot types and comprehensive tables, extending beyond the capabilities of commonly used statistical packages. To the best of our knowledge neither of them offers this extensive range of plots and tables without requiring advanced statistical programming.

2.7. Practical guidance

The framework supports several recommendations for applied research and clinical reporting:

a) Use thresholds aligned with diagnostic purpose-higher diagnostic thresholds for confirmatory diagnosis, lower for diagnosis for exclusion, balanced indices for screening.

b) Report uncertainty at both fixed and optimal thresholds, preferably with bootstrap or cross-validated adjustment for selection.

c) Stabilise ratio measures by working on the log-scale and presenting one-sided bounds when clinically relevant.

d) Periodically re-validate measurement-uncertainty models against quality-control data to ensure applicability across time, matrices, and lots.

2.8. Limitations and future directions

Limitations include reliance on parametric distributional forms, use of first-order Taylor approximations for uncertainty propagation, neglect of uncertainty in measurement-model parameters, and conditional CIs at estimated optimal thresholds. Future work should explore semiparametric and mixture models, full parametric bootstrapping, cross-validated threshold selection, and generalization to multi-class markers and decision-curve analysis.

3. Conclusion

This study introduced a unified, diagnostic threshold–based computational framework for sixteen DAMs. The framework integrates point estimation, uncertainty quantification, and optimization with respect to clinically meaningful objectives. By formulating each measure as a deterministic function of $n_D, m_D, s_D, n_{\bar{D}}, m_{\bar{D}}, and\ s_{\bar{D}}$ and by modelling class-conditional measurand distributions parametrically, the method generates smooth, interpretable profiles across diagnostic thresholds. Measurement uncertainty is incorporated through an explicit heteroscedastic model and propagated jointly with sampling variability, yielding CIs and threshold recommendations that transparently reflect both analytical imprecision and finite-sample uncertainty.

The empirical analyses showed that optimal thresholds vary systematically across DAM families: confirmatory diagnosis indices (e.g., LR^+, DOR) were characterised by higher thresholds, giving precedence to specificity; diagnosis for exclusion indices (e.g., LR^-) were characterised by lower thresholds, giving precedence to sensitivity; and geometric or association-based measures (e.g., $JS, ED, CZ, FMI, F1S, C\kappa, PABAK, MCC$) selected interior thresholds balancing errors. Prevalence-dependent measures ($PPV, NPV, ODA, F1S, FMI$) varied as expected with disease prevalence, whereas prevalence-invariant measures remained stable. The decomposition of uncertainty highlighted contexts where measurement uncertainty predominates (e.g., high-threshold regions where denominators approach zero) versus where sampling variability dominates (e.g., low-threshold regions with limited diseased cases).

The framework's methodological strengths include its unified taxonomy across heterogeneous measures, explicit treatment of measurement uncertainty alongside sampling variability, and transparent implementation in the Wolfram Language. Together, these features enable principled selection of thresholds for confirmatory diagnosis versus diagnosis for exclusion, with explicit trade-offs and uncertainty statements aligned to clinical decision-making.

Nevertheless, inferences remain contingent on appropriate distributional assumptions, correct orientation of decision rules, and robust handling of prevalence. Ratio-based measures require caution near boundary conditions, and threshold selection uncertainty should be explicitly addressed, for example by bootstrap or cross-validation procedures that re-estimate the threshold in each replicate. Periodic validation of the measurement uncertainty model against quality-control data is recommended to ensure transferability across settings and time.

In conclusion, this framework offers a coherent, practical pathway from parametrically fitted biomarker distributions in diseased and non-diseased groups to a comparison of diagnostic objectives with explicitly quantified uncertainty. By making clear the respective roles of disease prevalence, analytical imprecision, and threshold selection, it enables transparent reporting and supports more reliable clinical decisions in both laboratory and bedside settings. Future work should evaluate semiparametric and mixture formulations, extend to multi-marker panels, and incorporate decision-curve or net-benefit analyses to link threshold choices to expected clinical utility.

4. Supplemental material

The following supplemental files are available for download as a ZIP archive at: https://www.hcsl.com/Supplements/SDAMU.zip (accessed on September 26, 2025):

a) Supplemental File I:

DiagAccU.nb: The program as a Wolfram Mathematica Notebook.

b) Supplemental File II:

DiagAccUCalculations.nb: The calculations for the estimation of the DAMs and their standard uncertainty in a Wolfram Mathematica Notebook.

c) Supplemental File III:

DiagAccUInterface.pdf: A brief documentation of the interface of the program.

5. Declarations

Author Contributions: Conceptualization: RCA and TC; methodology: RCA and ATH; software: RCA, TC and ATH; validation: TC; formal analysis: RCA and ATH; investigation: TC; resources: ATH; data curation: TC; writing—original draft preparation: RCA; writing—review and editing ATH; visualization: RCA and TC; supervision: ATH; project administration: RCA. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Data collection was conducted following the rules of the Declaration of Helsinki. The National Center for Health Statistics Ethics Review Board approved data collection and posting of the data online for public use (Protocols #2005-06 and #2011-17). Refer to: https://www.cdc.gov/nchs/nhanes/about/erb.html (accessed on September 26, 2025).

Informed Consent Statement: Written consent was obtained from each subject participating in the survey.

Consent for publication: Not Applicable.

Data Availability Statement: The data presented in this study are available at https://wwwn.cdc.gov/nchs/nhanes/default.aspx (accessed on September 26, 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

6. References

1. Stanley DE, Campos DG. The logic of medical diagnosis. Perspect Biol Med. 2013;56:300–15.

2. Weiner ESC, Simpson JA, Oxford University Press. The Oxford English dictionary. Oxford, Oxford: Clarendon Press ; Melbourne; 1989 2004.

3. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation. 2007;115:654–7.

4. Djulbegovic B, van den Ende J, Hamm RM, Mayrhofer T, Hozo I, Pauker SG, et al. When is rational to order a diagnostic test, or prescribe treatment: the threshold model as an explanation of practice variation. Eur J Clin Invest. 2015;45:485–93.

5. Šimundić A-M. Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC. 2009;19:203–11.

6. Ayyub BM, Klir GJ. Uncertainty Modeling and Analysis in Engineering and the Sciences. Chapman and Hall/CRC; 2006.

7. Kallner A, Boyd JC, Duewer DL, Giroud C, Hatjimihail AT, Klee GG, et al. Expression of Measurement Uncertainty in Laboratory Medicine; Approved Guideline. Clinical and Laboratory Standards Institute; 2012.

8. Ellison SLR, Williams A. Quantifying Uncertainty in Analytical Measurement. 3rd edition. EURACHEM/CITAC; 2012.

9. M H Ramsey S L R Ellison P Rostron. Measurement uncertainty arising from sampling - A guide to methods and approaches. 2nd edition. EURACHEM/CITAC; 2019.

10. Chatzimichail T, Hatjimihail AT. A Software Tool for Calculating the Uncertainty of Diagnostic Accuracy Measures. Diagnostics (Basel). 2021;11. https://doi.org/10.3390/diagnostics11030406.

11. Schmoyer RL, Beauchamp JJ, Brandt CC, Hoffman FO. Difficulties with the lognormal model in mean estimation and testing. Environ Ecol Stat. 1996;3:81–97.

12. Bhaumik DK, Kapur K, Gibbons RD. Testing Parameters of a Gamma Distribution for Small Samples. Technometrics. 2009;51:326–34.

13. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, et al. Evaluation of measurement data --- Guide to the expression of uncertainty in measurement. 2008. https://doi.org/10.59161/JCGM100-2008E.

14. Larner AJ. The 2x2 matrix: Contingency, confusion and the metrics of binary classification. Cham: Springer International Publishing; 2024.

15. Agresti A, Franklin C, Klingenberg B. Statistics: The art and science of learning from data, global edition. 4th edition. London, England: Pearson Education; 2023.

16. Miller J, Miller JC. Statistics and Chemometrics for Analytical Chemistry. 7th edition. London, England: Pearson Education; 2018.

17. J. Aitchison JACB. The Lognormal Distribution with special reference to its uses in econometrics. Cambridge: Cambridge University Press; 1957.

18. Agresti A, Coull BA. Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. Am Stat. 1998;52:119–26.

19. Welch BL. The Generalization of `Student's' Problem when Several Different Population Variances are Involved. Biometrika. 1947;34:28–35.

20. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics. 1946;2:110–4.

21. American Diabetes Association Professional Practice Committee. 2. Diagnosis and classification of diabetes: Standards of care in diabetes-2025. Diabetes Care. 2025;48 1 Suppl 1:S27–49.

22. National Center for Health Statistics. National Health and Nutrition Examination Survey Data. Centers for Disease Control and Prevention. 2005-20016. https://wwwn.cdc.gov/nchs/nhanes/default.aspx. Accessed 4 Sept 2023.

23. Petrone S, Rousseau J, Scricciolo C. Bayes and empirical Bayes: do they merge? Biometrika. 2014;101:285–302.

24. Kelly DL, Smith CL. Bayesian inference in probabilistic risk assessment—The current state of the art. Reliab Eng Syst Saf. 2009;94:628–43.

25. Horvath AR, Bossuyt PMM, Sandberg S, John AS, Monaghan PJ, Verhagen-Kamerbeek WDJ, et al. Setting analytical performance specifications based on outcome studies - is it possible? Clin Chem Lab Med. 2015;53:841–8.

26. Hatjimihail AT. Estimation of the optimal statistical quality control sampling time intervals using a residual risk measure. PLoS One. 2009;4:e5770.

27. James H. Nichols, PhD, DABCC, FACB Sousan S. Altaie, PhD Greg Cooper, CLS, MHA Paul Glavina Abdel-Baset Halim, PharmD, PhD, DABCC Aristides T. Hatjimihail, MD, PhD et al. Laboratory Quality Control Based on Risk Management; Approved Guideline. Clinical and Laboratory Standards Institute; 2011.

28. Rao SS, Disraeli P, McGregor T. Impaired glucose tolerance and impaired fasting glucose. Am Fam Physician. 2004;69:1961–8.

29. Meneilly GS, Elliott T. Metabolic alterations in middle-aged and elderly obese patients with type 2 diabetes. Diabetes Care. 1999;22:112–8.

30. Geer EB, Shen W. Gender differences in insulin resistance, body composition, and energy balance. Gend Med. 2009;6 Suppl 1 Suppl 1:60–75.

31. Van Cauter E, Polonsky KS, Scheen AJ. Roles of circadian rhythmicity and sleep in human glucose regulation. Endocr Rev. 1997;18:716–38.

32. Colberg SR, Sigal RJ, Fernhall B, Regensteiner JG, Blissmer BJ, Rubin RR, et al. Exercise and type 2 diabetes: the American College of Sports Medicine and the American Diabetes Association: joint position statement. Diabetes Care. 2010;33:e147-67.

33. Salmerón J, Manson JE, Stampfer MJ, Colditz GA, Wing AL, Willett WC. Dietary fiber, glycemic load, and risk of non-insulin-dependent diabetes mellitus in women. JAMA. 1997;277:472–7.

34. Surwit RS, van Tilburg MAL, Zucker N, McCaskill CC, Parekh P, Feinglos MN, et al. Stress management improves long-term glycemic control in type 2 diabetes. Diabetes Care. 2002;25:30–4.

35. Pandit MK, Burke J, Gustafson AB, Minocha A, Peiris AN. Drug-induced disorders of glucose tolerance. Ann Intern Med. 1993;118:529–39.

36. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet. 2010;42:105–16.

37. Berlin KS, Williams NA, Parra GR. An introduction to latent variable mixture modeling (part 1): overview and cross-sectional latent class and latent profile analyses. J Pediatr Psychol. 2014;39:174–87.

38. Wilson JMG, Jungner G. Principles and practice of screening for disease. Geneva: World Health Organization; 1968.

39. Petersen PH, Horder M. 2.3 Clinical test evaluation. Unimodal and bimodal approaches. Scand J Clin Lab Invest. 1992;52:51–7.

40. Fischer NI, Mammen E, Marron JS. Testing for multimodality. Comput Stat Data Anal. 1994;18:499–512.

41. West RM. Best practice in statistics: The use of log transformation. Ann Clin Biochem. 2022;59:162–5.

42. Fieller EC. The distribution of the index in a normal bivariate population. Biometrika. 1932;24:428–40.

43. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A. 2002;99:6562–6.

44. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. Ann Intern Med. 2011;154:253–9.

45. Baudrit C, Couso I, Dubois D. Joint propagation of probability and possibility in risk analysis: Towards a formal framework. Int J Approx Reason. 2007;45:82–105.

46. Pawitan Y. In all likelihood: Statistical modelling and inference using likelihood. Oxford University PressOxford; 2001.

47. Dikta G, Scheer M. Bootstrap methods: With applications in R. 2021st edition. Cham, Switzerland: Springer Nature; 2021.

Appendices

A.1. List of abbreviations

DAM: diagnostic accuracy measure

OGTT: oral glucose tolerance test

ADA: American Diabetes Association

ROC: Receiver operating characteristic

A.2. Notation

A.2.1. Populations

$\overline{\mathrm{D}}$: nondiseased population

D: diseased population

A.2.2. Test outcomes

$\bar{T}$: negative test result

T : positive test result

TN : true negative test result

TP : true positive test result

FN : false negative test result

FP : false positive test result

A.2.3. Diagnostic accuracy measures

Se : sensitivity

Sp : specificity

PPV : positive predictive value

NPV : negative predictive value

ODA : overall diagnostic accuracy

DOR : diagnostic odds ratio

LR^{+}: likelihood ratio for a positive test result

LR^{-}: likelihood ratio for a negative test result

JS : Youden's index

ED : Euclidean distance

CZ : CZ

FMI: Fowlkes–Mallows index

$C\kappa$: Cohen’s kappa coefficient

$PABAK$: Prevalence-adjusted bias-adjusted kappa

$F1S$: F1 Score

MCC: Matthews correlation coefficient

A.2.4. Parameters

$\hat{\mu}_P$: estimate of the mean of the measurand of a test in the population P

$\hat{\sigma}_P$: estimate of the standard deviation of the measurand of a test in the population P

m_P: mean of the measurand of a test in a sample of the population P

s_P: standard deviation of the measurand of a test in a sample of the population P

n_P: size of a sample of the population P

v : prevalence of the disease

t : diagnostic threshold of a test

t^*: optimal diagnostic threshold of a test

p : confidence level

A.2.5. Functions and relations

$u_s(x)$: standard sampling uncertainty of x

$u_m(x)$: standard measurement uncertainty of x

$u_c(x)$: standard combined uncertainty of x

$u_i(x)$: the i^{th} component of the standard combined uncertainty of x

$f(x, \mu, \sigma)$: probability density function of a distribution with mean μ and standard deviation σ, evaluated at x

$F(x, \mu, \sigma)$: cumulative distribution function of a probability distribution with mean μ and standard deviation σ, evaluated at x

$P(a)$: probability of an event a

$P(a|b)$: probability of an event a given the event b

$CI_p(x)$: confidence interval of x at confidence level p

$Var(x)$: variance of x

$F^{-1}(\dots)$: the inverse function F

A.3. DAMs

A.3.1. Descriptions

Table A.T.1 presents short descriptions of the 16 DAMs

Table A.T.1: DAMs short descriptions

Sensitivity (Se): Probability that the test is positive among diseased individuals. Higher values reduce false negatives and support diagnosis for exclusion; range 0–1; optimised by maximization.

Specificity (Sp): Probability that the test is negative among nondiseased individuals. Higher values reduce false positives and support confirmatory diagnosis; range 0–1; optimised by maximization.

Overall Diagnostic Accuracy (ODA): Proportion of correct classifications at a given threshold; depends on disease prevalence. Targets overall correctness given the prevalence; range 0–1.

Positive Predictive Value (PPV): Probability of disease among those with a positive test result. Central for confirmatory diagnosis; increases with higher prevalence and stronger test performance; range 0–1.

Negative Predictive Value (NPV): Probability of no disease among those with negative test results. Central for diagnosis for exclusion; decreases as prevalence rises when test characteristics are fixed; range 0–1.

Diagnostic Odds Ratio (DOR): Odds of a positive test in diseased individuals divided by the odds in nondiseased individuals; does not depend on prevalence. Higher values indicate stronger separation of groups; unstable near boundary regions; lower-bounded by zero and unbounded above.

Likelihood Ratio for a Positive Test (LR^+): Multiplicative change in disease odds produced by a positive result; independent of prevalence. Useful for confirmatory diagnosis; larger values convey stronger evidence in favour of disease (context-dependent thresholds $t > 10$ are often considered strong).

Likelihood Ratio for a Negative Test (LR^-): Multiplicative change in disease odds produced by a negative result; independent of prevalence. Useful for diagnosis for exclusion; smaller values convey stronger evidence against disease (context-dependent thresholds $t < 0.1$ are often considered strong).

Youden's Index or J statistic (JS): Prevalence-invariant index that rewards simultaneous increases in Se and Sp. Common criterion for single-threshold selection; higher is better; range −1 to 1 (typically 0–1 in practice).

Euclidean Distance to the Ideal Point (ED): Distance from the ideal classifier (top-left corner) in receiver operating characteristic (ROC) space; optimised by minimization; range 0 to $\sqrt{2}$; robust to prevalence but sensitive to the joint behaviour of Se and Sp.

CZ (CZ): Measure that increases when both sensitivity and Sp are simultaneously high. Complements JS by emphasizing joint elevation of Se and Sp; prevalence-invariant; higher is better.

Fowlkes–Mallows Index (FM): Balance measure that increases with both PPV and Se. Useful when confirming positives is important while maintaining detection among diseased cases; range 0–1.

Cohen's Kappa Coefficient ($C\kappa$): Chance-corrected agreement between the test result and the true disease state. Interpretable on −1 to 1; sensitive to class imbalance; complements accuracy-type measures by adjusting for chance agreement.

Prevalence-Adjusted Bias-Adjusted Kappa ($PABAK$): Agreement index adjusted for both prevalence and marginal bias, derived from the observed agreement. Stabilises agreement estimates when prevalence is extreme, or marginals are unbalanced; range −1 to 1.

F1 Score ($F1S$): Balance measure that increases when both PPV and Se are high. Useful when both missed cases and false alarms carry consequences; range 0–1.

Matthews Correlation Coefficient (MCC): Correlation between the test classification and the disease status, accounting for all four cells of the contingency table. Robust to class imbalance; range −1 to 1; complements $C\kappa$ by not relying on an explicit chance-agreement model.

Note: ROC curve: The parametric plot of Se versus $(1 - Sp)$ as threshold t varies

A.3.2. Definitions

Tables A.T.2. and A.T.3. present mathematical definitions of the sixteen DAMs

Table A.T.2: Mathematical definitions of DAMs

measure	**natural frequency definition**	**probability definition**	**definition versus $\boldsymbol{Se, Sp, and\ v}$**
Se	$\frac{TP}{FN + TP}$	$P(T\|D)$	Se
Sp	$\frac{TN}{TN + FP}$	$P(\overline{T}\|\overline{D})$	Sp
PPV	$\frac{TP}{FP + TP}$	$P(D\|T)$	$\frac{Se\ v}{Se\ v + (1 - Sp)(1 - v)}$

NPV	$\frac{TN}{TN+FN}$	$P\left(\overline{D}\middle\|\overline{T}\right)$	$\frac{Sp\,(1-r)}{Sp\,(1-v)+(1-Se)v}$
ODA	$\frac{TN+TP}{TN+FN+TP+FP}$	$P(D)\;P(T\|D)$ $+\,P\left(\overline{D}\right)\,P\left(\overline{T}\middle\|\overline{D}\right)$	$Se\,v+Sp\,(1-v)$
DOR	$\frac{TN\,TP}{FN\,FP}$	$\frac{\frac{P(T\|D)}{P\left(\overline{T}\|D\right)}}{\frac{P\left(T\|\overline{D}\right)}{P\left(\overline{T}\|\overline{D}\right)}}$	$\frac{\frac{Se}{1-Se}}{\frac{1-Sp}{Sp}}$
LR^{+}	$\frac{TP\,(FP+TN)}{FP\,(FN+TP)}$	$\frac{P(T\|D)}{P\left(T\|\overline{D}\right)}$	$\frac{Se}{1-Sp}$
LR^{-}	$\frac{FN\,(FP+TN)}{TN\,(FN+TP)}$	$\frac{P\left(\overline{T}\|D\right)}{P\left(\overline{T}\|\overline{D}\right)}$	$\frac{1-Se}{Sp}$
JS	$\frac{TN\,TP-FN\,FP}{(TN+FP)(FN+TP)}$	$P(T\|D)+\;P\left(\overline{T}\|\overline{D}\right)-1$	$Se+Sp-1$
ED	$\sqrt{\left(\frac{FN}{FN+TP}\right)^{2}+\left(\frac{FP}{TN+FP}\right)^{2}}$	$\sqrt{P\left(\overline{T}\|D\right)^{2}+P\left(T\|\overline{D}\right)^{2}}$	$\sqrt{(1-Se)^{2}+(1-Sp)^{2}}$
CZ	$\frac{TN\;\;TP}{(TN+FP)(FN+TP)}$	$P(T\|D)\;P\left(\overline{T}\|\overline{D}\right)$	$Se\,Sp$

Table A.T.3: Mathematical definitions of DAMs

measure	natural frequency definition	probability definition	definition versus $Se, Sp, and\ v$
FMI	$\frac{TP}{\sqrt{(FN+TP)(FP+TP)}}$	$\sqrt{P(T\|D)P(D\|T)}$	$\sqrt{\frac{Se^2 v}{(1-Sp)(1-v)+Se\ v}}$
$C\kappa$	$\frac{\frac{TP+TN}{(TP+FP+TN+FN)}-\frac{(TP+FP)(TP+FN)+(FN+T}{(TP+FP+TN+FN}}{1-\frac{(TP+FP)(TP+FN)+(FN+TN)(FP+T}{(TP+FP+TN+FN)^2}}$	$P(D\|T)P(T)+\big(1-P(\bar{T}\|\bar{D})P(\bar{T})\big)P(T)+\Big(\big(1-P(D\|T)\big)P(T)+P(\bar{T}\|\bar{D})P(\bar{T})\Big)P(\bar{T})$	$\frac{2(-1+Se+Sp)(-1+v)v}{-1+Sp-(-2+Se+3Sp)v+2(-1+Se}$
$PABAK$	$\frac{TP+TN-FP-FN}{TP+FP+TN+FN}$	$2\big(P(D\|T)P(T)+P(\bar{T}\|\bar{D})P(\bar{T})\big)-1$	$-1+2(Sp(1-v)+Sev)$
$F1S$	$\frac{2TP}{FN+FP+2TP}$	$\frac{2P(T\|D)P(D\|T)}{P(T\|D)+P(D\|T)}$	$\frac{2\mathrm{Se}\ v}{1+Sp\ (-1+v)+Se\ v}$
MCC	$\frac{TP\ TN-FP\ FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	$\frac{P(D\|T)-P(D)}{\sqrt{P(D)\big(1-P(D)\big)P(T)(1-}}$	$\frac{-\big((1-Se)(1-v)\big)+Spv+(-1+\mathrm{se}+Sp}{\sqrt{(1-v)v\big((1-Sp)(1-v)+v\big)(1-v+(1}}$

Note: $P(T)=P(T|D)P(D)+P(T|\bar{D})P(\bar{D}), P(\bar{T})=P(\bar{T}|\bar{D})P(\bar{D})+P(\bar{T}|D)P(D)$

A.4.3. Classification

a) Fundamental measures

Se, Sp, PPV, NPV, ODA.

b) Composite indices

DOR, LR^{+}, LR^{-}, JS.

c) Extended indices

$FM, MCC, C\kappa, PABAK$.

d) Distance & concordance measures

$ED, CZ, F1S$.

e) Multi-axis classification

Table A.T.4. classifies the sixteen measures along multiple orthogonal axes (prevalence dependence, conditionality, and conceptual family).

Table A.T.4: Core DAMs' classification axes

Measure	Prevalence invariant	Disease-conditional	Test-conditional	Error-based	Information-based	Association-based
Sensitivity (Se)	✓	✓	—	✓	—	—
Specificity (Sp)	✓	✓	—	✓	—	—
Overall Diagnostic Accuracy (ODA)	—	—	—	✓	—	—
Positive Predictive Value (PPV)	—	—	✓	✓	—	—
Negative Predictive Value (NPV)	—	—	✓	✓	—	—
Diagnostic Odds Ratio (DOR)	✓	✓	—	—	✓	✓
Likelihood Ratio Positive (LR^{+})	✓	✓	—	—	✓	—
Likelihood Ratio Negative (LR^{-})	✓	✓	—	—	✓	—
Youden's Index (JS)	✓	✓	—	✓	—	—
Euclidean Distance (ED)	✓	✓	—	✓	—	—
CZ (CZ)	✓	✓	—	✓	—	—
Fowlkes–Mallows Index (FMI)	—	Hybrid	Hybrid	✓	—	—
Cohen's Kappa Coefficient ($C\kappa$)	—	—	—	—	—	✓
Prevalence-Adjusted Bias-Adjusted Kappa ($PABAK$)	—	—	—	—	—	✓
F1 Score ($F1S$)	—	Hybrid	Hybrid	✓	—	—
Matthews Correlation Coefficient (MCC)	—	—	—	—	—	✓

Note: Hybrid: depends jointly on Se/Sp and on PPV/NPV.

A.5. Uncertainty

A.5.1. Measurement Uncertainty

Measurement uncertainty is modelled as a function of the measurand value t. Two functional forms are considered, representing magnitude-dependent uncertainty [13]:

a) Linear:

$$u_m(t) \cong b_0 + b_1 t$$

b) Nonlinear:

$$u_m(t) = \sqrt{b_0^2 + (b_1 \cdot t)^2}$$

where b_0 is the constant contribution and $b_1$is the proportionality constant.

The model is heteroscedastic; it becomes homoscedastic for $b_1 = 0$.

These forms accommodate both additive and proportional components of imprecision, consistent with analytical performance of laboratory assays. The uncertainty of the constants b_0 and b_1 is considered negligible, although in practice they may contribute significantly if derived from limited quality-control data.

The functional form $u_m(t)$ was integrated into the parametric class-conditional models of diseased (D) and nondiseased $(\bar{D})$ populations, propagating into the uncertainty of DAMs.

A.5.2. Sampling uncertainty

Sampling uncertainty arises from the finite number of individuals in the diseased (n_D) and nondiseased $(n_{\bar{D}})$ groups. The uncertainty of estimated means $u(\hat{\mu})$, derived from the central limit theorem, and the uncertainty of variances $u^2(\hat{\sigma}^2)$, derived from the chi-squared distribution, are estimated as follows [15–17]:

$$u(\hat{\mu}) = \frac{s}{\sqrt{n}}$$

$$u^2(\hat{\sigma}^2) = \frac{2\sigma^4}{n-1}$$

A.5.3. Prevalence uncertainty

Prevalence v is modelled as the proportion $\frac{n_D}{n_{\bar{D}}+n_D}$ of diseased cases in the total study population. Uncertainty is approximated as:

$$u_s(v) \cong \sqrt{\frac{(2+n_{\bar{D}})(2+n_D)}{(4+n_{\bar{D}}+n_D)^3}}$$

according to the Agresti–Coull adjusted Wald interval, which offers improved coverage at small case numbers and extreme probabilities [18].

A.5.4. Combined uncertainty

Combined uncertainty in DAMs is propagated using the first-order Taylor expansion (delta method). Assuming uncorrelated parameters $\boldsymbol{\theta} = (x_1, x_2, \dots, x_l)$ with standard uncertainties $u_i(t)$, we have

$$u_c(t|\boldsymbol{\theta}) \approx \sqrt{\sum_{i=1}^{l} \left(\partial_{x_i} g(t|\boldsymbol{\theta})\right)^2 u_i(t)^2}$$

where $u_c(t|\boldsymbol{\theta})$ is the combined standard uncertainty of a DAM, denoted as $g(t|\boldsymbol{\theta})$.

Effective degrees of freedom are estimated using the Welch–Satterthwaite formula:

$$\nu_{eff}(t|\mathbf{\theta}) \cong \frac{u_c(t|\mathbf{\theta})^4}{\sum_{i=1}^{l} \frac{u_i(t)^4}{\nu_i}}$$

where $u_i$and $\nu_i$denote the uncertainty and degrees of freedom of each contributing component.

A.5.5 Expanded uncertainty and CIs

Expanded uncertainty is obtained by multiplying the combined standard uncertainty $u_c(t)$ by a coverage factor k determined from Student's t-distribution with $\nu_{eff}\ (t)$ degrees of freedom.

Uncertainty estimates and CIs are truncated at the respective DAM bounds.

A.6. Software availability and requirements

Program name: *DiagAccU*

Version: 1.0.0

Project home page: https://www.hcsl.com/Tools/DiagnosticAccuracy/ (accessed on September 26, 2025)

Program source: *DiagAccU.nb*

Available to download as a ZIP archive at:
https://www.hcsl.com/Tools/DiagnosticAccuracy/DiagAccU.zip (accessed on September 26, 2025)

Operating systems: Microsoft Windows 10+, Linux 3.15+, Apple macOS 11+

Programming language: Wolfram Language

Other software requirements: To run the program and read the *DiagAccUCalculations.nb* file, Wolfram Player® ver. 14.0+ is required, freely available at https://www.wolfram.com/player/ (accessed on September 26, 2025) or Wolfram Mathematica® ver. 14.3.

System requirements: Intel® i9™ or equivalent CPU and 32 GB of RAM

License: Attribution—Noncommercial—ShareAlike 4.0 International Creative Commons License

7. Permanent Citation:

Chatzimichail RA, Chatzimichail T, Hatjimihail AT. *Uncertainty Estimation of Diagnostic Accuracy Measures under Parametric Distributions.* Hellenic Complex Systems Laboratory*.* Technical Report XXIX. Hellenic Complex Systems Laboratory; 2025. Available at: https://www.hcsl.com/TR/hcsltr29/hcsltr29.pdf

8. License

First Published: September 21, 2025

Revised: September 28, 2025