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:

Accurate evaluation of diagnostic accuracy measures (DAMs) is essential for ensuring the safety and effectiveness of threshold-based screening or diagnostic tests. Conventional approaches often assume normally distributed measurands and homoscedastic measurement uncertainty, which may not reflect the true behavior of many clinical measurands. Skewed distributions, such as lognormal and gamma, and heteroscedastic measurement processes can substantially influence both DAM estimates and their associated uncertainties.

Methods:

We developed a computational framework to estimate the measurement, sampling, and combined uncertainty of 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, and effective degrees of freedom and expanded uncertainty are estimated via the Welch–Satterthwaite formula. Optimality conditions are derived analytically where applicable. The framework has been implemented in the program *DiagAccU,* in Wolfram Language, allowing parameter specification, DAM estimation and optimization, uncertainty decomposition, and graphical visualization.

Results:

Distributional assumptions had a substantial effect on DAM values and uncertainties. Measurement uncertainty predominated for specificity-related DAMs, while sampling uncertainty was more influential for sensitivity-related measures in small samples. Non-linear heteroscedasticity amplified uncertainty at extreme measurand values, shifting the diagnostic threshold that minimized combined uncertainty. Sample size increases reduced sampling uncertainty, with convergence rates varying by distribution.

Conclusions:

The framework enables realistic and flexible uncertainty quantification for DAMs by incorporating alternative distributional assumptions and heteroscedastic measurement models. This allows distribution-aware threshold optimization, supports targeted quality improvement, and improves the robustness of diagnostic performance interpretation in clinical and research settings.

Keywords:

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

1. Introduction

Medical diagnosis is the process of identifying a disease by analyzing 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.

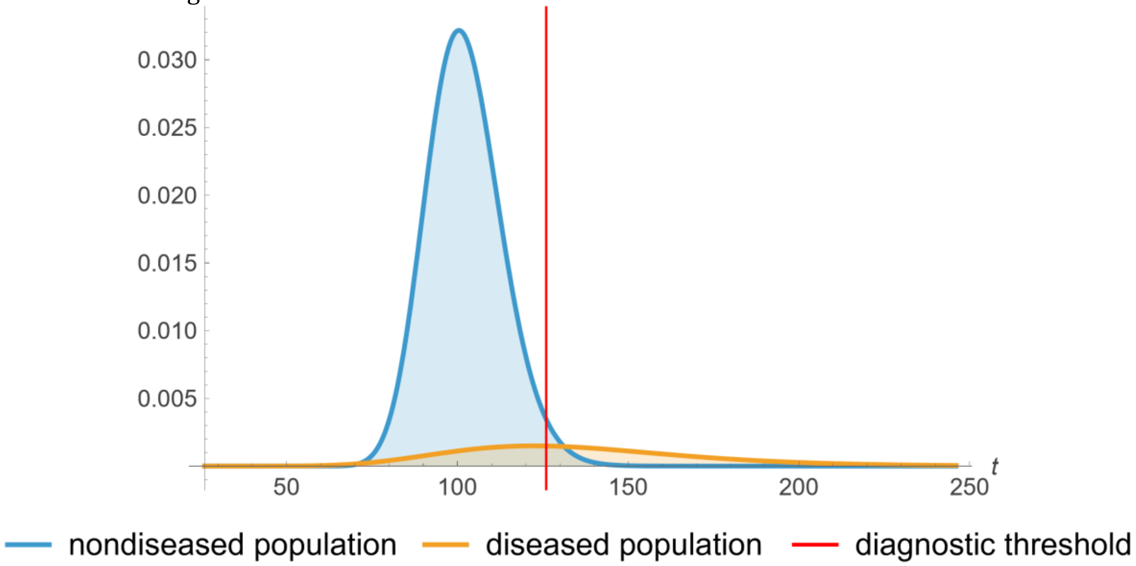


Figure 1. Probability density functions plots of fasting plasma glucose (FPG) in a diabetic (diseased) and nondiabetic (nondiseased) population.

Many in vitro screening and diagnostic tests are used as binary classifiers to partition individuals into mutually exclusive diseased (D) and nondiseased ($\bar{D}$) populations. These may be quantitative or qualitative; quantitative tests and some qualitative tests are based on direct measurement of a measurand. For a given measurand, the distributions of its values in D and $\bar{D}$ typically overlap. A diagnostic threshold t dichotomizes results: values above t indicate a positive test result (T) and values below indicate a negative result ($\bar{T}$) (or vice versa) [3] (refer to Figure 1). This operationally simple approach introduces uncertainty due to overlapping in the class-conditional distributions. Nonetheless, dichotomization has transformed clinical decision-making by mapping continuous evidence into binary actions, such as whether to initiate treatment [4].

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)

1.1. Diagnostic accuracy measures

The correctness of this threshold-based classification (refer to Table 1) is evaluated using diagnostic accuracy measures (DAMs). Although many DAMs exist [5], a smaller subset is most widely used in clinical research and practice. 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), concordance probability (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) [6] (refer to Table 2 and Appendix A.3).

Within this framework, *confirmatory diagnosis* denotes the application of a higher threshold to establish the presence of disease with high specificity, ensuring that positive results are rarely false. In contrast, *diagnosis for exclusion* denotes the application of a lower threshold to exclude disease with high sensitivity, ensuring that negative results are rarely false. These complementary strategies capture the dual clinical roles of diagnostic thresholds in laboratory medicine and provide context for interpreting DAMs in practice.

Uncertainty

From a metrological perspective, DAM uncertainty describes the dispersion of values that could reasonably be attributed to a DAM given available evidence [7]. Two primary components contribute to the combined uncertainty: measurement uncertainty and sampling uncertainty [8–10]. Both components can materially affect DAMs [11]. Uncertainty can be propagated from input parameters — sample means m_P, standard deviations s_P, prevalence v, and measurement uncertainty u_m — to DAMs values using first-order Taylor-series expansion [12].

1.1.Measurements distributions

The assumed distribution of the measurand affects both DAM estimates and uncertainties. While a normal distribution is often assumed, many measurands follow lognormal or gamma distributions due to biological and analytical factors [13, 14]. Modelling the measurand as normal, lognormal, or gamma allows uncertainty estimation under symmetric, positively skewed, or more flexible assumptions, respectively. Both linear and nonlinear heteroscedastic measurement uncertainty models can also be incorporated to better represent analytical performance [12].

2. Methods

2.1.Overview

We developed a computational framework extending prior work on uncertainty estimation for DAMs of threshold-based tests [11, 15] by incorporating three commonly encountered in laboratory medicine measurand distributional models—normal, lognormal, and gamma—and by allowing homoscedastic and both linear and nonlinear heteroscedastic measurement uncertainty functions.

This framework has been implemented in the program *DiagAccU,* which estimates the measurement, sampling, and combined uncertainties of DAMs and their CIs, applying first-order Taylor-series uncertainty propagation as specified in GUM [12] and 'Expression of measurement uncertainty in laboratory medicine' [8].

2.2.Calculations

2.2.1. Diagnostic accuracy measures

The 16 DAMS, defined in Tables 3-4, are estimated accordingly (refer to Supplemental File II: *DiagAccUCalculations.nb*).

2.2.2. Measurand distributions

For each of the diseased (D) and nondiseased ($\overline{D}$) populations, the measurand distribution $F_P(x|\boldsymbol{\theta})$ is modelled as one of: (i) normal, (ii) lognormal, and (iii) gamma. The sample parameters $\boldsymbol{\theta}$ are estimated from empirical data or set by the user. Parametric transformations ensures consistent uncertainty propagation under GUM.

2.2.3. Optimization

First-order optimality conditions are derived analytically for selected DAMs.

Table 2: Definitions of the sixteen DAM's

Sensitivity (Se): Probability that the test is positive among diseased individuals. Higher values reduce false negatives and support diagnosis for exclusion; range 0–1; optimized 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; optimized 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 0 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 favor of disease (context-dependent thresholds such as >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 such as <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; optimized by minimization; range 0 to $\sqrt{2}$; robust to prevalence but sensitive to the joint behavior of Se and Sp.

Concordance Probability (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. Stabilizes 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.

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

Table 3: Mathematical definitions of diagnostic accuracy measures

measure	natural frequency definition	probability definition	definition versus 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	$\frac{TN}{TN+FN}$	$P(\overline{D}\|\overline{T})$	$\frac{Sp\ (1-r)}{Sp\ (1-v)+(1-Se)v}$
ODA	$\frac{TN+TP}{TN+FN+TP+FP}$	$P(D)\ P(T\|D)+P(\overline{D})\ P(\overline{T}\|\overline{D})$	$Se\ v+Sp\ (1-v)$
DOR	$\frac{TN\ TP}{FN\ FP}$	$\frac{\frac{P(T\|D)}{P(\overline{T}\|D)}}{\frac{P(T\|\overline{D})}{P(\overline{T}\|\overline{D})}}$	$\frac{\frac{Se}{1-Se}}{\frac{1-Sp}{Sp}}$
LR^{+}	$\frac{TP\ (FP+TN)}{FP\ (FN+TP)}$	$\frac{P(T\|D)}{P(T\|\overline{D})}$	$\frac{Se}{1-Sp}$
LR^{-}	$\frac{FN\ (FP+TN)}{TN\ (FN+TP)}$	$\frac{P(\overline{T}\|D)}{P(\overline{T}\|\overline{D})}$	$\frac{1-Se}{Sp}$
JS	$\frac{TN\ TP-FN\ FP}{(TN+FP)(FN+TP)}$	$P(T\|D)+\ P(\overline{T}\|\overline{D})-1$	$Se+Sp-1$
ED	$\sqrt{\left(\frac{FN}{FN+TP}\right)^{2}+\left(\frac{FP}{TN+FP}\right)^{2}}$	$\sqrt{P(\overline{T}\|D)^{2}+P(T\|\overline{D})^{2}}$	$\sqrt{(1-Se)^{2}+(1-Sp)^{2}}$
CZ	$\frac{TN\ TP}{(TN+FP)(FN+TP)}$	$P(T\|D)\ P(\overline{T}\|\overline{D})$	$Se\ Sp$

2.2.4. Measurement uncertainty

Measurement uncertainty $u_m(t)$ is modelled using linear and nonlinear functions, consistent with analytical measurement guidelines [12]:

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

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

Here b_0 is the constant contribution and b_1 the proportionality constant. These forms represent homoscedastic (for $b_1 = 0$) and heteroscedastic (magnitude-dependent) uncertainty profiles commonly encountered in clinical assays.

2.2.5. Sampling uncertainty

Sampling uncertainty arises from the finite number of individuals in the diseased (n_D) and nondiseased ($n_{\overline{D}}$) 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 [16–18]:

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

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

Table 4: Mathematical definitions of diagnostic accuracy measures

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

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

Uncertainty of the prevalence or prior probability $v = \frac{n_D}{n_{\bar{D}}+n_D}$ is approximated as:

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

according tothe Agresti–Coull adjusted Wald interval, which offers improved coverage at small n and extreme probabilities [19].

2.2.6. Combined uncertainty

When there are l components of uncertainty, each with standard uncertainty $u_i(t)$, then:

$$u_c(t) = \sqrt{\sum_{i=1}^{l} u_i(t)^2}$$

2.2.7. Combined uncertainty propagation

Assuming uncorrelated parameters $\boldsymbol{\theta} = (x_1, x_2, \dots, x_l)$ with standard uncertainties $u_i(t)$ the combined standard uncertainty $u_c(t|\boldsymbol{\theta})$ of a DAM denoted as $g(t|\boldsymbol{\theta})$, is obtained via first-order Taylor-series expansion under GUM:

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

If inputs are correlated, covariance terms are included according to GUM-specified formulations. Partial derivatives $\partial_{x_i} g(t|\boldsymbol{\theta})$ are obtained analytically for each parameter x_i of $g(t|\boldsymbol{\theta})$ for each DAM and each distributional model.

2.2.8. Effective degrees of freedom and expanded uncertainty

Effective degrees of freedom ν_{eff} for the combined standard uncertainty $u_c(t)$ with l components $u_i(t)$, each with ν_i degrees of freedom, are estimated by the Welch–Satterthwaite formula [20, 21]:

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

The expanded combined uncertainty $U_c(t)$ at a confidence level p is estimated as:

$$U_c(t) \cong \left(F_\nu^{-1}\left(\frac{1-p}{2}\right)u_c(t), F_\nu^{-1}\left(\frac{1+p}{2}\right)u_c(t)\right)$$

where $F_\nu(z)$ is the Student's t-distribution cumulative distribution function with ν degrees of freedom and $u_c(t)$ is the standard combined uncertainty of a diagnostic accuracy measure.

Consequently, the confidence interval (CI) at the same confidence level p is approximated as:

$$CI_p(t) \cong \left(x + F_\nu^{-1}\left(\frac{1-p}{2}\right)u_c(t), x + F_\nu^{-1}\left(\frac{1+p}{2}\right)u_c(t)\right)$$

The estimated numerical values of the DAMs, their uncertainties and CIs are truncated to their ranges (refer to Table 2).

2.3. Software implementation

2.3.1. Software

2.3.1.1. About the program

The software program *DiagAccU* was developed in Wolfram Language, using Wolfram Mathematica® Ver 14.3 (Wolfram Research, Inc., Champaign, IL, USA).

The program is freely available as a Wolfram Language computable notebook (.nb) (Supplemental File I: *DiagAccU.nb*). It can be executed using Wolfram Player® or Wolfram Mathematica® (see Appendix A.7). Owing to the complexity of the underlying calculations, the notebook is extensive and may require substantial computational resources.

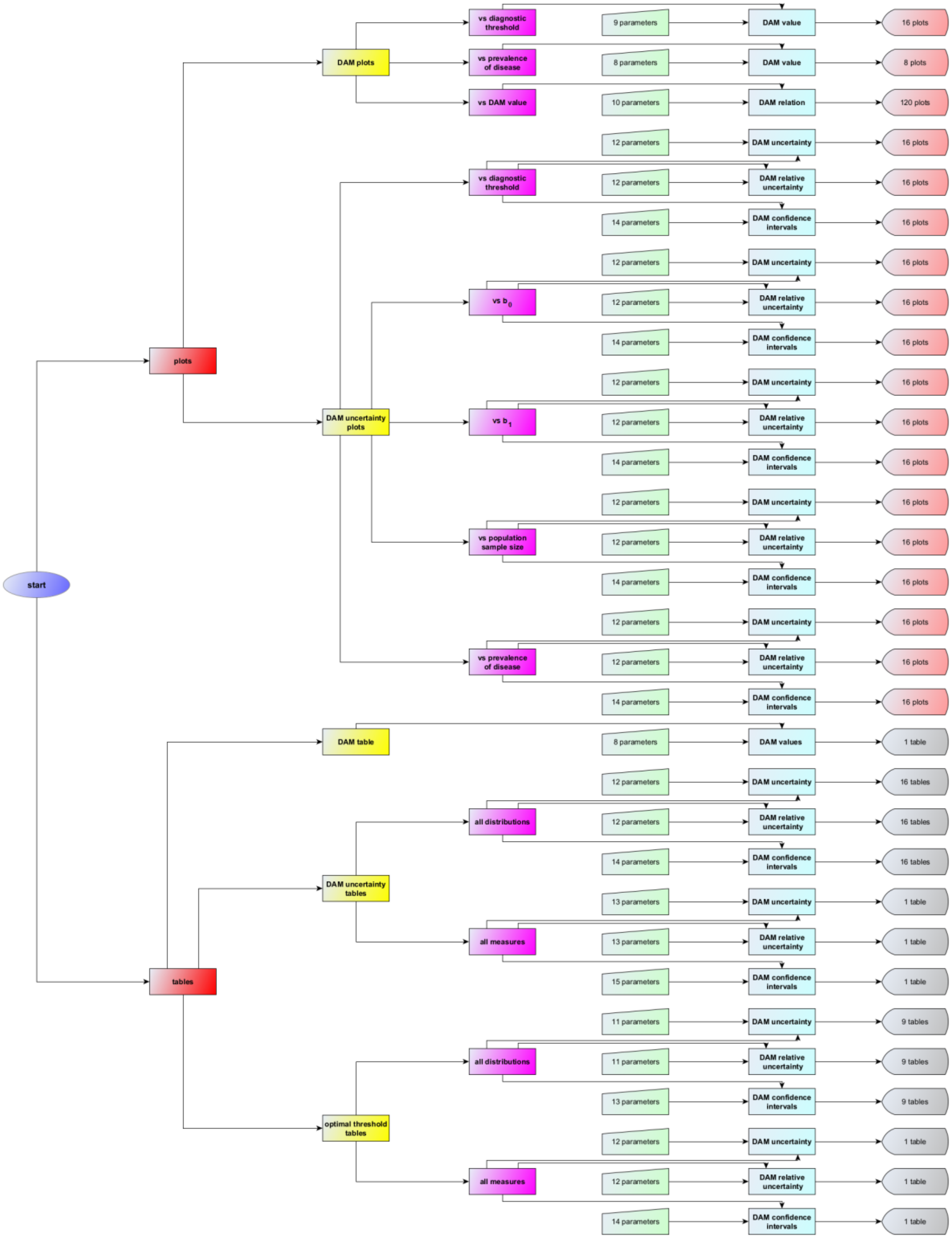


Figure 2. A simplified interface flowchart of the program *DiagAccU*.

2.3.1.2. Input of the program

2.3.1.2.1. Parametric distributions

Users select the distribution of the measurand for a diseased and nondiseased population from a predefined list of parametric distributions:

a) Normal distribution
b) Lognormal distribution
c) Gamma distribution.

2.3.1.2.2. Diagnostic accuracy measures

a) Sensitivity (Se)
b) Specificity (Sp)
c) Overall diagnostic accuracy (ODA)
d) Positive predictive value (PPV)
e) Negative predictive value (NPV)
f) Diagnostic odds ratio (DOR)
g) Likelihood ratio for a positive test result (LR^+)
h) Likelihood ratio for a negative test result (LR^-)
i) Youden's index (JS)
j) Euclidean distance (ED)
k) Concordance probability (CZ)
l) Fowlkes–Mallows index (FMI)
m) Cohen's kappa coefficient $(C\kappa)$
n) Prevalence-adjusted bias-adjusted kappa $(PABAK)$
o) $F1$ Score $(F1S)$
p) Matthews correlation coefficient (MCC)

2.3.1.2.3. Definition of populations samples statistics

For each population sample, users define its size n_P, the mean m_P, and the standard deviation s_P of the measurements. The statistics m_P and s_P are specified in arbitrary units.

2.3.1.2.4. Measurement uncertainty

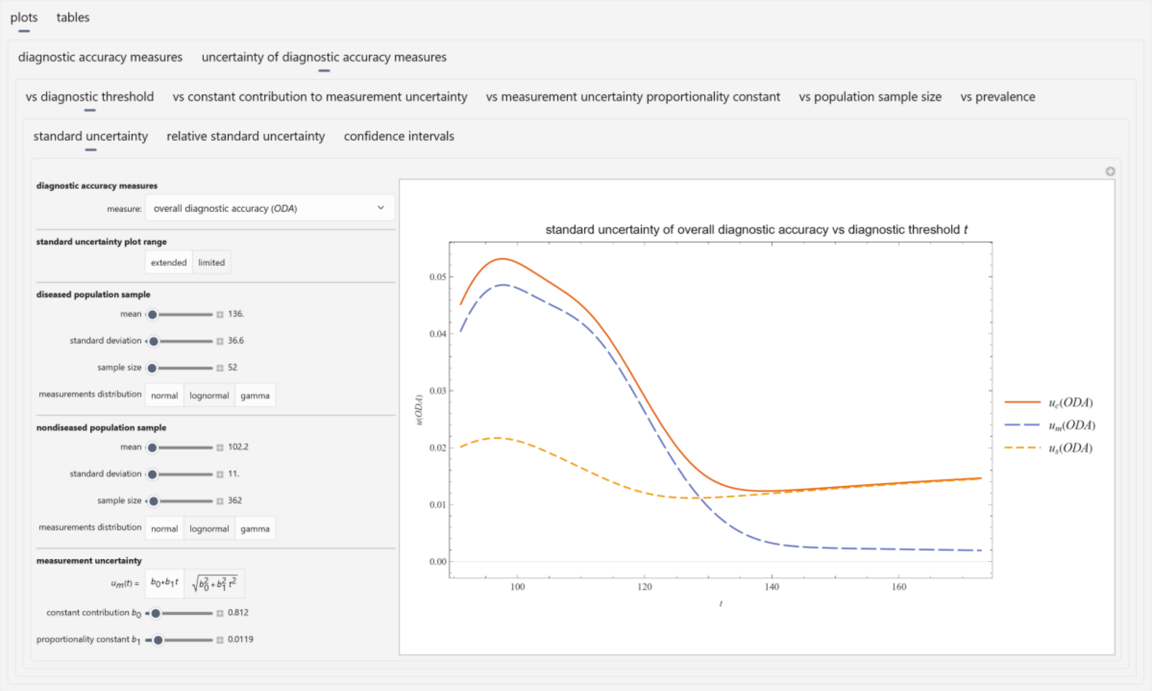
Users select a linear or nonlinear equation of the measurement uncertainty versus the value t of the measurand. They define the constant contribution b_0 to the standard measurement uncertainty, the proportionality constant b_1, and the number of quality control samples analyzed for its estimation.

For more details about the program's input, please refer to Appendix A.4.

2.3.1.3. Output

The program generates plots and tables detailing diagnostic measures, including their standard sampling, measurement, combined uncertainty, and associated CIs. By providing this extensive array of input parameters, output plots, and tables, the program presents a robust platform for exploring and comparing diagnostic accuracy measures and their uncertainties, utilizing parametric distributions of medical diagnostic measurands.

Figure 2 presents a simplified flowchart of the program interface, and Figure 3 provides a representative snapshot. More detailed documentation of the interface is available in Supplemental File III: *DiagAccUInterface.pdf*.


plots
tables
diagnostic accuracy measures
uncertainty of diagnostic accuracy measures
vs diagnostic threshold
vs constant contribution to measurement uncertainty
vs measurement uncertainty proportionality constant
vs population sample size
vs prevalence
standard uncertainty
relative standard uncertainty
confidence intervals
standard uncertainty of overall diagnostic accuracy vs diagnostic threshold t
$u(ODA)$
0.00
0.01
0.02
0.03
0.04
0.05
100
120
140
160
t
$u_c(ODA)$
$u_m(ODA)$
$u_s(ODA)$

diagnostic accuracy measures

measure: overall diagnostic accuracy (*ODA*)

standard uncertainty plot range

extended | limited

diseased population sample

mean 136.

standard deviation 36.6

sample size 52

measurements distribution: normal | lognormal | gamma

nondiseased population sample

mean 102.2

standard deviation 11.

sample size 362

measurements distribution: normal | lognormal | gamma

measurement uncertainty

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

constant contribution b_0 0.812

proportionality constant b_1 0.0119

Figure 3. A screenshot of the program *DiagAccU*.

3. Results

3.1. Illustrative case study

As previously described, we undertook an illustrative case study to demonstrate the program's application [22]. Fasting plasma glucose (FPG) was used as the diagnostic test measurand for the diagnosis of diabetes mellitus (hereafter "diabetes"), with the oral glucose tolerance test (OGTT) as the reference method. Diabetes diagnosis was confirmed if the 2-hour plasma glucose value (2-h PG), measured two hours after oral administration of 75 g of glucose during an OGTT, was equal to or greater than 200 mg/dl [23]. The study focused on individuals aged 65 to 68 years, reflecting the significant correlation between age and diabetes prevalence [24].

Data were obtained from participants in the National Health and Nutrition Examination Survey (NHANES) from 2005 to 2016 (n = 60,936), as described previously [22]. NHANES is a comprehensive survey assessing the health and nutritional status of adults and children in the United States [25].

The inclusion criteria were valid FPG and OGTT results (n = 13,836), no prior diagnosis of diabetes [26] (n = 13,465), and age 65–68 years (n = 414).

Participants with a 2-h PG measurement ≥200 mg/dL were classified as diabetic (n = 52), according to American Diabetes Association (ADA) [23].

The prevalence (prior probability) of diabetes, along with the probability distributions for FPG in both diabetic and nondiabetic individuals, were estimated using empirical Bayes methods [27], as follows:

$$v \cong \frac{52}{414} \cong 0.126$$

Table 5 presents the descriptive statistics of the FPG datasets (hereafter, FPG and its uncertainty are expressed in mg/dl).

Table 5. 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

Lognormal distributions were used to model FPG measurands in diabetic and nondiabetic participants using the maximum likelihood estimation method [28]. Parametrized for their means m_D and $m_{\bar{D}}$, and standard deviations s_D and $s_{\bar{D}}$, were defined as:

$$L_D = Lognormal(m_D, s_D) = Lognormal(136.000, 36.673)$$

$$L_{\bar{D}} = Lognormal(m_{\bar{D}}, s_{\bar{D}}) = Lognormal(102.225, 11.144)$$

Quality control data for FPG measurements in NHANES over the same period (2005–2016) included 1,350 QC samples. Nonlinear least squares regression [29, 30] provided the following function for standard measurement uncertainty $u_m(t)$ relative to the measurement value t:

$$u_m(t) = \sqrt{b_0^2 + b_1^2 t^2} = \sqrt{0.6600 + 0.00014t^2}$$

where $b_0 = 0.8124$ and $b_1 = 0.0119$.

The means of the standard measurement uncertainty of FPG of the diabetic and nondiabetic participants were estimated as:

$$\hat{u}_D \cong 1.863 \text{ mg/dL}$$

$$\hat{u}_{\bar{D}} \cong 1.469 \text{ mg/dL}$$

Consequently, the distributions of the measurands, assuming negligible measurement uncertainty, were estimated as:

$$d_D \cong Lognormal\left(m_D, \sqrt{s_D^2 - \hat{u}_D^2}\right) \cong Lognormal(136.000, 36.625)$$

$$d_{\bar{D}} \cong Lognormal\left(m_{\bar{D}}, \sqrt{s_{\bar{D}}^2 - \hat{u}_{\bar{D}}^2}\right) \cong Lognormal(102.642, 11.047)$$

Table 5 presents the descriptive statistics of the estimated lognormal distributions for diabetic and nondiabetic populations and the respective p-values from the Cramér–von Mises goodness-of-fit test [31]. Figures 4 and 5 present the estimated PDFs of FPG in the diabetic and nondiabetic populations, under the lognormal assumption (with negligible measurement uncertainty), alongside histograms of the respective NHANES datasets.

Figure 1 shows both PDFs and the ADA diagnostic threshold $t = 126$ mg/dl for FPG for diabetes [23].

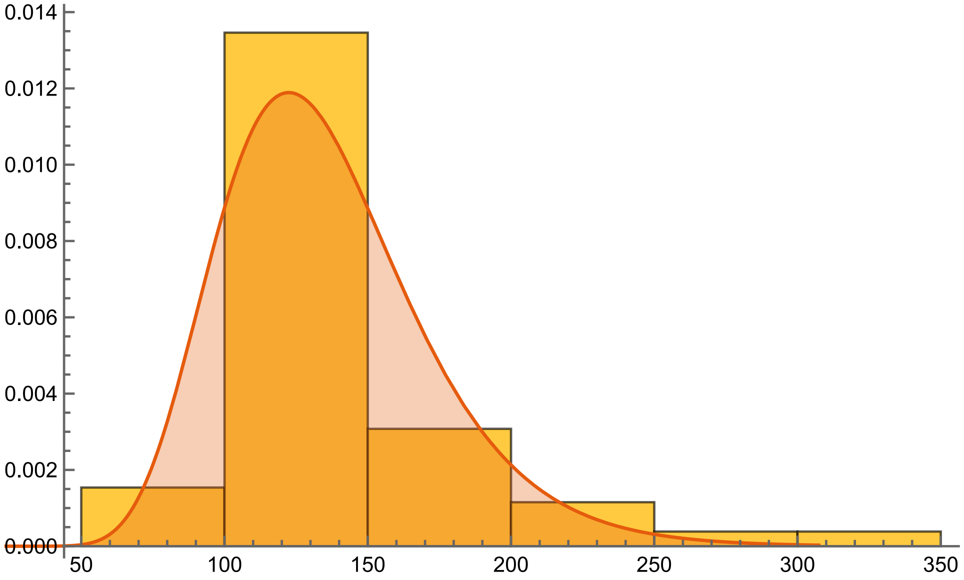


Figure 4. The estimated PDF of the FPG (mg/dL) in diabetic participants.

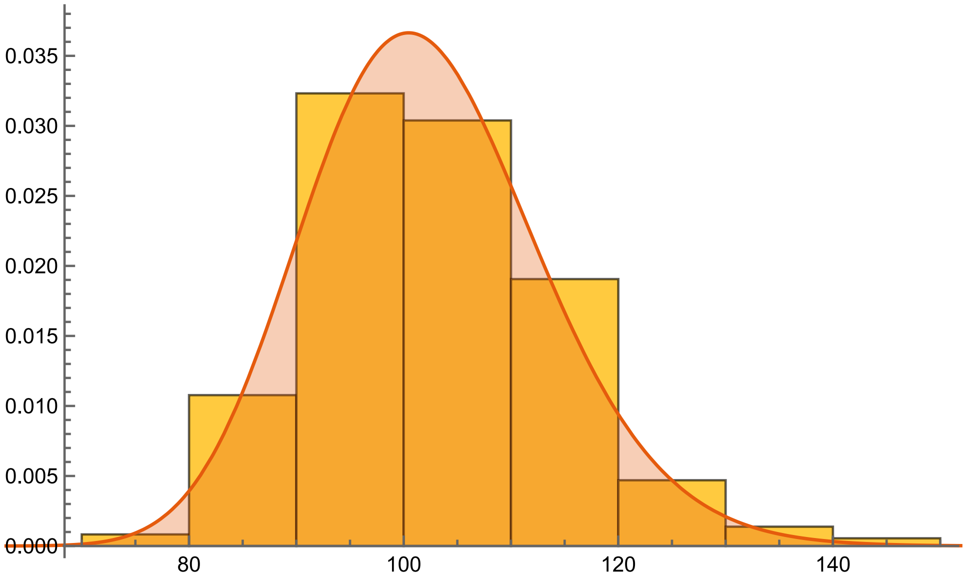


Figure 5. The estimated PDF of the FPG (mg/dL) in nondiabetic participants.

The DAMs and their uncertainties and CIs were estimated accordingly.

3.2.Application of the program

The results of the application of the program on the illustrative case study dataset are presented in Figures 6-26. Unless otherwise noted, all figures use the settings in Table 6. The selected diagnostic threshold $t = 126$ mg/dl of Figures 10-11 and 22-25 is the ADA diagnostic threshold of FPG for diabetes (refer to Figure 1).

Table 6. The settings of the program *DiagAccU* for the figures 6-26

	Units	Fig 6-9	Fig 10-11	Fig 12-13	Fig 14-17	Fig 18-21	Fig 22	Fig 23	Fig 24	Fig 25	Fig 26
t	mg/dL	91.0–173.0	126	-	91.0–173.0	91.0–173.0	126	126	126	126	-
μ_D	mg/dL	136.0	136.0	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	17.7	17.7
$\mu_{\bar{D}}$	mg/dL	102.2	102.2	102.2	102.2	102.2	102.2	102.2	102.2	102.2	102.2
$\sigma_{\bar{D}}$	mg/dL	11.0	11.0	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	0.126	0.126
b_0		-	-	-	0.8124	0.8124	-	0.8124	0.8124	0.8124	0.8124
b_1		-	-	-	0.0119	0.0119	-	0.0119	0.0119	0.0119	0.0119
p		-	-	-	-	0.95	-	0.95	0.95	0.95	0.95
d_D		lognormal					normal lognormal gamma	lognormal	normal	gamma	lognormal
$d_{\bar{D}}$		lognormal					normal lognormal gamma	lognormal	normal	gamma	lognormal

3.2.1. Threshold-dependent behavior of diagnostic accuracy measures

Figures 6–9 show the threshold t profiles for the 16 DAMs under the lognormal model and the parameter settings of Table 6. As expected for a "higher-is-positive" rule, Se decreases monotonically with t, whereas Sp increases monotonically; their intersection maps the classical trade-off between false negatives and false positives.

PPV increases with t and NPV decreases, reflecting the changing post-test probabilities as the diagnostic threshold moves along the ROC curve at fixed prevalence. ODA exhibits a single interior maximum where the weighted densities $(1 - v)F_{\bar{D}}(t)$ and $v\ F_D(t)$ balance, while JS attains its maximum near the point where $F_{\bar{D}}(t) = F_D(t)$, consistent with its prevalence-invariance. Euclidean distance (ED) to ROC space (0,1) is minimized at a nearby threshold, and $CZ\ =\ Se\ \cdot\ Sp$ reaches a peak where both components are simultaneously large rather than extreme in only one direction.

DOR and LR^{+} increase rapidly as Sp approaches 1 (right-tail thresholds), whereas LR^{-} declines toward 0 as Se approaches 1 (left-tail thresholds). Composite association measures (*FM, F1S, Cκ, PABAK, MCC*) peak at interior thresholds that balance true-positive yield ($\frac{TP}{TP+FN}$) against false discoveries $(FP + FN)$(Figures 6–9).

3.2.2. Prevalence sensitivity analyses

Varying v while holding the fitted distributions fixed (Figures 10-11) show that *PPV, NPV, ODA, F1S*, and *FMI* are substantially prevalence-dependent: *PPV* rises and *NPV* falls with increasing v, with ODA exhibiting a concave response whose maximum can shift in t. In contrast, $Se, Sp, JS, ED, CZ, LR^{+}, LR^{-}$, $DOR, C\kappa, PABAK$, and MCC retain their prevalence invariance by definition. These patterns confirm the multi-axis taxonomy summarized in Appendix A.3.

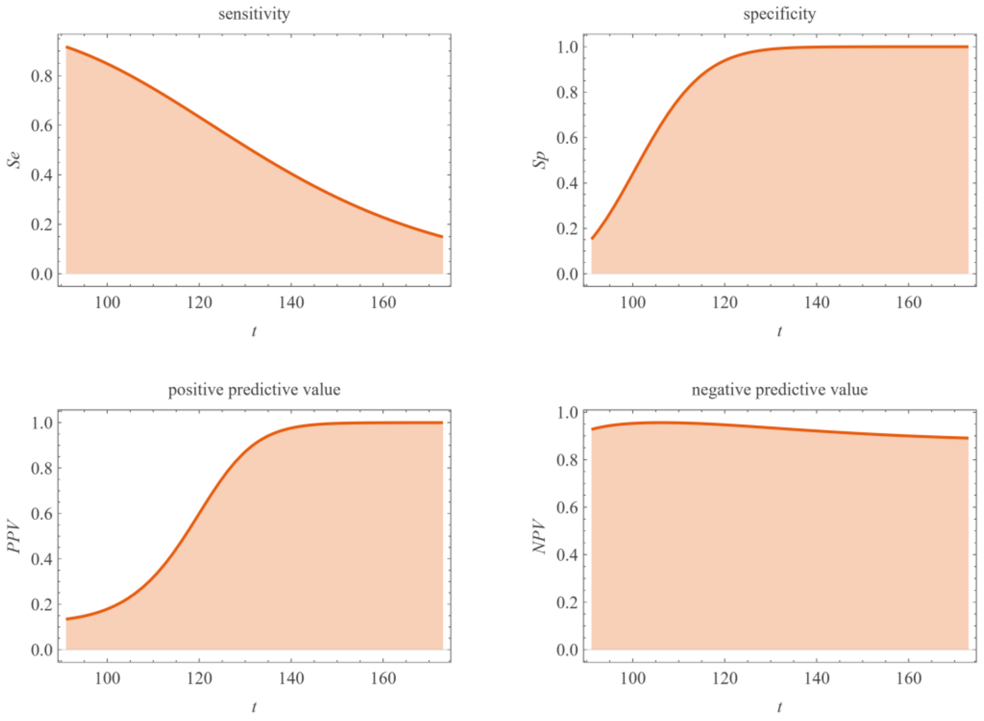


Figure 6. Sensitivity (Se), specificity (Sp), positive predictive value (PPV), and negative predictive value (NPV) versus threshold $t.$

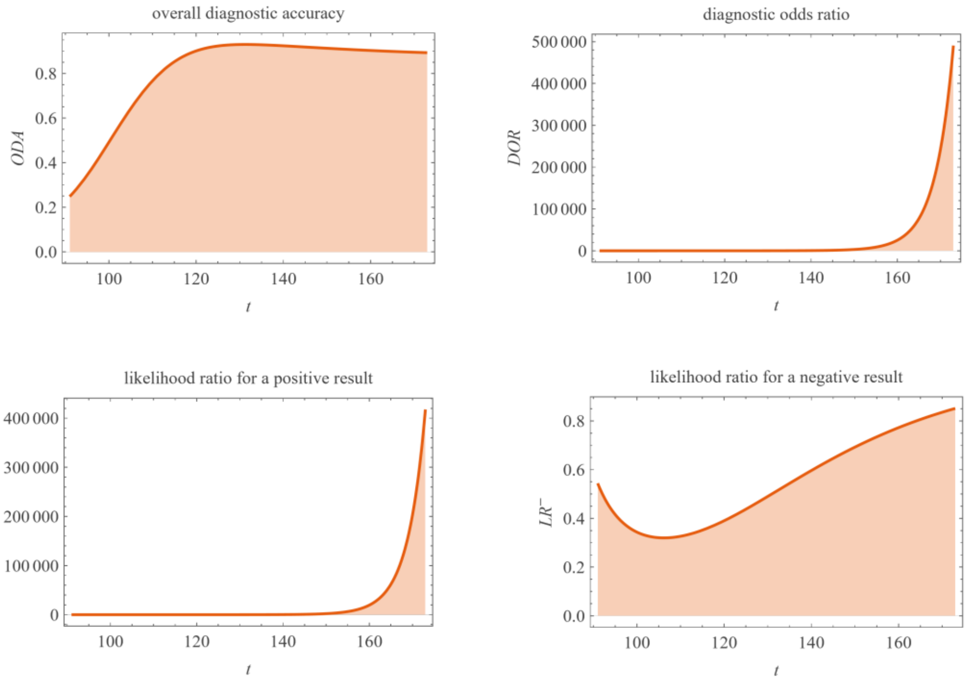


Figure 7. Overall diagnostic accuracy (ODA), diagnostic odds ratio (DOR), likelihood ratio for a positive result (LR^+*),* likelihood ratio for a negative result (LR^-) versus threshold $t.$

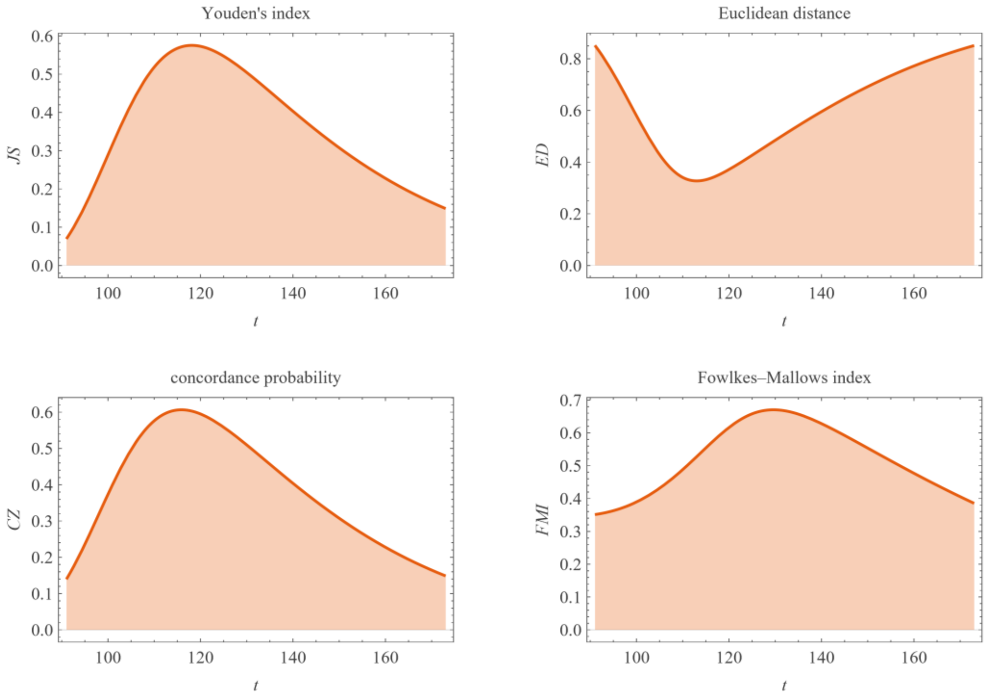


Figure 8. Youden's index (JS), Euclidean distance (ED), concordance probability (CZ), and Fowlkes–Mallows index (FMI) versus threshold t.

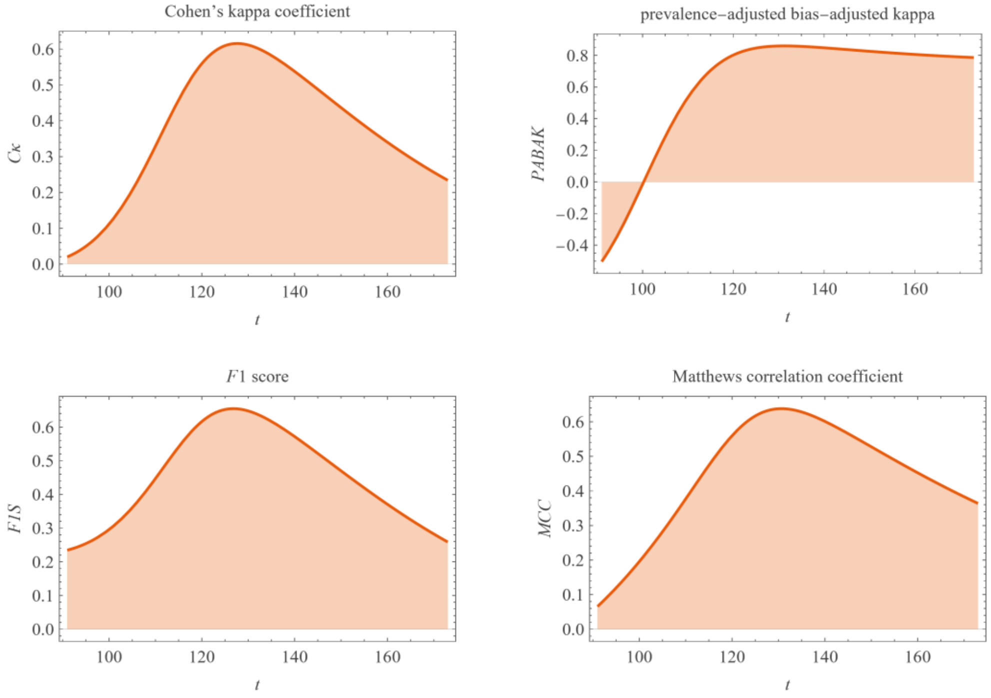


Figure 9. Cohen's kappa coefficient ($C\kappa$), prevalence-adjusted bias-adjusted kappa ($PABAK$), $F1$ score ($F1S$), and Matthews correlation coefficient (MCC) versus threshold $t.$

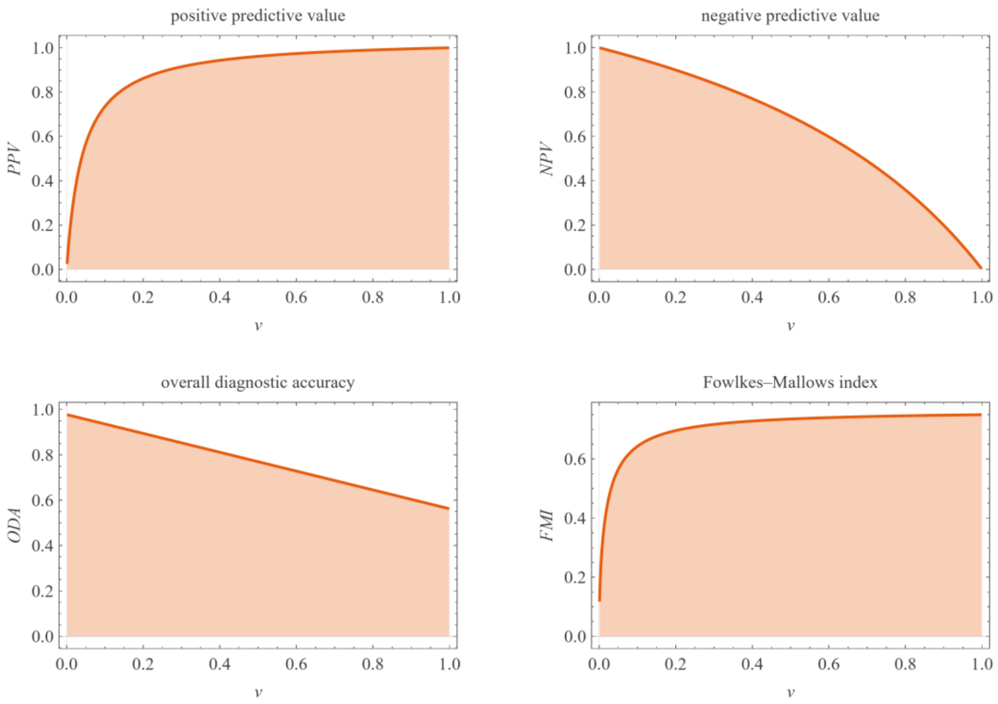


Figure 10. Positive predictive value (PPV), negative predictive value (NPV), overall diagnostic accuracy (ODA), and Fowlkes–Mallows index (FMI) versus prevalence of diabetes v.

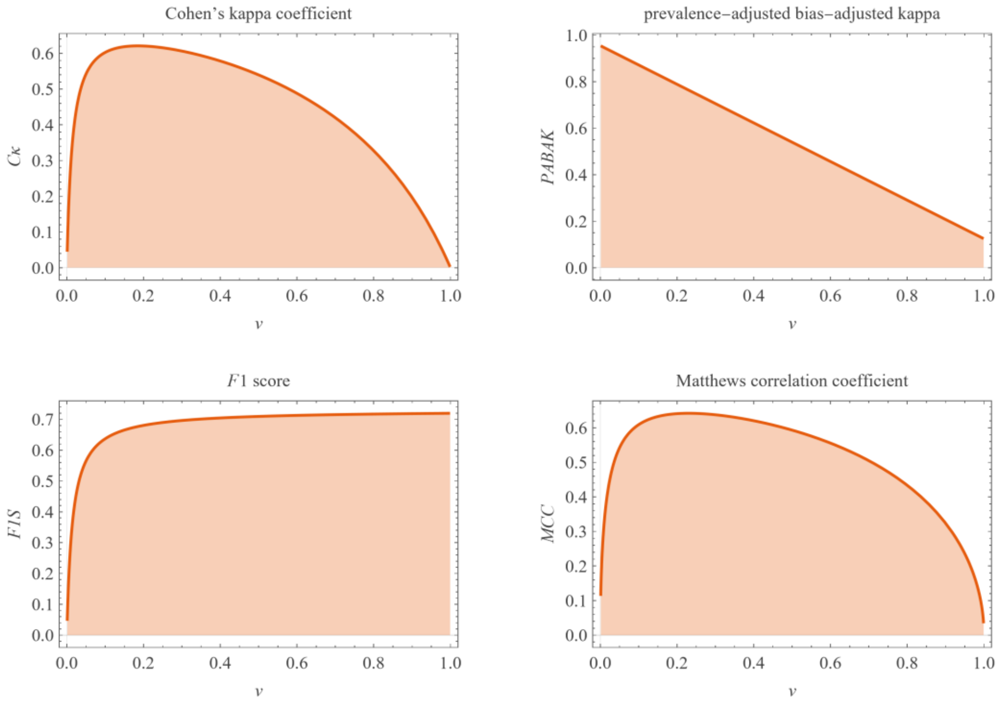


Figure 11. Cohen's kappa coefficient ($C\kappa$), prevalence-adjusted bias-adjusted kappa ($PABAK$), $F1$ score ($F1S$), and Matthews correlation coefficient (MCC) versus prevalence of diabetes v.

3.2.3. DAM relations

Figures 12 and 13 present eight interesting relations between pairs of diagnostic accuracy measures among the 120 provided by the program. Figure 12 shows that ODA increases with JS but is not one-to-one because ODA depends on prevalence. As ODA weights Se and Sp by disease prevalence, the ODA interior maximum appears near the weighted density balance, whereas JS peaks close to the unweighted crossing. $C\kappa$ versus JS is sharply nonlinear with large dynamic range at low JS. ED and CZ show a strong inverse relation (approach to the ROC ideal point reduces ED and increases CZ). $F1S$ increases with FMI and is typically smaller, consistent with harmonic $\leq$ geometric mean.

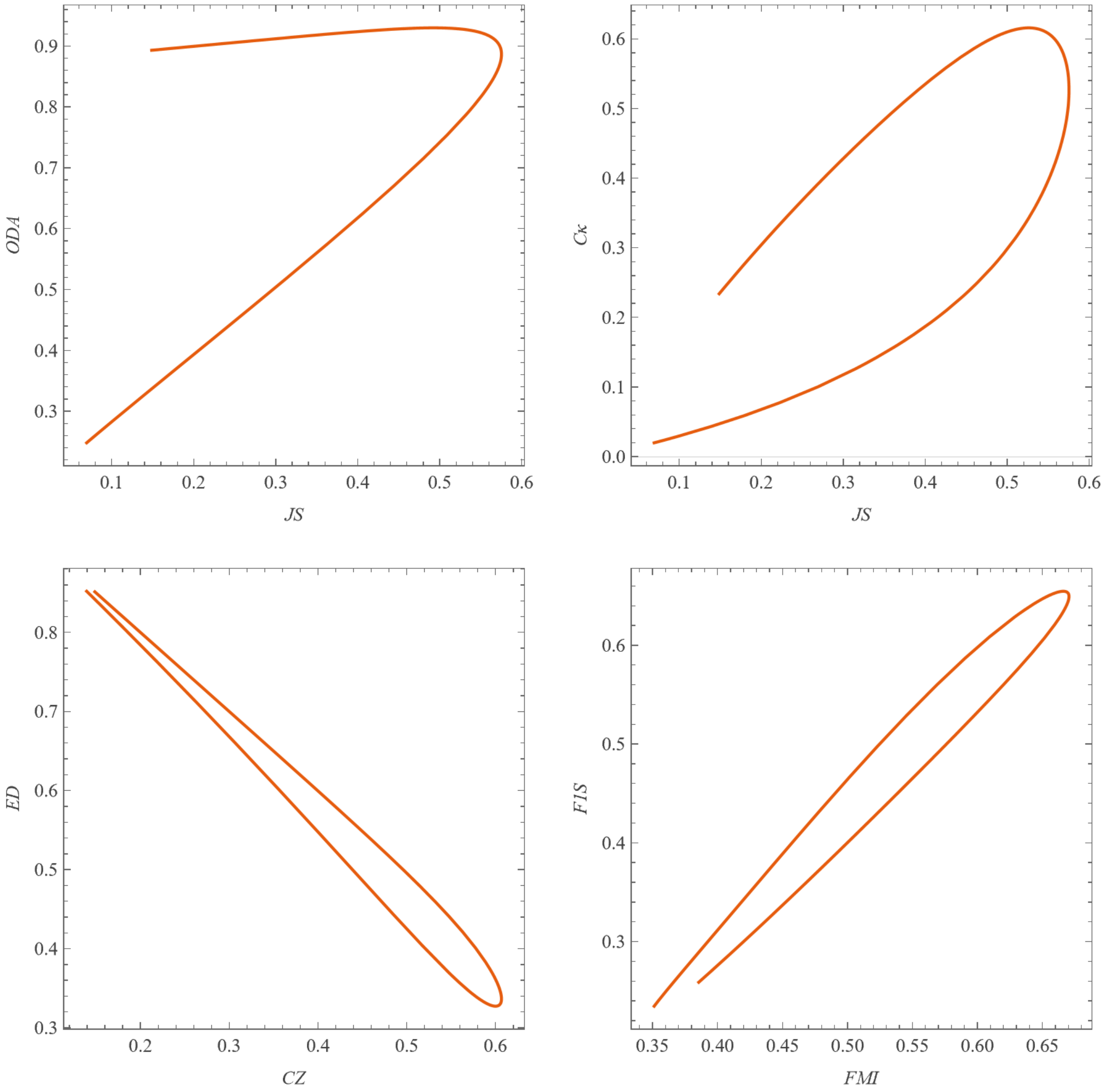


Figure 12. Plots of overall diagnostic accuracy (ODA) and Cohen's kappa coefficient ($C\kappa$) versus Youden's index (JS), Euclidean distance (ED) versus concordance probability (CZ), and $F1$ score ($F1S$) versus Fowlkes–Mallows index (FMI).

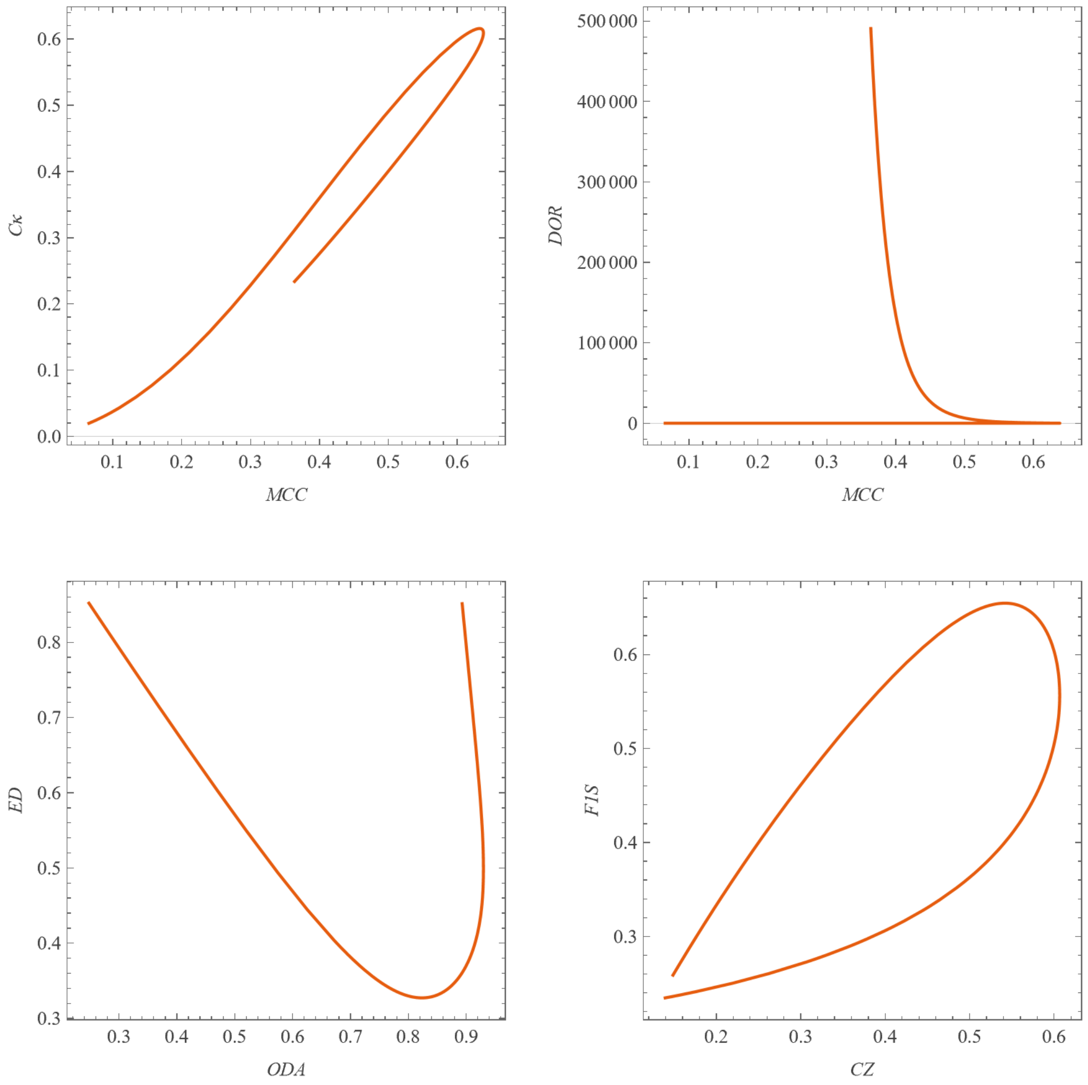


Figure 13. Plots of Cohen's kappa coefficient ($C\kappa$) and diagnostic odds ratio (DOR) versus Matthews correlation coefficient (MCC), positive predictive value (PPV) versus overall diagnostic accuracy (ODA), and concordance probability (CZ) versus Fowlkes–Mallows index (FMI).

3.2.4. Uncertainty decomposition across thresholds

Figures 14–17 partition the relative standard uncertainties of each DAM into sampling, measurement, and combined components.

Three regularities emerge:

a) Sp-anchored measures (e.g., LR^{+} at high t, DOR when $Sp \rightarrow 1$) are dominated by measurement uncertainty near the right tail, where small absolute errors in t propagate strongly through $(1 - Sp)$ in denominators.

b) Se-anchored measures (e.g., LR^{-} at low t) show larger sampling uncertainty near the left tail when the effective number of diseased positives becomes small.

c) Prevalence-dependent post-test measures (PPV and NPV) exhibit heightened combined uncertainty in regions where the opposite class is rare (e.g., PPV near low v unless t is stringently high), consistent with binomial variance scaling and the curvature of Bayes' formula.

Nonlinear heteroscedasticity ($u_m(t)$ increasing with t) amplifies the right-tail measurement component and can shift the combined-uncertainty minimum away from the JS-optimal threshold.

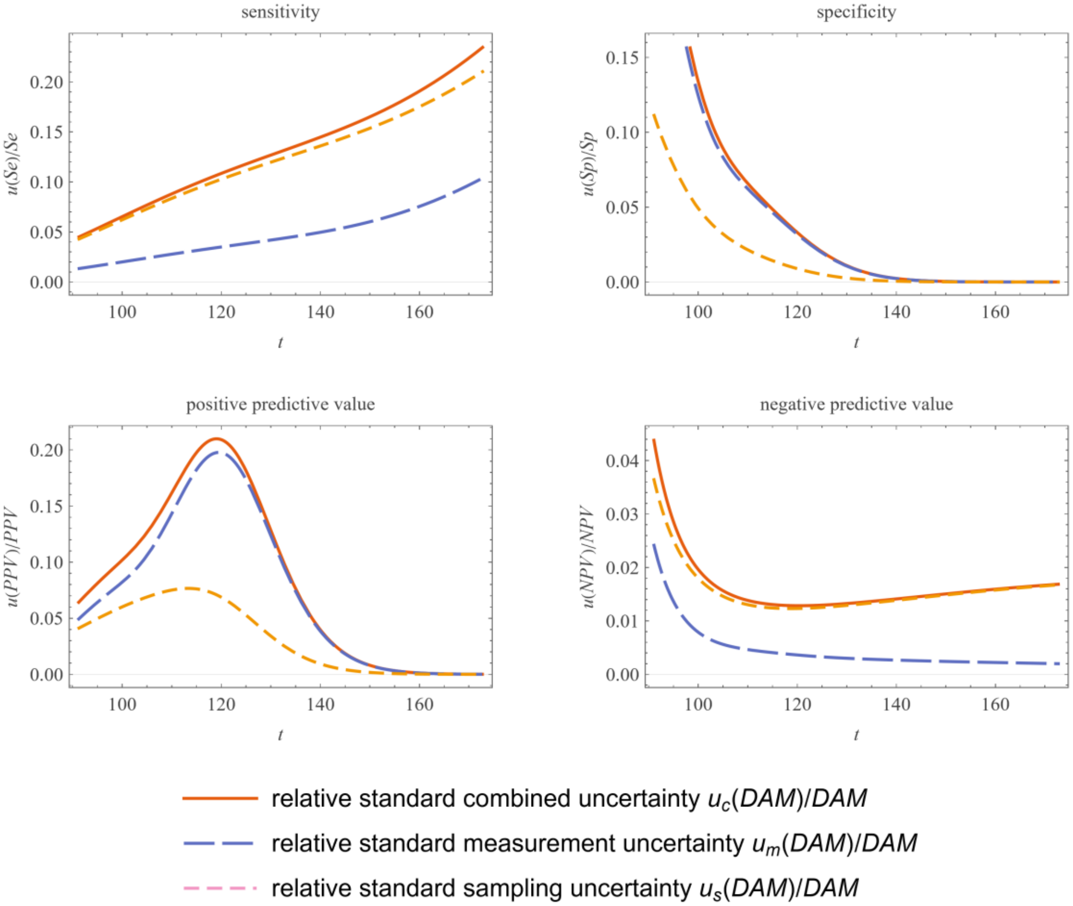


Figure 14. Relative standard sampling, measurement, and combined uncertainty of sensitivity (Se), specificity (Sp), positive predictive value (PPV), and negative predictive value (NPV) versus threshold t.

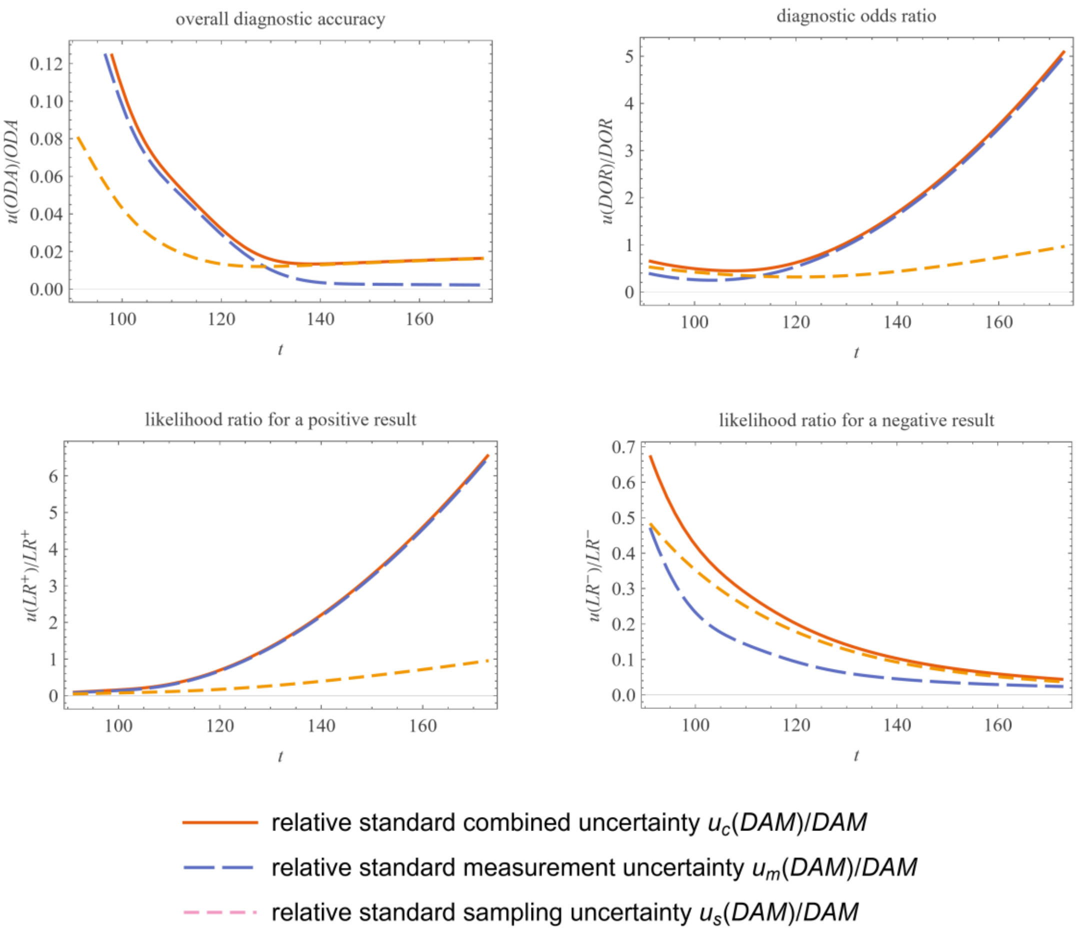


Figure 15. Relative standard sampling, measurement, and combined uncertainty of overall diagnostic accuracy (ODA), diagnostic odds ratio (DOR), likelihood ratio for a positive result (LR^{+}), and likelihood ratio for a negative result (LR^{-}) versus threshold t.

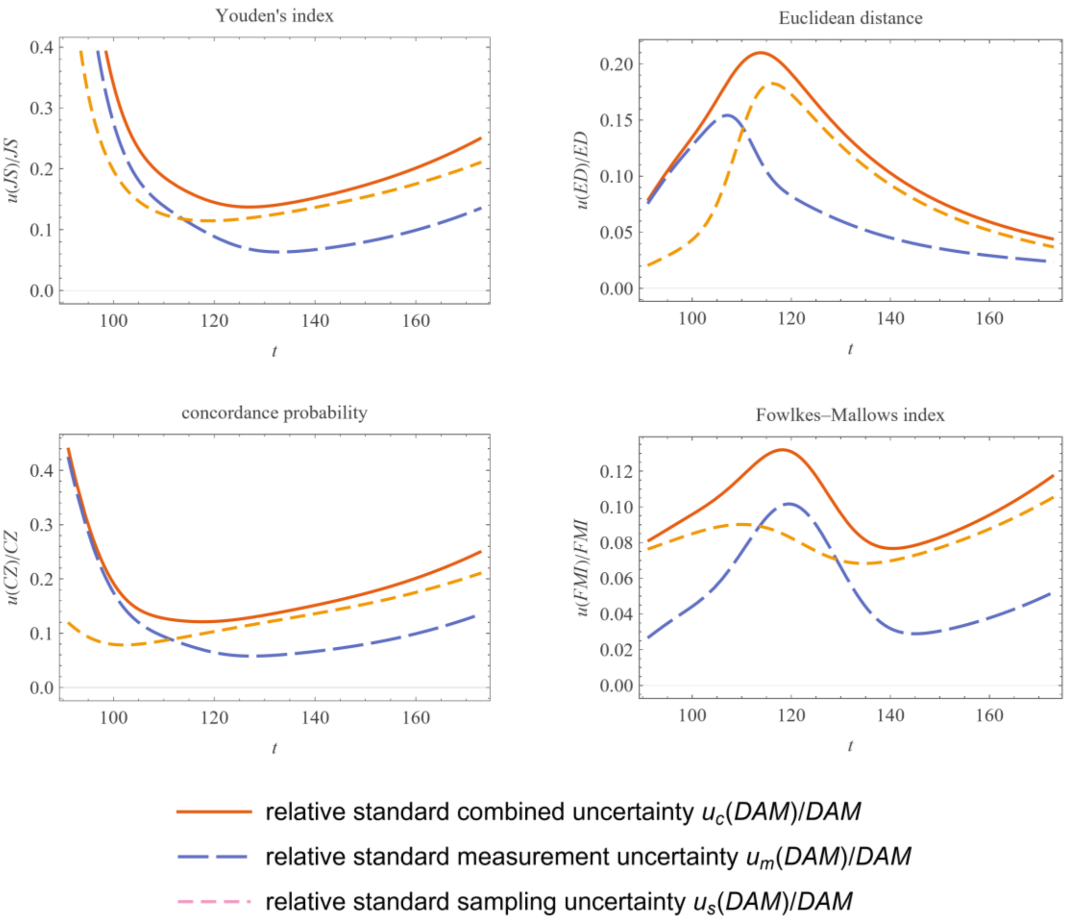


Figure 16. Relative standard sampling, measurement, and combined uncertainty of Cohen's kappa coefficient ($C\kappa$), prevalence-adjusted bias-adjusted kappa ($PABAK$), $F1$ score ($F1S$), and Matthews correlation coefficient (MCC) versus threshold t.

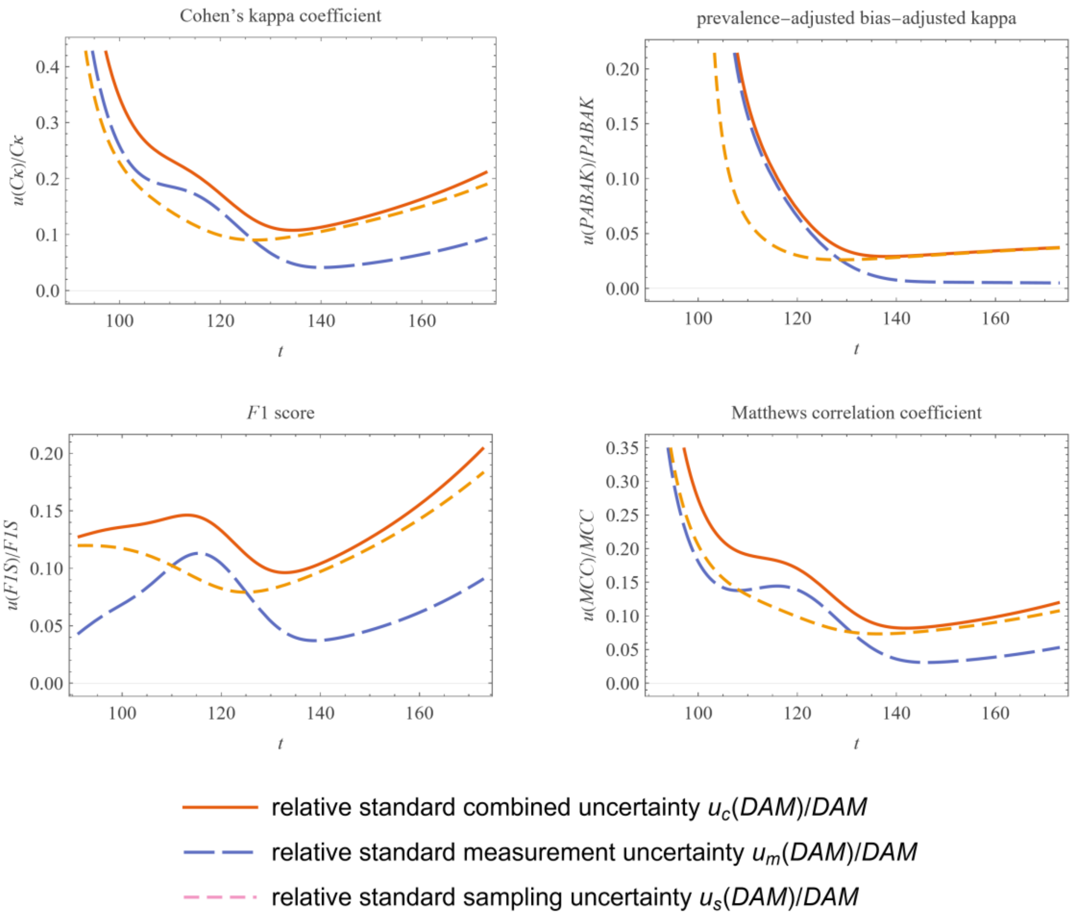


Figure 17. Relative standard sampling, measurement, and combined uncertainty of Cohen's kappa coefficient ($C\kappa$), prevalence-adjusted bias-adjusted kappa ($PABAK$), $F1$ score ($F1S$), and Matthews correlation coefficient (MCC) versus threshold t.

3.2.5. Confidence intervals

Figures 18–21 present 95% CIs for the DAMs as a function of t. CIs widths broaden in tail regions where one of Se or Sp approaches boundary values, with the largest expansions observed for ratio measures (LR^{+}, LR^{-}, DOR) due to denominator instability, and for PPV and NPV under mismatched prevalence and diagnostic threshold.

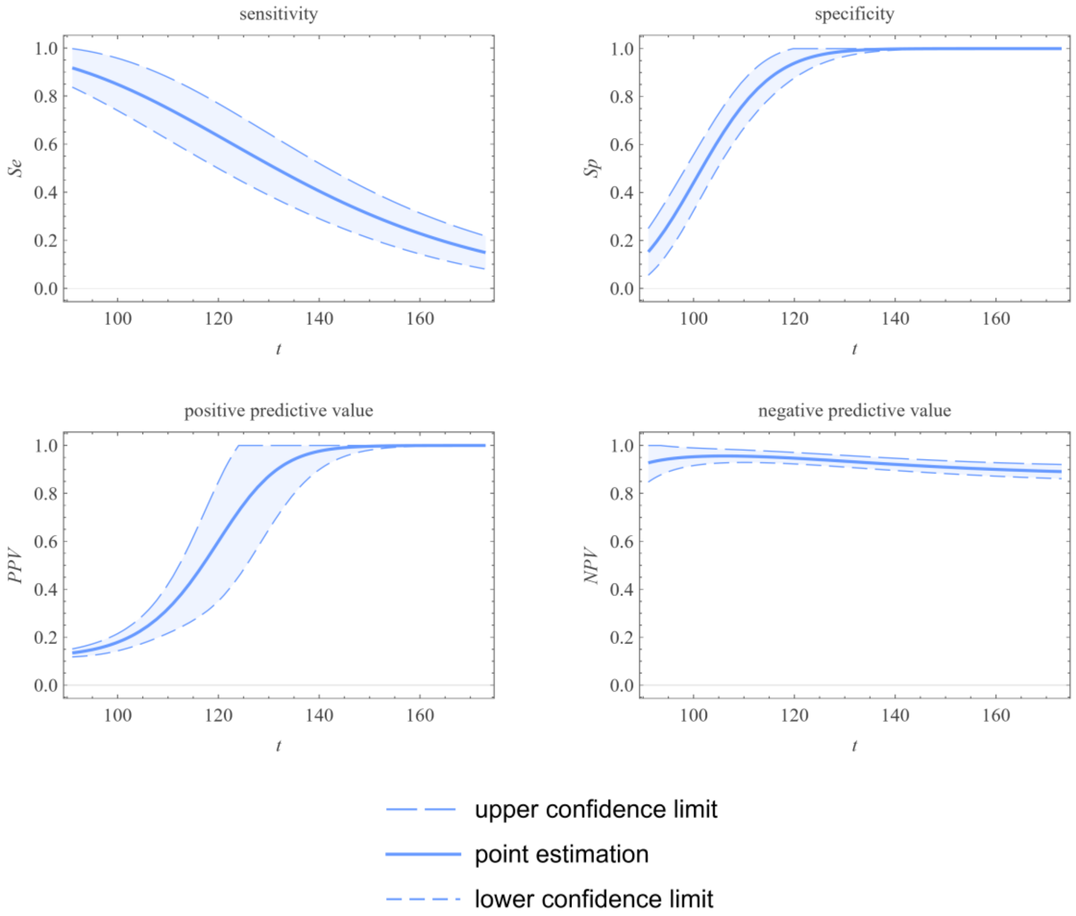


Figure 18. Confidence intervals of sensitivity (Se), specificity (Sp), positive predictive value (PPV), and negative predictive value (NPV) versus threshold t.

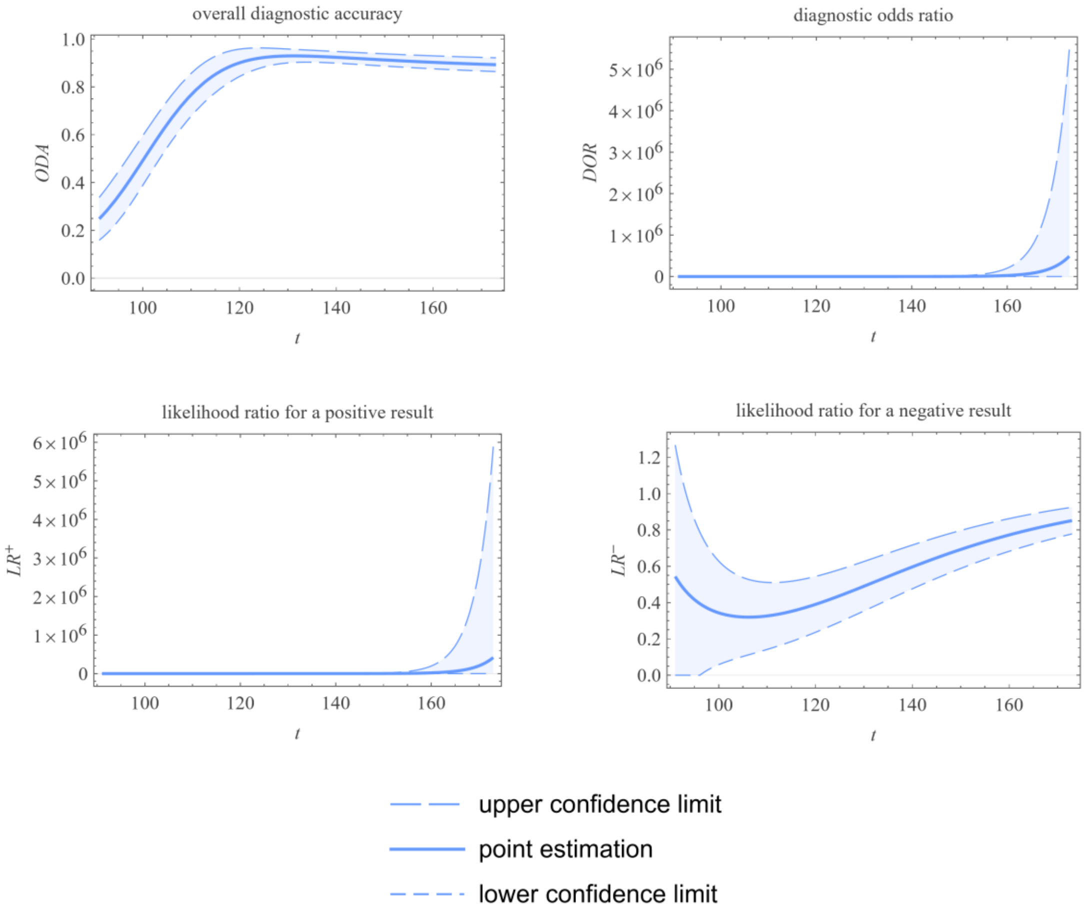


Figure 19. Confidence intervals of overall diagnostic accuracy (ODA), diagnostic odds ratio (DOR), likelihood ratio for a positive result (LR^+) and likelihood ratio for a negative result (LR^-) versus threshold t.

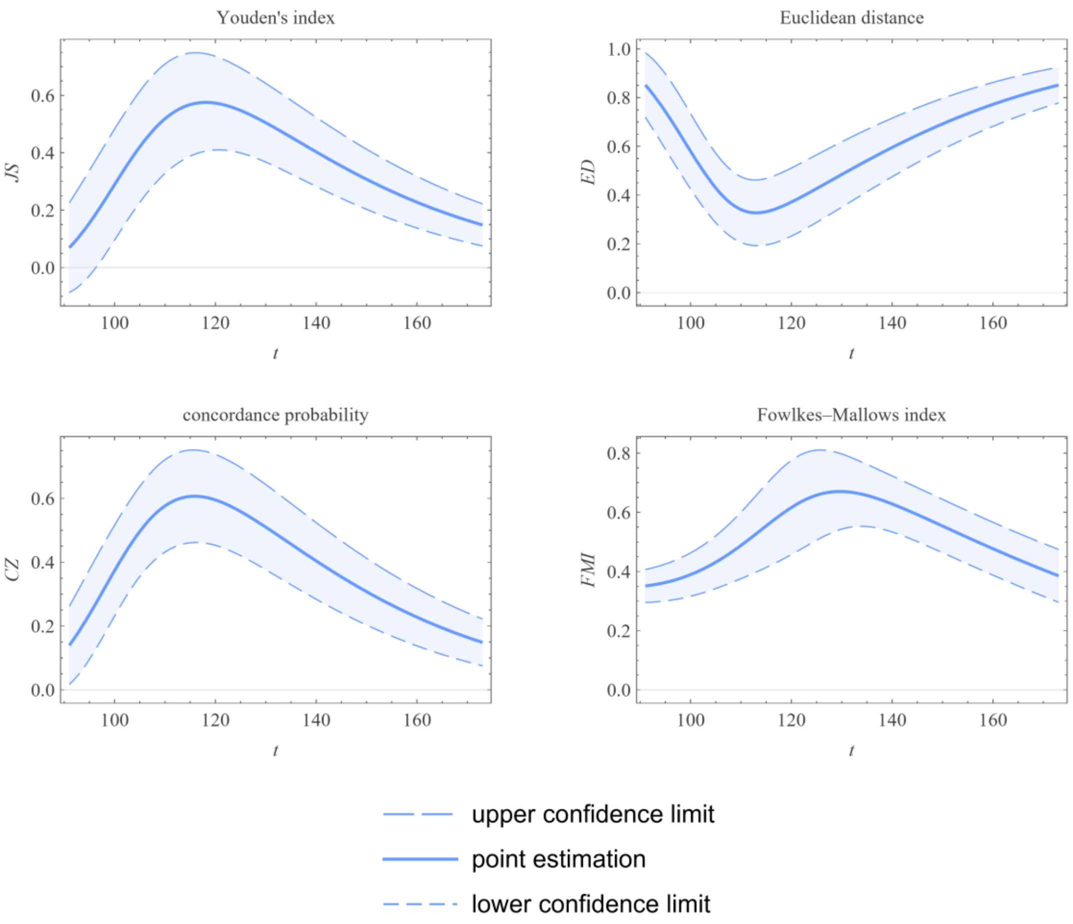


Figure 20. Confidence intervals of Youden's index (JS), Euclidean distance (ED), concordance probability (CZ), and Fowlkes–Mallows index (FMI) versus threshold t.

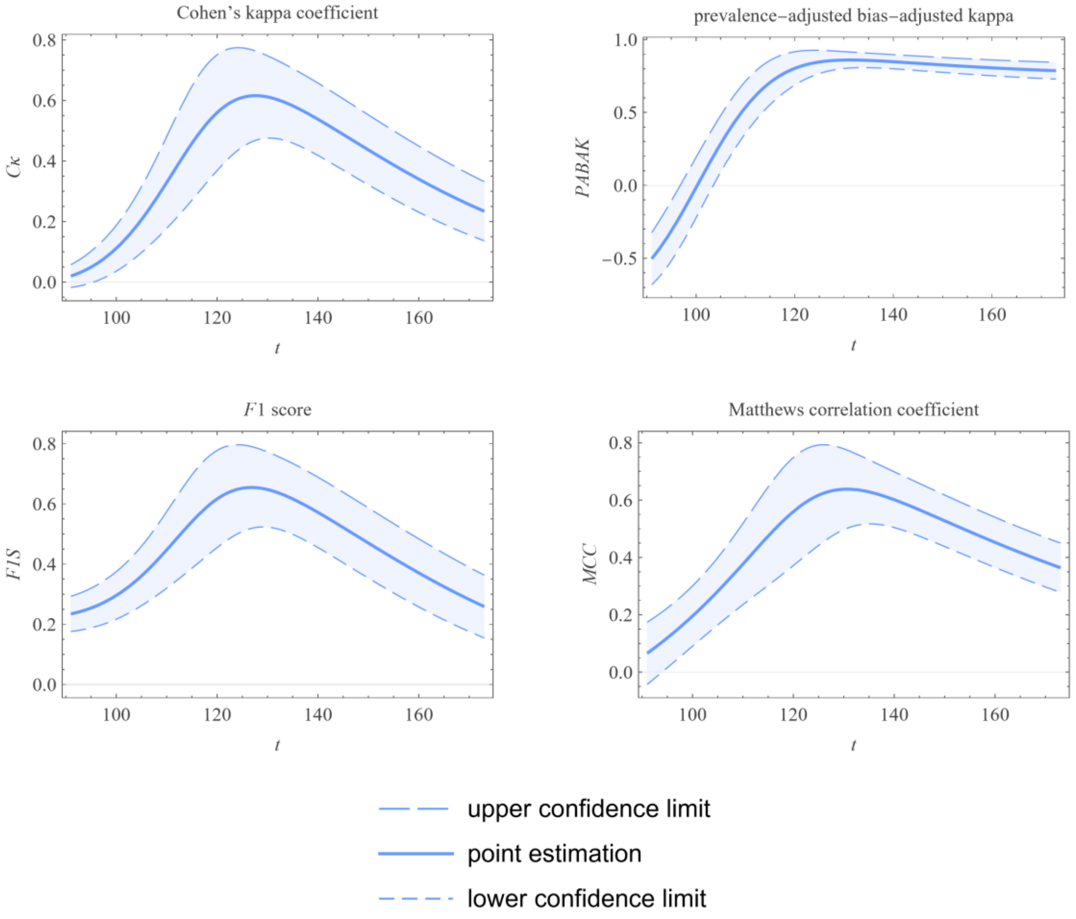
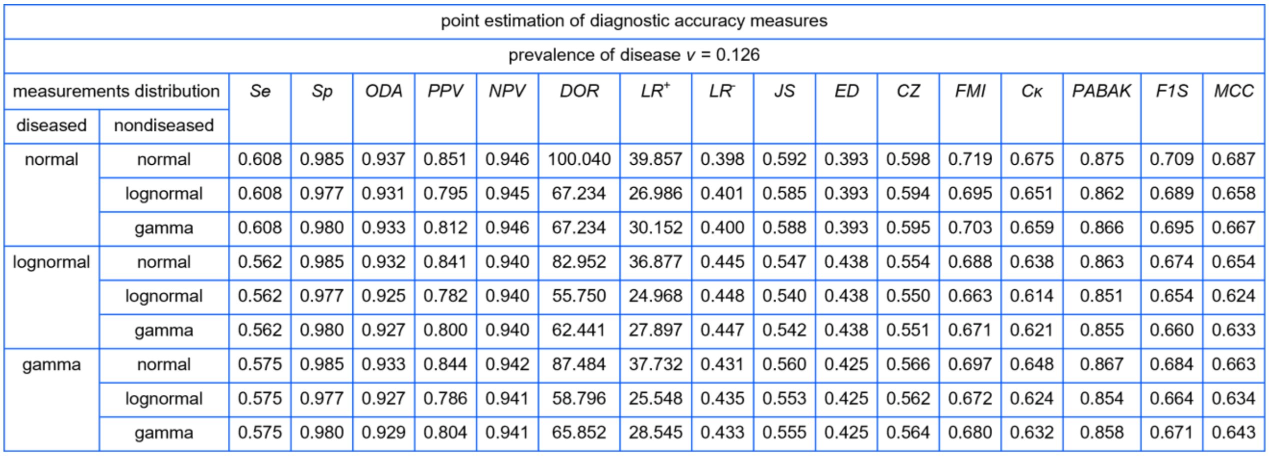


Figure 21. Confidence intervals of Cohen's kappa coefficient ($C\kappa$), prevalence-adjusted bias-adjusted kappa ($PABAK$), $F1$ score ($F1S$), and Matthews correlation coefficient (MCC) versus threshold t.

3.2.6. Point estimates and confidence intervals at a clinically relevant diagnostic threshold, and robustness to distributional assumptions

At the conventional screening diagnostic threshold $t = 126$ mg/dL, Figure 22 reports point estimates and Figure 23 95% CIs for all 16 DAMs under the fitted lognormal model and prevalence $v \cong 0.126$. These results highlight the asymmetric information content of the diagnostic threshold: PPV remains modest because diabetes is relatively uncommon in the target age band, whereas NPV is high; association measures ($C\kappa, PABAK, MCC$) and concordance-geometry indices ($JS, ED, CZ, FM, F1S$) fall in the mid-range, indicating imperfect but clinically useful discrimination.

To examine robustness to modeling assumptions, Figures 24 and 25 present parallel analyses under normal and gamma distributions (parameterized to match the observed means and standard deviations). Overall rankings and qualitative threshold behavior were largely preserved across parametric forms, reproducing the profiles seen in Figures 6–9. However, prevalence-dependent measures ($PPV, NPV, ODA, F1S, FM$) and ratio indices (LR^{+}, DOR) showed greater sensitivity, with confidence-interval widths changing by several mg/dL in some scenarios. These findings underscore the need for distributional diagnostics and, where indicated, mixture or nonparametric alternatives to mitigate misspecification.

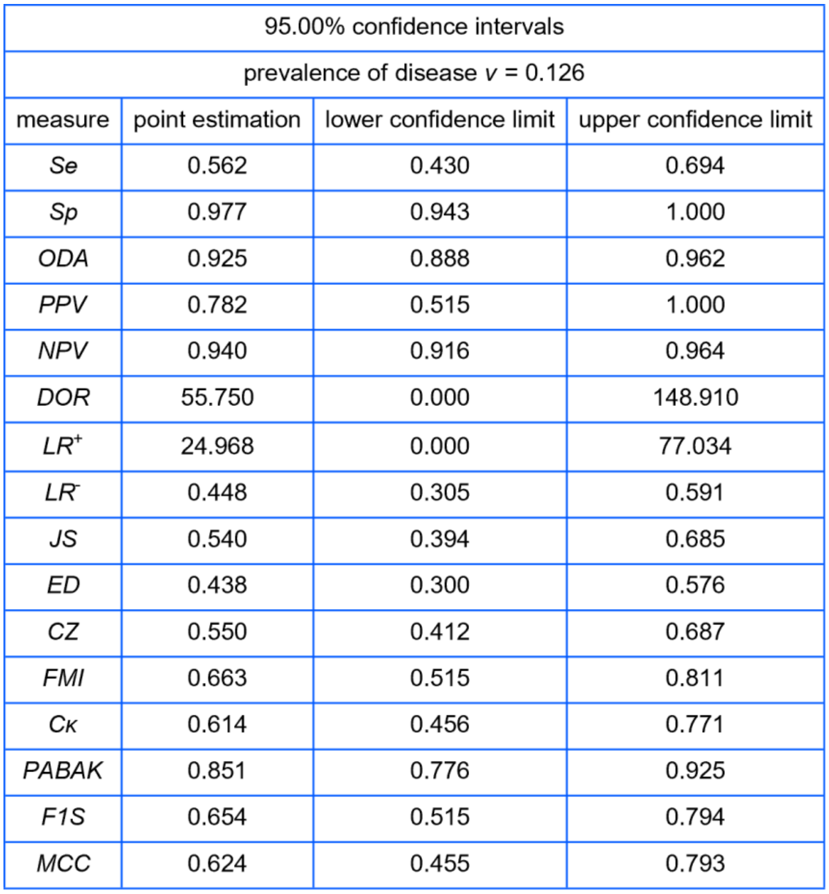


point estimation of diagnostic accuracy measures

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 22. Table of the point estimations of the diagnostic accuracy measures for an FPG value $t = 126$ mg/dL at prevalence $v \cong 0.126$, with the other settings of the program in Table 6.

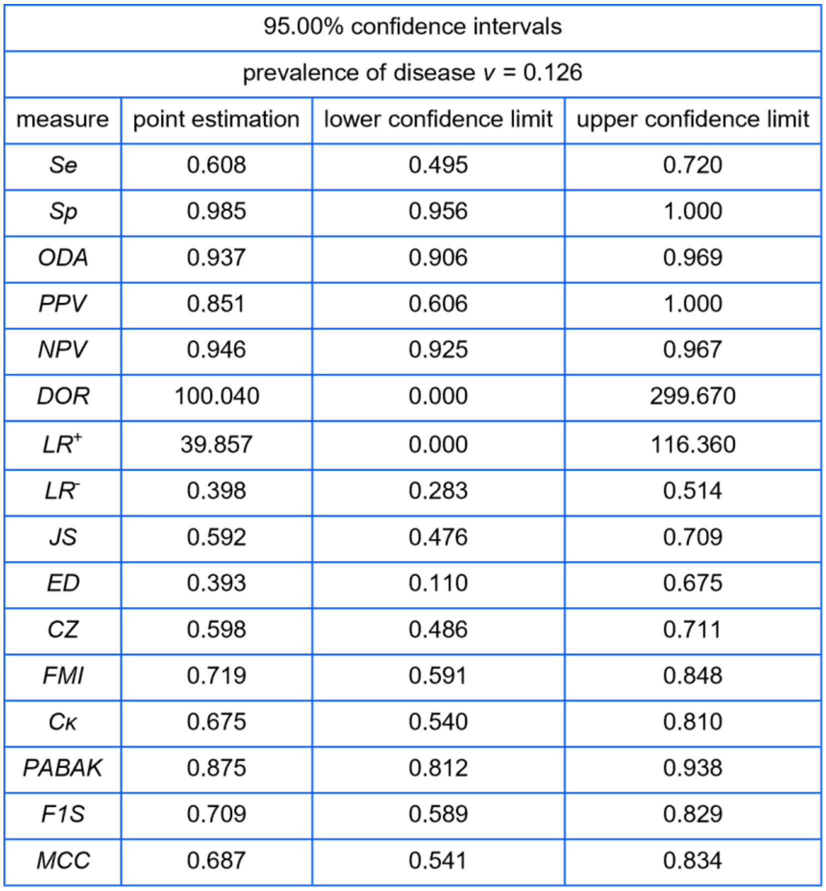


95.00% confidence intervals

prevalence of disease *v* = 0.126

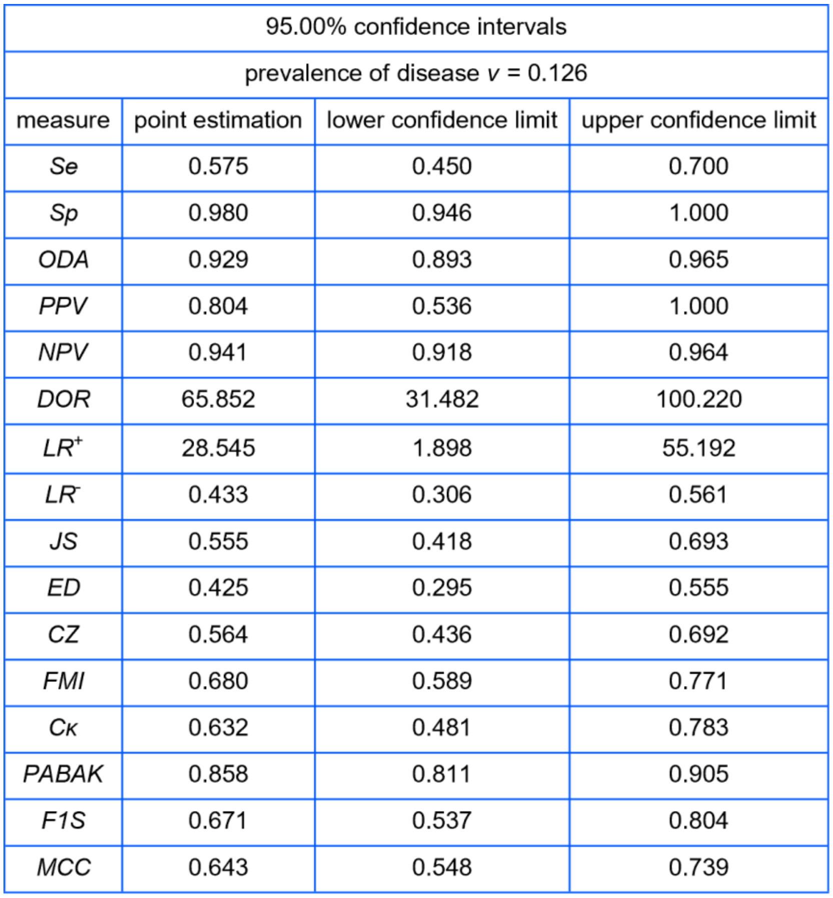
measure	point estimation	lower confidence limit	upper confidence limit
Se	0.562	0.430	0.694
Sp	0.977	0.943	1.000
ODA	0.925	0.888	0.962
PPV	0.782	0.515	1.000
NPV	0.940	0.916	0.964
DOR	55.750	0.000	148.910
LR^+	24.968	0.000	77.034
LR^-	0.448	0.305	0.591
JS	0.540	0.394	0.685
ED	0.438	0.300	0.576
CZ	0.550	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 23. Table of the point estimations and the 95% confidence intervals of the diagnostic accuracy measures for an FPG value $t = 126$ mg/dL at prevalence $v \cong 0.126$, with the other settings of the program in Table 6. Both distributions of the diseased and nondiseased are assumed lognormal.



95.00% confidence intervals			
prevalence of disease *v* = 0.126			
measure	point estimation	lower confidence limit	upper confidence limit
Se	0.608	0.495	0.720
Sp	0.985	0.956	1.000
ODA	0.937	0.906	0.969
PPV	0.851	0.606	1.000
NPV	0.946	0.925	0.967
DOR	100.040	0.000	299.670
LR^+	39.857	0.000	116.360
LR^-	0.398	0.283	0.514
JS	0.592	0.476	0.709
ED	0.393	0.110	0.675
CZ	0.598	0.486	0.711
FMI	0.719	0.591	0.848
Cκ	0.675	0.540	0.810
PABAK	0.875	0.812	0.938
F1S	0.709	0.589	0.829
MCC	0.687	0.541	0.834

Figure 24. Table of the point estimations and the 95% confidence intervals of the diagnostic accuracy measures for an FPG value $t = 126$ mg/dL at prevalence $v \cong 0.126$, with the other settings of the program in Table 6. Both distributions of the diseased and nondiseased are assumed normal.



95.00% confidence intervals			
prevalence of disease *v* = 0.126			
measure	point estimation	lower confidence limit	upper confidence limit
Se	0.575	0.450	0.700
Sp	0.980	0.946	1.000
ODA	0.929	0.893	0.965
PPV	0.804	0.536	1.000
NPV	0.941	0.918	0.964
DOR	65.852	31.482	100.220
LR^+	28.545	1.898	55.192
LR^-	0.433	0.306	0.561
JS	0.555	0.418	0.693
ED	0.425	0.295	0.555
CZ	0.564	0.436	0.692
FMI	0.680	0.589	0.771
Cκ	0.632	0.481	0.783
PABAK	0.858	0.811	0.905
F1S	0.671	0.537	0.804
MCC	0.643	0.548	0.739

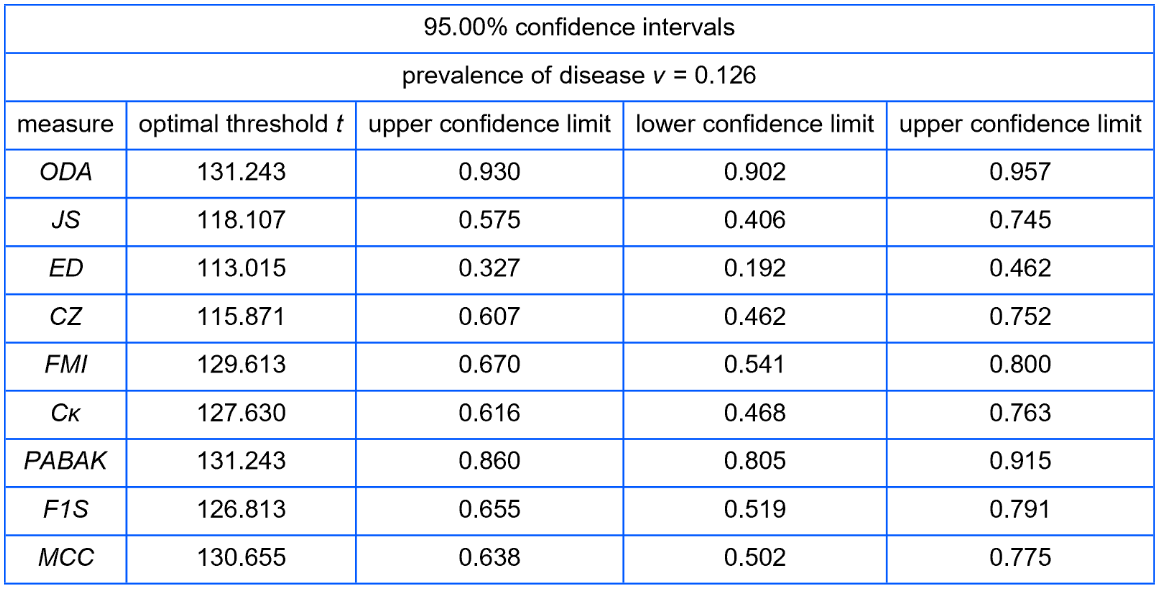
Figure 25. Table of the point estimations and the 95% confidence intervals of the diagnostic accuracy measures for an FPG value $t = 126$ mg/dL at prevalence $v \cong 0.126$, with the other settings of the program in Table 6. Both distributions of the diseased and nondiseased are assumed gamma.

3.2.7. Optimal diagnostic thresholds across measures

Figure 26 (right panel) tabulates, for each DAM, the diagnostic threshold t^* that optimizes the measure (maximize $ODA, JS, CZ, FM, F1S, C\kappa, PABAK;$ minimize ED), together with the corresponding point estimates and CIs. The optimal thresholds differ across measures, as expected from their distinct loss surfaces.

JS-optimal and ED-minimal thresholds cluster near the ROC point where the class-conditional densities cross. CZ-optimal (maximizing $Se \cdot Sp$) lies close but not identical to JS-optimal, favoring simultaneous, near-balanced gains in both Se and Sp. LR^+ and DOR-optimal thresholds are higher, leveraging extreme specificity; FMI and $F1S$ select interior diagnostic thresholds that balance recall (Se) against precision (PPV), while $C\kappa$, $PABAK$, and MCC select thresholds that best reconcile agreement or correlation with class imbalance.

Collectively, these results underscore that no single diagnostic threshold is universally optimal; selection should target the clinical objective (confirmatory diagnosis, diagnosis for exclusion [32], or balanced discrimination) and the downstream decision costs.



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 26. Table of the optimal diagnostic thresholds (in mg/dl), the respective point estimations and the 95% confidence intervals of the diagnostic accuracy measures at prevalence $v \cong 0.126$, with the other settings of the program in Table 6.

4. Discussion

4.1. Principal findings

We synthesized sixteen diagnostic accuracy measures (DAMs) within a unified, threshold-dependent framework and examined their diagnostic performance under parametric modelling of measurand distributions in diseased and non-diseased populations. Consistent with fundamental diagnostic principles, measures used for confirmatory purposes (LR^{+}, DOR) achieved their optima at higher thresholds emphasizing specificity, whereas measures used for diagnostic exclusion (LR^{-}) favored lower thresholds emphasizing sensitivity. Prevalence-dependent post-test measures (PPV, NPV) followed the anticipated monotonic responses to v, while prevalence-invariant measures ($Se, Sp, JS, ED, CZ, LR^{+}, LR^{-}, DOR, C\kappa, PABAK, MCC$) remained unaffected by v by definition. Importantly, the optimal diagnostic threshold t^{*} differed across measures, underscoring that no single diagnostic threshold can simultaneously maximize all clinically relevant goals.

4.2.Relations between diagnostic accuracy measures

Figures 12–13 illustrate two recurring principles. First, prevalence-invariant measures (JS, ED, CZ) capture the intrinsic separation between diseased and non-diseased populations, whereas prevalence-dependent measures ($ODA, PPV, NPV, F1, FMI$) vary markedly with prevalence v. Reporting both families prevents conflating diagnostic accuracy with post-test certainty. Second, ratio-type metrics (e.g., DOR) and related constructs ($C\kappa$) can be unstable near boundary values; bounded or log-scale displays are preferable for estimation and interval reporting. The contrast between ODA and JS underscores that threshold optimisation differs when guided by Bayes risk, which incorporates prevalence and misclassification costs, versus when assessed by prevalence-invariant measures of diagnostic accuracy. $ED - CZ$ and $F1S - FMI$ provide complementary perspectives on concordance: the former reflects geometric proximity, while the latter captures multiplicative or mean-based aggregation. These contrasts are particularly informative when Se and Sp improve asymmetrically. $C\kappa$ and MCC, bounded in $[-1, 1]$, serve as numerically stable anchors against which unbounded ratio measures may be interpreted. Overall, the software's relational plots act as stress tests across thresholds and prevalence settings, supporting criterion selection for diagnostic confirmation or diagnostic exclusion applications.

4.3. Appraisal of the uncertainty estimation approach

Our approach to uncertainty quantification separates sampling variability from measurement uncertainty via an explicit heteroscedastic function $u_m(t)$ and propagates both through closed-form mappings from $(n_D, m_D, s_D, n_{\bar{D}}, m_{\bar{D}}, s_{\bar{D}}, b_0, b_1)$ to each DAM. This decomposition is clinically meaningful, as it clarifies how analytical imprecision amplifies uncertainty in confirmatory, high diagnostic threshold settings (where $(1 - Sp)$ enters denominators of LR^+ and DOR), while low event counts magnify sampling error in the low-threshold region (Se-driven settings relevant to early detection and diagnosis for exclusion). The use of parametric models provides smooth, differentiable profiles over the diagnostic threshold, supports analytic first-order optimality conditions for DAMs, and enables computationally efficient interval estimation.

4.4. Potential sources of error and bias

4.4.1. 'Gold standard' assumption

In the absence of a 'gold standard' alternative approaches for classification are available [33–35].

4.4.2. Distributional misspecification.

Assuming a normal or lognormal distribution of the two populations is parsimonious but can be violated by latent mixtures [36], group-wise differences in skewness, or heavy tails. Systematic departures can bias Se and Sp at a fixed diagnostic threshold t and, by extension, any diagnostic-accuracy measure derived from them. In addition, although population-level bimodality (diseased versus non-diseased) is often assumed, unimodal or multimodal forms are empirically plausible [37–39].

Model checking (e.g., kernel-density overlays, Q–Q plots) and statistical modality tests could be applied [40].

4.4.3. Prevalence handling

For prevalence-dependent measures ($PPV, NPV, ODA, F1S, FMI$), ignoring the variance $Var(v)$ of prevalence leads to anticonservative intervals. When v is estimated, spectrum differences between the source and target populations should be considered.

4.4.4. Boundary behavior and ratio instability

Intervals for ratio measures (LR^+, LR^-, DOR) are sensitive to near-zero denominators ($Se, Sp, 1 - Sp$). Symmetric, untransformed intervals can exhibit poor coverage and non-physical bounds. Log-scale inference with back-transformation or Fieller-type methods is preferable [41, 42]; additionally, reporting one-sided lower bounds for LR^+ and upper bounds for LR^- can be more decision-relevant.

4.4.5. Selection (optimization) uncertainty

Reporting a CI for a DAM evaluated at an estimated optimum diagnostic threshold t^* implicitly conditions on a data-driven maximizer. The maximization step introduces additional variability (selection uncertainty) not captured by pointwise intervals at fixed t [43, 44]. Without adjusting for selection, uncertainty at t^* is typically understated. Bootstrap procedures that refit t^* in each replicate, or nested cross-validation for threshold selection, provide more honest diagnostic threshold uncertainty [45, 46].

4.4.6. Measurement uncertainty transferability

The heteroscedastic function $u_m(t)$ was fit to QC data; its transferability to patient samples assumes stability across matrices and time. Matrix effects or lot-to-lot shifts could understate or overstate measurement uncertainty if not periodically re-validated. Group-specific imprecision (different $u_m(t)$ for D and $\bar{D}$) and potential correlation structures also warrant scrutiny.

4.4.7. Dependence and joint estimation

Mean and standard deviation estimates of the diseased and nondiseased populations at a given threshold t, are not independent, particularly when derived from the same underlying parametric fit. Treating components as independent in interval estimation can misstate uncertainty. Joint propagation, likelihood-based intervals, or parametric bootstrap naturally respect dependence [45, 47, 48].

4.4.8. Truncation

Enforcing parametric bounds (e.g., [0,1] for probabilities, $[0, \sqrt{2}]$ for ED) after interval estimation prevents non-physical reporting but alters nominal coverage, especially near boundaries. Where feasible, constrained estimation on an appropriate link scale ($logit\ or\ log$) should precede transformation back to the natural scale.

4.4.9. Orientation misspecification.

If the true decision rule is 'smaller-is-positive' but the analysis treats 'larger-is-positive', DAM profiles invert, and optimization may drift to boundary regions with spuriously extreme ratios.

4.5. Limitations of the program

This program's limitations, which provide paths for further research, include:

4.5.1. Utilization of first-order Taylor-series approximations

First-order Taylor-series approximations are employed in the propagation of uncertainty calculations. While this method provides a baseline estimation, higher-order approximations or Monte Carlo simulations may yield more precise results [49, 50].

4.5.2. Uncertainty of terms of the measurement uncertainty equation.

The uncertainties associated with both the constant contribution b_0 and the proportionality constant b_1 in the measurement uncertainty equation are assumed to be negligible, although in practice they may be non-trivial.

4.5.3. Uncertainty approximation in disease prevalence

The uncertainty associated with the prevalence v of a disease is approximated using the Agresti–Coull adjusted Wald interval. Although this method is widely used, more accurate techniques are available [51].

4.5.4. Approximations of the sampling uncertainty for the sample means and standard deviations

These approximations can be refined for smaller sample sizes or in the presence of pronounced skewness, as observed in lognormal and gamma distributions [13, 14].

4.5.5. CIs based on the *t*-distribution

CIs are derived using the t-distribution, which for small effective degrees of freedom, may yield wide intervals but remains practical outside a Bayesian framework. [52, 53].

Although improvement of these approximations would substantially increase computational complexity, it is essential for future methodological refinement [49, 54]. We should, however, keep in mind that "all models will be based on assumptions and can only approach complex reality" [55], as "all models are wrong, but some models are useful" [56].

4.6. Limitations of the illustrative case study

The primary limitations of the illustrative case study are:

a) Selection of the OGTT as the reference diagnostic method for diabetes mellitus, even with various factors affecting glucose tolerance [57–65].

b) Approximation of the FPG measurement distributions from NHANES datasets by lognormal distributions.

c) The implicit assumption of simple random sampling.

4.7. Strengths of the framework

4.7.1. Comparability of DAMs

By expressing all DAMs as deterministic functions of parameters of the diseased and nondiseased populations $(n_D, m_D, s_D, n_{\bar{D}}, m_{\bar{D}}, s_{\bar{D}})$ the framework enables direct comparison of thresholds and facilitates principled choice of diagnostic thresholds for confirmatory versus exclusionary diagnosis.

4.7.2. Explicit incorporation of measurement uncertainty

Modelling $u_m(t)$ and propagating it through DAMs clarifies when analytical imprecision, rather than sampling error, dominates decision uncertainty—especially for ratio measures near Sp extremes.

4.7.3. Analytic structure and efficiency

Closed-form derivatives provide first-order optimality conditions for several DAMs (e.g., JS, ODA, CZ, ED), while parametric profiles yield smooth curves and enable stable optimisation without the computational burden of extensive Monte Carlo simulation.

4.7.4. Reproducibility and transparency

All computations are implemented in the Wolfram Language with documented functions, making the analysis auditable and easy to run under alternative assumptions or updated data.

4.8. Originality and positioning

To our knowledge, few applied studies in laboratory medicine present a single, integrated framework that (i) characterizes measurand distributions in diseased and non-diseased populations using alternative parametric models; (ii) propagates a heteroscedastic measurement-error model across sixteen DAMs; (iii) decomposes uncertainty into sampling and measurement components over the entire threshold continuum; and (iv) optimizes thresholds for each DAM with analytic guidance. While individual elements have precedents, their joint implementation and the breadth of DAM coverage appear uncommon in the applied diagnostics literature.

Besides, the presented software provides 386 types of plots and 66 types of comprehensive tables of the sixteen diagnostic accuracy measures, their uncertainty, and the associated CIs (Figure 1), many of which are novel. Although all major general or medical statistical software packages (JASP® ver. 0.20.0, Mathematica® ver. 14.3, Matlab® ver. R2025a, MedCalc® ver. 23.3.7, metRology ver. 0.9-29-2, NCSS® ver. 24.0.0, NIST Uncertainty Machine ver. 2.0.0, OpenBUGS ver. 3.2.3, R ver. 4.5.0, SAS® ver. 9.4, SPSS® ver. 31, CmdStan ver. 2.37.0, Stata® ver. 19, and UQLab ver. 2.1.0) include routines for calculating and plotting various diagnostic measures and their CIs, to the best of our knowledge, neither of them nor any other software offers this extensive range of plots and tables without requiring advanced statistical programming.

4.9. Practical guidance

4.9.1. Use purpose-matched diagnostic thresholds.

For confirmatory diagnosis applications, prioritize thresholds with large LR^{+} (and report robust lower bounds); for diagnosis for exclusion, favor small LR^{-} (with robust upper bounds). Balanced screening can target JS-optimal or ED-minimal thresholds, while precision-oriented tasks may prefer FMI or $F1S$ optima.

4.9.2. Mitigate selection bias in optimal diagnostic threshold t^{*}.

When reporting DAM values at an estimated optimal diagnostic threshold, supplement with cross-validated performance or bootstrap intervals that reestimate t^{*} within each replicate.

4.9.3. Stabilize ratio measures.

Work on the log-scale for LR^{+}, LR^{-}, and DOR; consider Fieller-compatible or profile-likelihood intervals; report one-sided bounds when aligned with clinical decision rules.

4.10. Limitations and future work

The main limitations are reliance on parametric distributional forms and conditional CIs at data-selected thresholds. Future work should examine semi-parametric or mixture-based models, implement full joint parametric bootstrap (including v) and cross-validated threshold selection, and validate the measurement uncertainty model across matrices and time. Extending the framework to multi-class settings, ordinal tests, or net-benefit-based decision analysis would further enhance applicability [66].

5. Conclusion

We presented an integrated, threshold-aware framework for sixteen diagnostic-accuracy measures (DAMs) that unifies point estimation, uncertainty quantification, and optimization with respect to clinically meaningful objectives. By expressing each measure as a deterministic function of $n_{\bar{D}}, m_{\bar{D}}, s_{\bar{D}}, b_0,\ and\ b_1$, and by modelling the class-conditional distributions parametrically, the approach provides smooth, interpretable profiles over the diagnostic threshold t. Measurement uncertainty is modelled via an explicit heteroscedastic model and propagated jointly with sampling variability, yielding CIs and diagnostic threshold recommendations that transparently reflect both analytical imprecision and finite-sample variability (refer to Supplemental File II: *DiagAccUCalculations.nb*).

The exploration of sampling, measurement, and combined uncertainty in DAMs may contribute to quality and risk management in laboratory medicine, affect analytical performance specifications, and guide the design and implementation of diagnostic test accuracy studies. [67, 68].

Empirically, optimal thresholds differed systematically across measures: confirmatory diagnosis indices (LR^{+}, DOR) favored higher diagnostic thresholds that emphasize specificity; diagnosis for exclusion indices (LR^{-}) favored lower diagnostic thresholds emphasizing sensitivity; and geometric- or association-based measures ($JS, ED, CZ, FMI, F1S, C\kappa, PABAK, MCC$) selected interior diagnostic thresholds that balance errors. Prevalence-dependent measures ($PPV, NPV, ODA, F1S, FMI$) behaved as expected with varying v, whereas prevalence-invariant measures remained stable. The decomposition of uncertainty clarified where measurement error dominates (e.g., high t regions where $(1 - Sp)$ appears in denominators) versus where sampling error dominates (e.g., low t regions with few true positives).

Methodologically, the framework's strengths lie in (a) its unified taxonomy across heterogeneous measures; (b) explicit treatment of measurement uncertainty alongside sampling variability; (c) analytic structure supporting first-order optimality conditions for the sixteen DAMs; and (d) reproducibility via a transparent Wolfram Language implementation. These features enable principled selection of confirmatory diagnosis versus diagnosis for exclusion thresholds, with clear trade-offs and uncertainty statements aligned to clinical decision-making.

Nonetheless, the validity of inferences remains contingent on distributional adequacy, correct orientation of the decision rule, and appropriate handling of prevalence. Ratio measures require care near boundaries; selection uncertainty at the estimated optimal diagnostic threshold t^{*} should be acknowledged and, where feasible, addressed through bootstrap or cross-validation that reestimates t^{*} in each replicate. Periodic verification of the measurement uncertainty model against quality control data is advisable to ensure transportability across time, matrices, and lots.

In summary, this work offers a coherent and practically useful route from fitted class-conditional models to an uncertainty-aware comparison of multiple diagnostic objectives. Its explicit articulation of the roles of prevalence, measurement uncertainty, and threshold optimization provides a basis for more transparent reporting and more reliable clinical decisions. Future extensions to semiparametric or mixture models, multivariate markers, and decision-curve or net-benefit analyses would further broaden the framework's applicability.

6. Supplemental material

The following supplemental files are available for download as a ZIP archive at: https://www.hcsl.com/Supplements/SDAMU.zip:

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.

7. Declarations

Funding: This research received no external funding.

Institutional Review Board Statement: Data collection was carried out 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 21, 2025).

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

Data Availability Statement: The data presented in this study are available at https://wwwn.cdc.gov/nchs/nhanes/default.aspx (accessed on September 21, 2024).

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

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

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

8. 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.

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

10. 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.

11. 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.

12. 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.

13. 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.

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

15. Chatzimichail T, Hatjimihail AT. A software tool for applying Bayes' theorem in medical diagnostics. BMC Med Inform Decis Mak. 2024;24:399.

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

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

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

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

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

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

22. Chatzimichail T, Hatjimihail AT. A Software Tool for Estimating Uncertainty of Bayesian Posterior Probability for Disease. Diagnostics (Basel). 2024;14. https://doi.org/10.3390/diagnostics14040402.

23. 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.

24. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract. 2022;183:109119.

25. 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.

26. National Center for Health Statistics. National Health and Nutrition Examination Survey Questionnaire. Centers for Disease Control and Prevention. 2005-2016. https://wwwn.cdc.gov/nchs/nhanes/Search/variablelist.aspx?Component=Questionnaire. Accessed 4 Sept 2023.

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

28. Myung IJ. Tutorial on maximum likelihood estimation. J Math Psychol. 2003;47:90–100.

29. Johnson ML. Nonlinear least-squares fitting methods. In: Methods Cell Biol. Academic Press; 2008. p. 781–805.

30. Bates DM, Watts DG. Nonlinear Regression Analysis and Its Applications. Hoboken, New Jersey: John Wiley & Sons, Inc.; 1988.

31. Darling DA. The Kolmogorov-Smirnov, Cramer-von Mises Tests. Ann Math Stat. 1957;28:823–38.

32. Fred HL. The diagnosis of exclusion: an ongoing uncertainty. Tex Heart Inst J. 2013;40:379–81.

33. Knottnerus JA, Dinant GJ. Medicine based evidence, a prerequisite for evidence based medicine. BMJ. 1997;315:1109–10.

34. Pfeiffer RM, Castle PE. With or without a gold standard. Epidemiology . 2005;16:595–7.

35. van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KGM, de Groot JAH. Latent class models in diagnostic studies when there is no reference standard--a systematic review. Am J Epidemiol. 2014;179:423–31.

36. 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.

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

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

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

40. Gramacki A. Nonparametric Kernel Density Estimation and Its Computational Aspects. Springer; 2017.

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

46. Parvandeh S, Yeh H-W, Paulus MP, McKinney BA. Consensus features nested cross-validation. Bioinformatics. 2020;36:3093–8.

47. 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.

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

49. Joint Committee for Guides in Metrology. Evaluation of measurement data — Supplement 1 to the “Guide to the expression of uncertainty in measurement”— Propagation of distributions using a

Monte Carlo method Joint Committee for Guides in Metrology. Pavillon de Breteuil, F-92312 Sèvres, Cedex, France: BIPM; 2008.

50. Joint Committee for Guides in Metrology. Evaluation of measurement data – Supplement 2 to the "Guide to the expression of uncertainty in measurement" – Extension to any number of output quantities. Pavillon de Breteuil, F-92312 Sèvres, Cedex, France: BIPM; 2011.

51. Pires AM, Amado C. Interval Estimators for a Binomial Proportion: Comparison of Twenty Methods. Revstat Stat J. 2008;6:165–97.

52. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. CRC Press; 2013.

53. Stephens M. The Bayesian lens and Bayesian blinkers. Philos Trans A Math Phys Eng Sci. 2023;381:20220144.

54. Joint Committee for Guides in Metrology. Guide to the expression of uncertainty in measurement — Part 6: Developing and using measurement models. Pavillon de Breteuil, F-92312 Sèvres, Cedex, France: BIPM; 2020.

55. Oosterhuis WP. Analytical performance specifications in clinical chemistry: the holy grail? J Lab Precis Med. 2017;2:78–78.

56. Box GEP. Robustness in the strategy of scientific model building. In: Robustness in Statistics. Elsevier; 1979. p. 201–36.

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

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

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

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

61. 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.

62. 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.

63. 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.

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

65. 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.

66. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18.

67. Horvath AR, Bell KJL, Ceriotti F, Jones GRD, Loh TP, Lord S, et al. Outcome-based analytical performance specifications: current status and future challenges. Clin Chem Lab Med. 2024. https://doi.org/10.1515/cclm-2024-0125.

68. Sitch AJ, Dekkers OM, Scholefield BR, Takwoingi Y. Introduction to diagnostic test accuracy studies. Eur J Endocrinol. 2021;184:E5–9.

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

$\overline{T}$: 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 : concordance probability

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 ith 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 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 classification

A.3.1 Fundamental measures

Se, Sp, PPV, NPV, ODA.

A.3.2. Composite indices

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

A.3.3. Extended indices

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

A.3.4. Distance & concordance measures

$ED, CZ, F1S$.

A.3.5. Multi-axis classification

The taxonomy below classifies the 16 measures along multiple orthogonal axes.

Table A.1. Core classification axes (prevalence dependence, conditionality, and conceptual family).

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)	✓	✓	—	✓	—	—
Concordance Probability (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.3.5. Additional orthogonal axes classifications

A.3.5.1 Scale, bounds, and reference points:

a. [0,1], larger is better: $Se, Sp, ODA, PPV, NPV, CZ, FM, F1$.
b. [-1,1], centered at 0: $C\kappa, PABAK, MCC$.
c. [0,∞), evidence ratios (null = 1): LR^+, LR^-, DOR.
d. [0,√2], smaller is better: ED.

A.3.5.2. Symmetry under label swapping ($D \leftrightarrow \bar{D}, T \leftrightarrow \bar{T}$):

a. Symmetric: $J, ED, CZ, \kappa, PABAK, MCC, DOR$ ($LR^+ \leftrightarrow LR^-$ swap roles).
b. Asymmetric: $Se, Sp, PPV, NPV LR^+, LR^-, F1, FM, ODA$.

A.3.5.3. Robustness to class imbalance (qualitative):

a. More robust: $CZ, JS, ED, LR^+, LR^-, DOR$; MCC is also robust but defined via all four margins.
b. Sensitive to imbalance/prevalence: $ODA, PPV, NPV, F1, FM, C\kappa, PABAK$.

A.3.5.4. Orientation to clinical use:

a. Emphasis on confirmatory diagnosis: LR^+, PPV, DOR (via LR^+).
b. Emphasis on diagnosis for exclusion: LR^-, NPV.
c. Balanced/overall discrimination/agreement: $Se, Sp, JS, ED, CZ, \kappa, PABAK, MCC, ODA, FM, F1S$.

A.3.5.5. Estimation and transportability across designs:

a. Transportable across prevalence: $Se, Sp, LR^+, LR^-, DOR, JS, ED, CZ, MCC, C\kappa, PABAK$.
b. Require target-population prevalence:

$ODA, PPV, NPV, F1S, FM$.

A.3.5.6. Boundary behavior and singularities

a. CZ is well-defined on $[0,1]^2$; no singularities.
b. LR^+ diverges as $Sp \rightarrow 1$;
c. LR^- diverges as $Se \rightarrow 0$;
d. DOR diverges as $Sp \rightarrow 1$;
e. DOR and MCC are undefined under zero margins.

A.3.5.7. ROC-geometry semantics

a. $CZ = Se \cdot Sp$ corresponds to rectangular-hyperbola level sets in ROC space; it rewards simultaneous vertical (Se) and horizontal (Sp) performance.
b. JS measures maximum vertical displacement; ED measures Euclidean distance to $(0,1)$; $LR^+$and LR^- relate to iso-likelihood slopes; DOR is an odds association.

A.4. Input

A.4.1. Range of input parameters

t: $maximum(0, minimum(m_{\bar{D}} - 6s_{\bar{D}}, m_D - 6s_{\bar{D}})) - maximum(m_{\bar{D}} + 6s_{\bar{D}}, m_D + 6s_{\bar{D}})$

n_D : 2 – 10,000

m_D: 0.1 – 10,000

s_D: 0.01 – 1,000

$n_{\bar{D}}$: 2 – 10,000

$m_{\bar{D}}$: 0.1 – 10,000

$s_{\bar{D}}$: 0.01 – 1,000

v : 0.001 – 0.999

n_U : 20 – 10,000

b_0 : 0 – $\sigma_{\bar{D}}$

b_1: 0 – 0.1000

p : 0.900 – 0.999

$t, m_D, s_D, m_{\bar{D}}$, and $s_{\bar{D}}$ are defined in arbitrary units.

A.4.2. Additional input options

A.4.2.1. Plots

Users can select between an extended and limited plot range.

A.4.2.2. Tables

Users can define the number of decimal digits for results, ranging from 1 to 10.

A.4.3. About program controls

The program features an intuitive tabbed user interface to streamline user interaction and facilitate effortless navigation across multiple modules and submodules.

Users may define the numerical settings with menus or sliders. Sliders can be finely manipulated by pressing the *alt* or *opt* key while dragging the mouse. The sliders can be even more finely manipulated by also holding the *shift* or *ctrl* keys.

Dragging with the mouse while pressing the *ctrl*, *alt*, or *opt* keys zooms plots in or out. When the mouse cursor is positioned over a point on a curve in a plot, the coordinates of that point are displayed, and vertical drop lines are drawn to the respective axes.

A.7. Software availability and requirements

Program name: DiagAccU

Version: 1.0.0

Project home page: https://www.hcsl.com/Tools/DiagnosticAccuracy/ (accessed on September 12, 2025)

Program source: *DiagAccU.nb*

Available to download as a ZIP archive at:
https://www.hcsl.com/Tools/DiagnosticAccuracy/DiagAccU.zip (accessed on September 21, 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 21, 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

9. 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

10. License

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

First Published: September 21, 2025