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*

Background: Diagnostic accuracy measures (DAMs) are widely used in evaluating medical 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 measurements 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 and their uncertainties and confidence intervals (CIs).

Results: We used fasting plasma glucose for diabetes diagnosis as an illustrative case study. At extreme thresholds, ratio-type measures showed widened CIs. Agreement, association, and concordance based indices exhibited comparatively stable behaviour and typically attained interior optima. Confirmatory objectives favoured higher thresholds emphasising specificity, whereas exclusionary objectives favoured lower thresholds emphasising sensitivity. These findings support reporting threshold-wise confidence intervals and aligning cut-off point selection with the intended clinical objective.

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 and diagnosis for exclusion.

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 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.

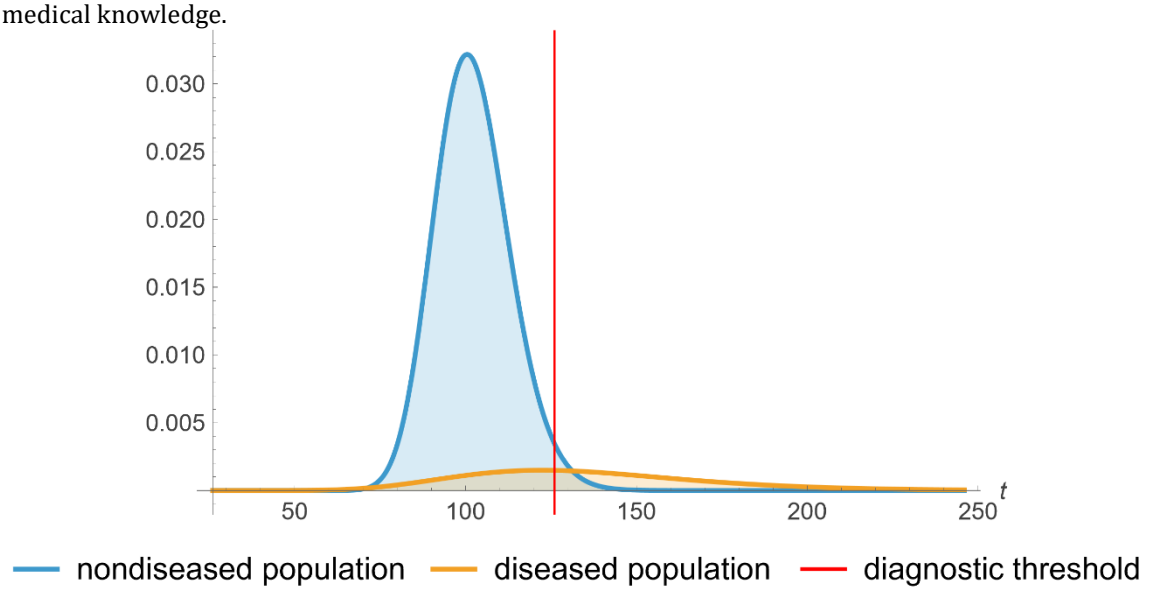

Figure 1. Probability density functions plots of fasting plasma glucose (FPG) in a diabetic (diseased) and nondiabetic (nondiseased) population.

Many in vitro screening and diagnostic tests are used as binary classifiers to partition individuals into mutually exclusive diseased (D) and nondiseased ($\overline{D}$) 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 $\overline{D}$ typically overlap. A diagnostic threshold t dichotomizes results: values above t indicate a positive test result (T) and values below indicate a negative result ($\overline{T}$) (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	
		diseased (D)	nondiseased ($\overline{D}$)
test	positive (T)	true positive (TP)	false positive (FP)
	negative ($\overline{T}$)	false negative (FN)	true negative (TN)

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),

Table 2: Definitions of the sixteen DAMs

Sensitivity (Se): Probability that the test is positive among diseased individuals. Higher values reduce false negatives and support diagnosis for exclusion; range 0–1.
Specificity (Sp): Probability that the test is negative among nondiseased individuals. Higher values reduce false positives and support confirmatory diagnosis; range 0–1.
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; optimised 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.

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

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).

Within this framework, *confirmatory diagnosis* denotes applying a higher decision threshold that prioritizes specificity thereby reducing the frequency of false-positive results among nondiseased individuals. In contrast, *diagnosis for exclusion* denotes applying a lower decision threshold that prioritizes sensitivity, thereby reducing the frequency of false-negative results among diseased

individuals. These complementary strategies capture the dual clinical roles of diagnostic thresholds in laboratory medicine and provide context for interpreting DAMs in practice.

1.2.Uncertainty

As there is inherent variability in any measurement process, there is measurement uncertainty, which is defined as a "parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand" [7]. The parameter may be the standard measurement uncertainty (u), expressed as a standard deviation and estimated as described in "Expression of Measurement Uncertainty in Laboratory Medicine [8]. Bias may be considered as a component of the standard measurement uncertainty [9].

Measurement uncertainty is gradually replacing the total analytical error concept [10].

From a metrological perspective, DAM uncertainty describes the dispersion of values that could reasonably be attributed to a DAM given available evidence [11]. Two primary components contribute to the combined uncertainty: measurement uncertainty and sampling uncertainty [8, 12, 13]. Both components can materially affect DAMs [14–17]. Uncertainty can be propagated from input parameters — sample means m_P, standard deviations s_P, prevalence v, and measurement uncertainty u_m — to DAM values using first-order Taylor-series expansion. Both linear and nonlinear heteroscedastic measurement uncertainty models can also be incorporated to better represent analytical performance [7]

1.3.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 [18, 19]. Modelling the measurand as normal, lognormal, or gamma allows uncertainty estimation under symmetric, positively skewed, or more flexible assumptions, respectively [20].

1.1.Previous research

A substantial literature addresses diagnostic accuracy estimation and ROC modelling, and a separate literature addresses measurement uncertainty in laboratory reporting and analytical performance specifications. Nevertheless, comparatively fewer studies integrate these two domains within a single uncertainty-propagation framework that is (a) threshold-wise across multiple DAMs, (b) compatible with non-normal measurand distributions, and (c) able to represent heteroscedastic analytical imprecision [21–23].

Our prior software tools addressed complementary elements of this integration: (i) exploration of relationships between diagnostic accuracy and measurement uncertainty, and (ii) estimation of uncertainty for diagnostic measures within earlier modelling constraints [17, 24, 25]. In parallel, the statistics literature includes foundational work on parametric ROC modelling for non-normal data and on measurement error effects in ROC analyses and cut-point [14–16, 20].

The present work extends this landscape by offering a unified, implementable approach for uncertainty propagation and component decomposition across a broad set of DAMs under practical distributional and measurement-uncertainty models relevant to laboratory medicine, explicitly aligned with metrological guidance for uncertainty evaluation (BIPM et al. 2008; Kallner et al. 2012).

2. Methods

2.1.Overview

We developed a computational framework that extends our prior work on uncertainty estimation for threshold-based DAMs [17, 25] in four specific ways:

2.1.1. Measurand distributional modelling

Support for three common parametric models for biomarker distributions in diseased and non-diseased populations—normal, lognormal, and gamma—to reflect frequent departures from normality in laboratory measurements.

2.1.2. Measurement-uncertainty modelling

Support for homoscedastic as well as linear and non-linear heteroscedastic measurement uncertainty functions, enabling concentration-dependent analytical imprecision consistent with routine laboratory performance patterns.

2.1.3. Threshold-continuum uncertainty profiling and decomposition

Estimation of measurement, sampling, and combined uncertainty components (and corresponding confidence intervals) over the full threshold continuum of 16 DAMs selected for completeness and implementation, enabling identification of threshold regions where uncertainty is dominated by analytical versus sampling sources.

2.1.4. Reproducible software implementation

Implementation of the framework in the program *DiagAccU*, providing an end-to-end workflow from distribution fitting and uncertainty-model specification to uncertainty propagation, decomposition, and reporting outputs.

Uncertainty propagation is performed using first-order Taylor-series propagation consistent with the Guide to the Expression of Uncertainty in Measurement (GUM) and related laboratory-medicine guidance, with confidence intervals constructed using appropriate effective degrees-of-freedom approximations where applicable [7, 8].

2.2.Calculations

2.2.1. Diagnostic accuracy measures

Following the definition of the sensitivity and specificity of a test (Table 3), the respective functions versus diagnostic threshold t are:

$$Se(t, \mu_D, \sigma_D) = F(t, \mu_D, \sigma_D)$$

$$Se(t, \mu_{\bar{D}}, \sigma_{\bar{D}}) = 1 - F(t, \mu_{\bar{D}}, \sigma_{\bar{D}})$$

The rest of the DAMS, defined in Tables 3-4, are calculated accordingly (refer to Supplemental File II: *DiagAccUCalculations.nb*).

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	$\frac{TN}{TN + FN}$	$P(\overline{D}\|\overline{T})$	$\frac{Sp\ (1-v)}{Sp\ (1-v) + (1-Se)v}$
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	$\frac{TN\ TP}{FN\ FP}$	$\frac{\frac{P(T\|D)}{P(\overline{T}\|D)}}{\frac{P(T\|\overline{D})}{P(\overline{T}\|\overline{D})}}$	$\frac{\frac{Se}{1-Se}}{\frac{1-Sp}{Sp}}$
$\boldsymbol{LR^+}$	$\frac{TP\ (FP + TN)}{FP\ (FN + TP)}$	$\frac{P(T\|D)}{P(T\|\overline{D})}$	$\frac{Se}{1-Sp}$
$\boldsymbol{LR^-}$	$\frac{FN\ (FP + TN)}{TN\ (FN + TP)}$	$\frac{P(\overline{T}\|D)}{P(\overline{T}\|\overline{D})}$	$\frac{1-Se}{Sp}$
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	$\frac{TN\ \ TP}{(TN + FP)(FN + TP)}$	$P(T\|D)\ P(\overline{T}\|\overline{D})$	$Se\ Sp$

Table 4: Mathematical definitions of diagnostic accuracy measures

measure	natural frequency definition	probability definition	definition versus $Sp, Se, and\ v$
FMI	$\frac{TP}{\sqrt{(FN+TP)(FP+TP)}}$	$\sqrt{P(T\|D)P(D\|T)}$	$\sqrt{\frac{Se^2 v}{(1-Sp)(1-v)+Se\ v}}$
Cκ	$\frac{\frac{TP+TN}{(TP+FP+TN+FN)} - \frac{(TP+FP)(TP+FN)+(FN+TN)(FP+TN)}{(TP+FP+TN+FN)^2}}{1-\frac{(TP+FP)(TP+FN)+(FN+TN)(FP+TN)}{(TP+FP+TN+FN)^2}}$	$\frac{2(1-P(T\|D)-P(\bar{T}\|\bar{D}))P(\bar{D})P(D)}{-1+P(\bar{T}\|\bar{D})-(-2+3P(\bar{T}\|\bar{D})+P(T\|D))P(D)+2(-1+P(T\|D)+P(\bar{T}\|\bar{D}))P(D)^2}$	$\frac{2(-1+Se+Sp)(-1+v)v}{-1+Sp-(-2+Se+3Sp)v+2(-1+Se+Sp)v^2}$
PABAK	$\frac{TP+TN-FP-FN}{TP+FP+TN+FN}$	$2\big(P(D\|T)P(T)+P(\bar{T}\|\bar{D})P(\bar{T})\big)-1$	$-1+2(Sp(1-v)+Se\ v)$
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	$\frac{TP\ TN-FP\ FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	$\frac{P(D\|T)-P(D)}{\sqrt{P(D)\big(1-P(D)\big)P(T)\big(1-P(T)\big)}}$	$\frac{(1-Se-Sp)(1-v)v}{\sqrt{(1-v)v\big((1-Sp)(1-v)+v\big)(1-v+(1-Se)v)}}$

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

2.2.2. Measurements distributions

For each of the diseased (D) and nondiseased ($\bar{D}$) populations, the measurements 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 ensure consistent uncertainty propagation under GUM.

2.2.3. Optimisation

Optimisation is performed in the mathematical sense, by identifying a local or global extremum of the diagnostic accuracy measures (DAMs) as a function of the diagnostic threshold. Such extrema should not be interpreted as operationally optimal decision thresholds for patient management, because operational optimality typically depends on additional decision-analytic and health-economic considerations.

For selected DAMs, first-order optimality conditions are derived analytically to characterize candidate extrema.

2.2.4. Measurement uncertainty

Measurement uncertainty $u_m(t)$ is modelled using linear and nonlinear functions, consistent with analytical measurement guidelines [7]:

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

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

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, for $b_1 > 0$) uncertainty profiles commonly encountered in clinical assays.

2.2.5. Sampling uncertainty

Sampling uncertainty for the sample mean m_P and sample standard deviation s_P of a population P with sample size n_P is computed as:

$$u_s(m_P) \cong \frac{s_P}{\sqrt{n_P}}$$

$$u_s(s_P) \cong \frac{s_P}{\sqrt{2(n_P - 1)}}$$

using the central limit theorem and the chi-square distribution (26–28).

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

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

using the Agresti–Coull adjusted Wald interval, which offers improved coverage at small n and extreme probabilities (29).

2.2.6. Combined uncertainty

When there are l uncorrelated components of uncertainty, each with standard uncertainty $u_i(t)$, then the combined standard uncertainty is approximated as:

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

2.2.7. Combined uncertainty propagation

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

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

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 $g(t|\boldsymbol{\theta})$.

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 [30, 31]:

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

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 DAMs, their uncertainties and CIs are truncated to their ranges (refer to Table 2).

2.3. Software implementation

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.4). Owing to the complexity of the underlying calculations, the notebook is extensive and may require substantial computational resources.

Detailed documentation of its interface is available in Supplemental File III: *DiagAccUInterface.pdf*.

3. Results

3.1.Illustrative case study

As previously described, we undertook an illustrative case study to demonstrate the program's application [32]. 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 [33]. The study focused on individuals aged 65 to 68 years, reflecting the significant correlation between age and diabetes prevalence [34].

Data were obtained from participants in the National Health and Nutrition Examination Survey (NHANES) from 2005 to 2016 (n = 60,936), as described previously [32]. NHANES is a comprehensive survey assessing the health and nutritional status of adults and children in the United States [35].

The inclusion criteria were valid FPG and OGTT results (n = 13,836), no prior diagnosis of diabetes [36] (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) [33].

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 [37], as follows:

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

Table 5 presents the summary 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 measurements in diabetic and nondiabetic participants using the maximum likelihood estimation method [38]. Parametrised 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 [39, 40]) 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 measurements, assuming negligible measurement uncertainty, were estimated as:

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

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

Table 6. The settings of the program *DiagAccU* for the Figures 3-11

	Units	Fig 3	Fig 4	Fig 5	Fig 6	Fig 7	Fig 8	Fig 9	Fig 10	Fig 11
t	mg/dL	91.0– 173.0	126		91.0– 173.0	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	136.0
σ_D	mg/dL	36.6	17.7	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	102.2
$\sigma_{\overline{D}}$	mg/dL	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
b_0		-	-	-	0.8124	0.8124	0.8124	-	0.8124	0.8124
b_1		-	-	-	0.0119	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

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 [41]. Figure 2 presents 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.

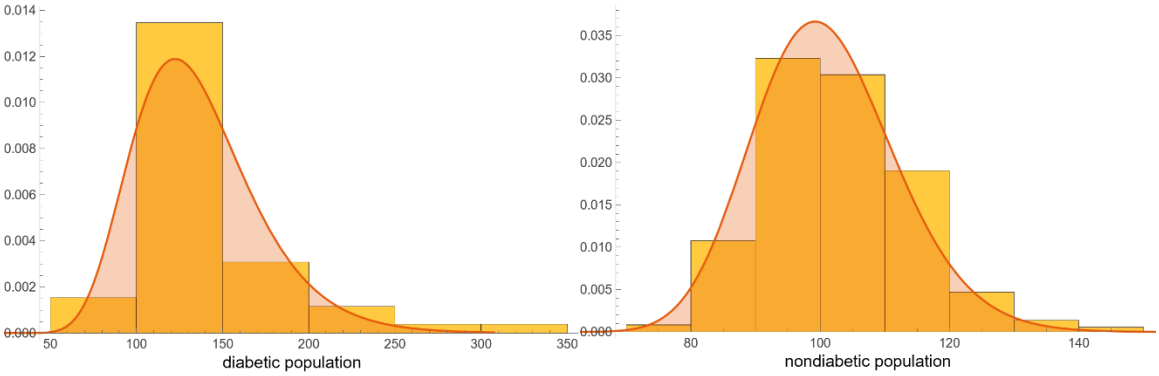

Figure 2. The estimated PDFs of the FPG (mg/dL) in diabetic and nondiabetic participants, assuming negligible measurement uncertainty.

Figure 1 shows both PDFs and the ADA diagnostic threshold $t = 126$ mg/dl for FPG for diabetes [33].

The DAMs, and their uncertainties and CIs were estimated accordingly.

1.1.Application of the program

The results of the application of the program on the illustrative case study dataset are presented in Figures 3-11. Unless otherwise noted, all figures use the settings in Table 6. The selected diagnostic threshold $t = 126$ mg/dl of Figures 4 and 8-11 is the ADA diagnostic threshold of FPG for diabetes (refer to Figure 1).

1.2.Threshold-dependent profiles

Across operating diagnostic thresholds Se decreased monotonically with diagnostic threshold t, while Sp increased to a high plateau (refer to Figure 3). 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 showed 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.

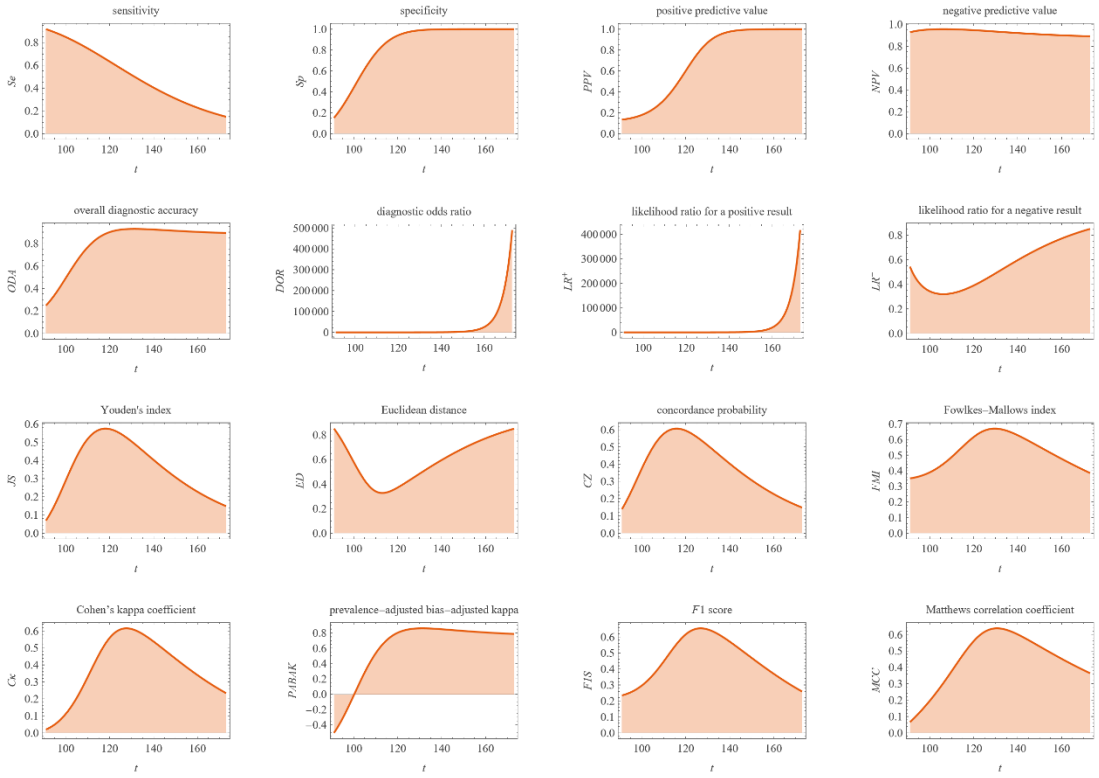

Figure 3. The 16 DAMs versus threshold t, with the settings of the program in Table 6.

1.3.Prevalence-dependent profiles

At a fixed diagnostic threshold, predictive values were strongly prevalence-dependent (refer to Figure 4): PPV increased steeply at low ν and approached an upper plateau, whereas NPV decreased monotonically toward zero. ODA declined approximately linearly with increasing ν, reflecting increasing class imbalance. FMI and $F1S$ rose steeply at low ν and then levelled. In contrast, $C\kappa$ and the MCC were unimodal, each attaining an interior maximum at moderate ν and decreasing at high ν. $PABAK$ decreased roughly linearly across the range of ν.

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

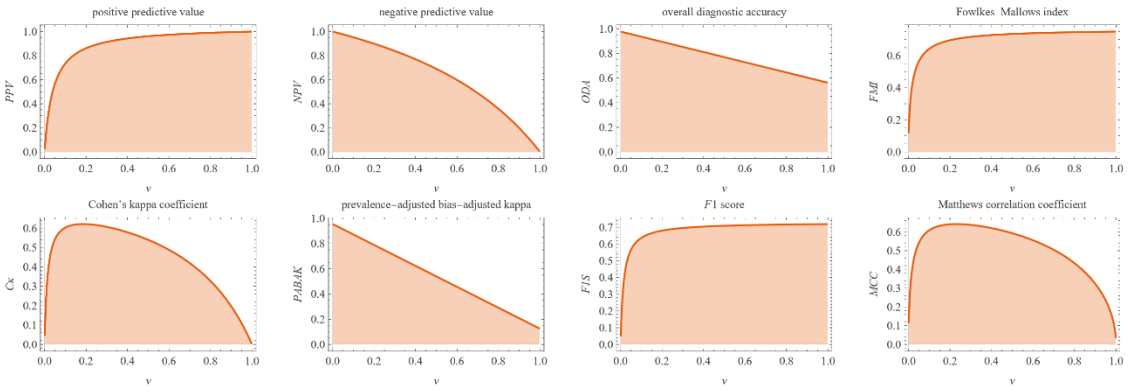

Figure 4. Positive predictive value (PPV), negative predictive value (NPV), overall diagnostic accuracy (ODA), Fowlkes–Mallows index (FMI), Cohen's kappa coefficient ($C\kappa$), prevalence-adjusted bias-adjusted kappa ($PABAK$), $F1$ score ($F1S$), and Matthews correlation coefficient (MCC) versus prevalence of diabetes v, with the settings of the program in Table 6.

1.4.DAM relations

Pairwise relations between measures (refer to Figure 5) 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 $C\kappa$ 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.

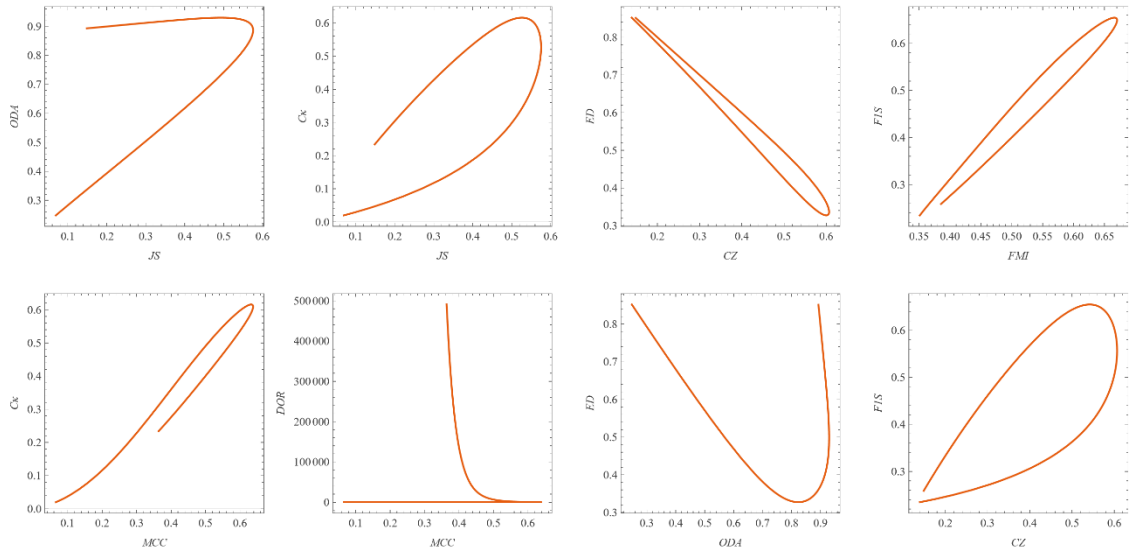

Figure 5. Plots of overall diagnostic accuracy (ODA) and Cohen's kappa coefficient ($C\kappa$) versus Youden's index (JS), Euclidean distance (ED) versus concordance probability (CZ), $F1$ score ($F1S$) versus Fowlkes–Mallows index (FMI), Cohen's kappa coefficient ($C\kappa$) and diagnostic odds ratio (DOR) versus Matthews correlation coefficient (MCC), Euclidean distance (ED) versus overall diagnostic accuracy (ODA), and $F1$ score ($F1S$) versus concordance probability (CZ), with the settings of the program in Table 6.

1.5.Uncertainty

The standard combined uncertainty $u_c(t)$ exhibited three patterns (refer to Figure 6):

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 for LR^+ and DOR). For PPV, $u_m(t)$ displayed a distinct interior maximum.

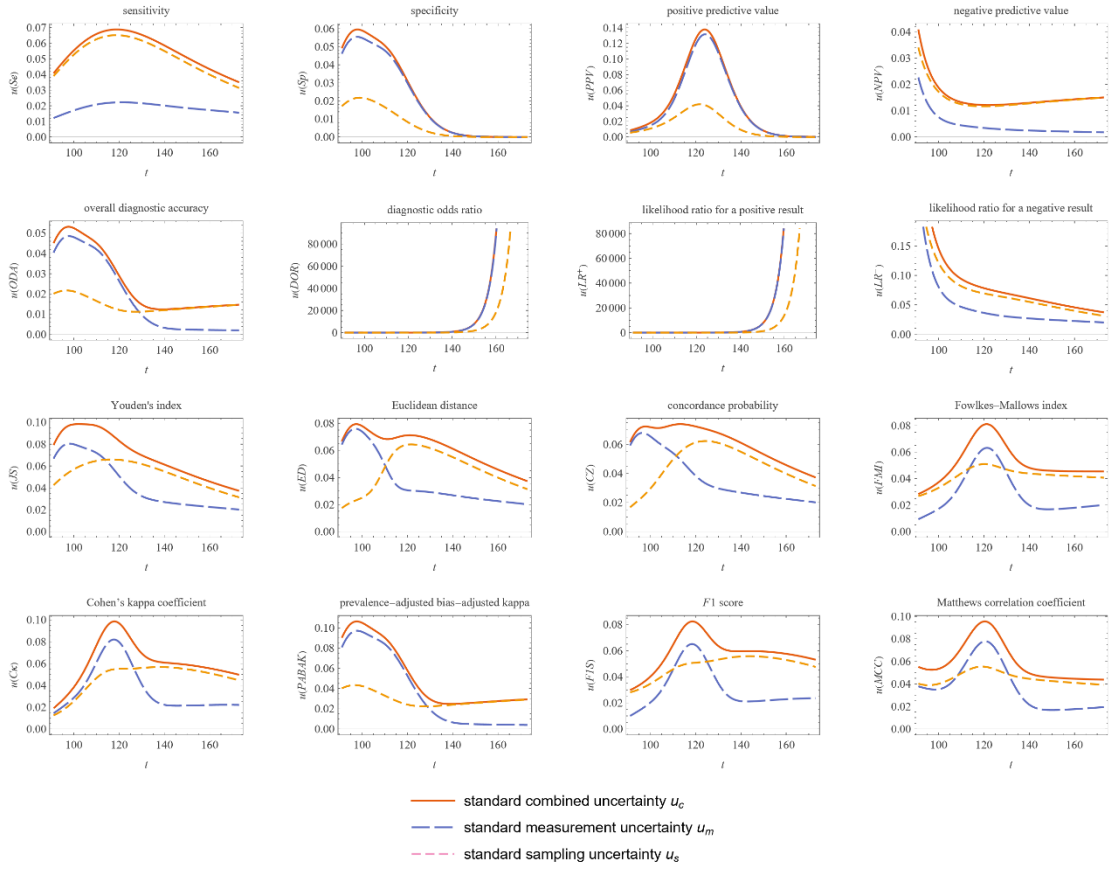

Figure 6. Standard sampling, measurement, and combined uncertainty of the sixteen DAMs versus threshold t, with the settings of the program in Table 6.

1.6.Relative uncertainty

The relative standard combined uncertainty (refer to Figure 7) 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 relative measurement uncertainty 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. ED 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).

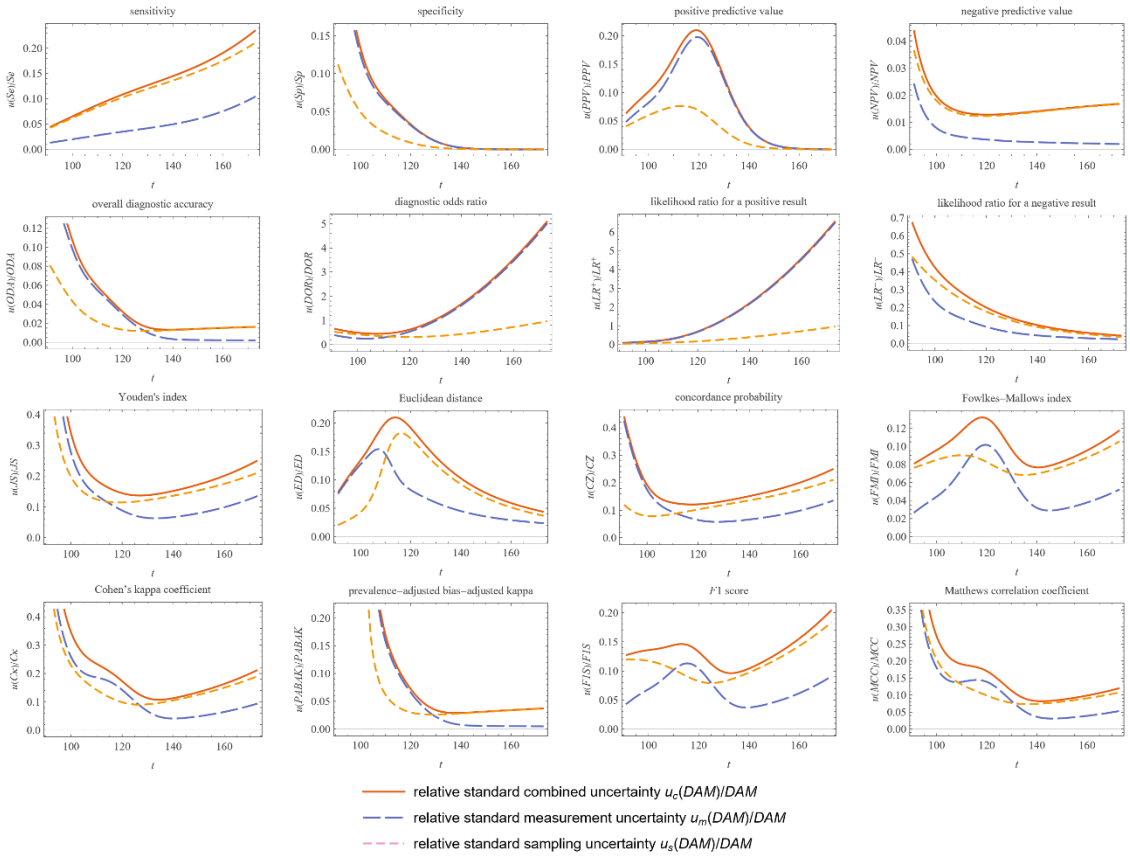

Figure 7. Relative standard sampling, measurement, and combined uncertainty of the sixteen DAMs versus threshold t, with the settings of the program in Table 6.

1.7.CIs

Figure 8 presents 95% CIs across operating diagnostic thresholds. 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).

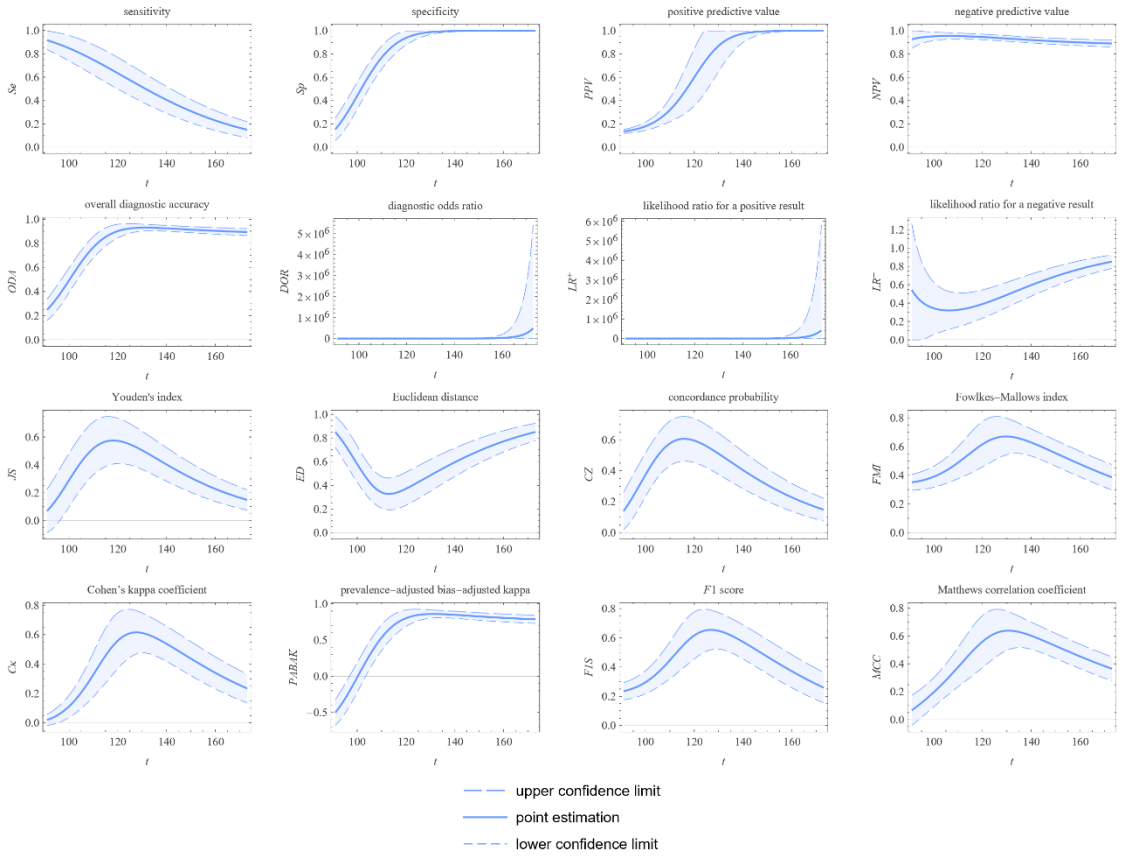

Figure 8. CIs of the 16 DAMs versus threshold t, with the settings of the program in Table 6.

1.8.Point estimates and CIs at a clinical threshold

Point estimates and 95% CIs for all 16 DAMs are shown in Figures 9-11, at the ADA screening cut-off of 126 mg/dL. *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 9. 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 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 10. 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 6.

1.9.Optimal diagnostic thresholds

Optimal thresholds for each DAM are summarised in Figure 11. 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 diagnostic threshold is universally optimal; selection should target the clinical objective (confirmatory diagnosis, diagnosis for exclusion [42], or balanced discrimination) and the downstream decision costs of false-positive and false-negative results.

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 11. 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 6.

2. Discussion

2.1.Principal findings

We synthesised sixteen DAMs within a unified, threshold-dependent framework and examined their behaviour under parametric modelling of measurements distributions in diseased and nondiseased populations. Prevalence-dependent measures ($PPV, NPV, ODA, F1S, FMI$) followed the anticipated monotonic responses to disease prevalence v. Across measures, uncertainty was not uniform over the decision axis. For most non-ratio type indices (e.g. $JS, CZ, C\kappa, PABAK, F1S, MCC$) combined uncertainty tended to be minimised at interior thresholds, consistent with their interior optima. In contrast, at extreme thresholds, ratio-type measures showed widened CIs. Decomposition of the combined uncertainty indicated that sampling variability predominated at low thresholds—where diseased positives were fewer—whereas measurement uncertainty predominated in the upper tail. At extreme thresholds, ratio-type measures showed widened CIs. 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. Importantly, optimal diagnostic thresholds t^{*} differed systematically across measures.

These results emphasise that threshold selection should be guided jointly by the intended clinical role and the local uncertainty profile, rather than by point estimates alone.

2.2. Relations between measures

Our relational analyses highlighted complementary insights. Prevalence-invariant indices captured the intrinsic separation between diseased and nondiseased 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 optimisation differs when guided by Bayes risk [43], 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.

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, the framework clarifies when analytical uncertainty dominates (e.g., at high thresholds affecting LR^+ and DOR) and when sampling uncertainty (e.g., at low thresholds with few diseased positives).

Quantifying and partitioning diagnostic uncertainty is imperative in laboratory medicine to define analytical performance specifications [44], manage quality and risk [45, 46], and design and implement test accuracy studies. Enhancing assay precision and standardisation 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 [47–55].

b) *Distributional misspecification:* Normal or lognormal assumptions are parsimonious but can be violated by latent mixtures [56], skewness differences, or heavy tails. Both unimodal and multimodal forms remain plausible [57–59]; model checking and sensitivity analyses are advisable.

c) *Prevalence handling:* Spectrum (case-mix) effects between source and target populations undermine the transportability of disease prevalence.

d) *Boundary behaviour:* Ratio measures exhibit unstable CIs at extreme thresholds, limiting interpretability; log-scale or Fieller-type intervals provide more reliable inference [60, 61].

e) *Sampling uncertainty:* Sampling-uncertainty approximations are exact under normality and used as first-order approximations under lognormal/gamma parameterisations; a bootstrap option is advisable.

f) *Selection uncertainty:* Reporting CIs at data-selected optimal thresholds understates uncertainty [62, 63]; bootstrap or cross-validated threshold re-estimation is recommended.

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

h) *Dependence among estimators:* Treating sample means and variances as independent can misstate uncertainty; likelihood-based or bootstrap methods better respect dependence [64–66].

i) *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 optimisation 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 nondiseased 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 revalidate 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 measurement uncertainty model parameters, and *t*-based CIs for lognormal and gamma distributions. Future work should explore semiparametric and mixture models, full parametric bootstrapping, cross-validated threshold selection, and generalisation to multi-class markers and decision-curve analysis [67–73].

3. Conclusion

This study introduced a unified, diagnostic threshold based computational framework for sixteen DAMs. The framework integrates point estimation, uncertainty quantification, and optimisation 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 measurements 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 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). Optimal thresholds vary systematically across DAM families: confirmatory diagnosis measures (e.g., LR^+, DOR) were characterised by higher thresholds, giving precedence to specificity; diagnosis for exclusion measures (e.g., LR^-) were characterised by lower thresholds, giving precedence to sensitivity; and agreement and association indices (e.g., JS, CZ, FMI, $F1S$, $C\kappa$, $PABAK$, MCC) selected interior thresholds balancing classification errors.

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. 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 medical measurements distributions in diseased and nondiseased 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: Conceptualisation: 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; visualisation: 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. Larner AJ. The 2x2 matrix: Contingency, confusion and the metrics of binary classification. Cham: Springer International Publishing; 2024.
7. 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.
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. White GH. Basics of estimating measurement uncertainty. Clin Biochem Rev. 2008;29 Suppl 1:S53-60.
10. Oosterhuis WP, Theodorsson E. Total error vs. measurement uncertainty: revolution or evolution? Clin Chem Lab Med. 2016;54:235–9.
11. Ayyub BM, Klir GJ. Uncertainty Modeling and Analysis in Engineering and the Sciences. Chapman and Hall/CRC; 2006.
12. Ellison SLR, Williams A. Quantifying Uncertainty in Analytical Measurement. 3rd edition. EURACHEM/CITAC; 2012.
13. 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.
14. Faraggi D. The effect of random measurement error on receiver operating characteristic (ROC) curves. Stat Med. 2000;19:61–70.
15. Reiser B. Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves. Stat Med. 2000;19:2115–29.
16. Perkins NJ, Schisterman EF. The Youden Index and the optimal cut-point corrected for measurement error. Biom J. 2005;47:428–41.
17. 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.
18. 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.
19. Bhaumik DK, Kapur K, Gibbons RD. Testing Parameters of a Gamma Distribution for Small Samples. Technometrics. 2009;51:326–34.

20. Goddard MJ, Hinberg I. Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. Stat Med. 1990;9:325–37.
21. Theodorsson E. Uncertainty in Measurement and Total Error: Tools for Coping with Diagnostic Uncertainty. Clin Lab Med. 2017;37:15–34.
22. Padoan A, Sciacovelli L, Aita A, Antonelli G, Plebani M. Measurement uncertainty in laboratory reports: A tool for improving the interpretation of test results. Clin Biochem. 2018;57:41–7.
23. Smith AF, Shinkins B, Hall PS, Hulme CT, Messenger MP. Toward a Framework for Outcome-Based Analytical Performance Specifications: A Methodology Review of Indirect Methods for Evaluating the Impact of Measurement Uncertainty on Clinical Outcomes. Clin Chem. 2019;65:1363–74.
24. Chatzimichail T, Hatjimihail AT. A Software Tool for Exploring the Relation between Diagnostic Accuracy and Measurement Uncertainty. Diagnostics (Basel). 2020;10. https://doi.org/10.3390/diagnostics10090610.
25. Chatzimichail T, Hatjimihail AT. A software tool for applying Bayes' theorem in medical diagnostics. BMC Med Inform Decis Mak. 2024;24:399.
26. Agresti A, Franklin C, Klingenberg B. Statistics: The art and science of learning from data, global edition. 4th edition. London, England: Pearson Education; 2023.
27. Miller J, Miller JC. Statistics and Chemometrics for Analytical Chemistry. 7th edition. London, England: Pearson Education; 2018.
28. J. Aitchison JACB. The Lognormal Distribution with special reference to its uses in econometrics. Cambridge: Cambridge University Press; 1957.
29. Agresti A, Coull BA. Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. Am Stat. 1998;52:119–26.
30. Welch BL. The Generalization of `Student's' Problem when Several Different Population Variances are Involved. Biometrika. 1947;34:28–35.
31. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics. 1946;2:110–4.
32. 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.
33. 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.
34. 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.
35. 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.
36. 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.
37. Petrone S, Rousseau J, Scricciolo C. Bayes and empirical Bayes: do they merge? Biometrika. 2014;101:285–302.
38. Myung IJ. Tutorial on maximum likelihood estimation. J Math Psychol. 2003;47:90–100.
39. Johnson ML. Nonlinear least-squares fitting methods. In: Methods Cell Biol. Academic Press; 2008. p. 781–805.
40. Bates DM, Watts DG. Nonlinear Regression Analysis and Its Applications. Hoboken, New Jersey: John Wiley & Sons, Inc.; 1988.
41. Darling DA. The Kolmogorov-Smirnov, Cramer-von Mises Tests. Ann Math Stat. 1957;28:823–38.
42. Fred HL. The diagnosis of exclusion: an ongoing uncertainty. Tex Heart Inst J. 2013;40:379–81.
43. Kelly DL, Smith CL. Bayesian inference in probabilistic risk assessment—The current state of the art. Reliab Eng Syst Saf. 2009;94:628–43.

44. 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.
45. Hatjimihail AT. Estimation of the optimal statistical quality control sampling time intervals using a residual risk measure. PLoS One. 2009;4:e5770.
46. Nichols, James H., Altaie, Sousan S., Cooper Greg,Glavina Paul, Halim Abdel-Baset, Hatjimihail Aristides T., et al. Laboratory Quality Control Based on Risk Management; Approved Guideline. Clinical and Laboratory Standards Institute; 2011.
47. Rao SS, Disraeli P, McGregor T. Impaired glucose tolerance and impaired fasting glucose. Am Fam Physician. 2004;69:1961–8.
48. Meneilly GS, Elliott T. Metabolic alterations in middle-aged and elderly obese patients with type 2 diabetes. Diabetes Care. 1999;22:112–8.
49. Geer EB, Shen W. Gender differences in insulin resistance, body composition, and energy balance. Gend Med. 2009;6 Suppl 1 Suppl 1:60–75.
50. Van Cauter E, Polonsky KS, Scheen AJ. Roles of circadian rhythmicity and sleep in human glucose regulation. Endocr Rev. 1997;18:716–38.
51. 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.
52. 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.
53. 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.
54. Pandit MK, Burke J, Gustafson AB, Minocha A, Peiris AN. Drug-induced disorders of glucose tolerance. Ann Intern Med. 1993;118:529–39.
55. 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.
56. 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.
57. Wilson JMG, Jungner G. Principles and practice of screening for disease. Geneva: World Health Organization; 1968.
58. Petersen PH, Horder M. 2.3 Clinical test evaluation. Unimodal and bimodal approaches. Scand J Clin Lab Invest. 1992;52:51–7.
59. Fischer NI, Mammen E, Marron JS. Testing for multimodality. Comput Stat Data Anal. 1994;18:499–512.
60. West RM. Best practice in statistics: The use of log transformation. Ann Clin Biochem. 2022;59:162–5.
61. Fieller EC. The distribution of the index in a normal bivariate population. Biometrika. 1932;24:428–40.
62. 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.
63. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. Ann Intern Med. 2011;154:253–9.
64. 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.
65. Pawitan Y. In all likelihood: Statistical modelling and inference using likelihood. Oxford University PressOxford; 2001.
66. Dikta G, Scheer M. Bootstrap methods: With applications in R. 2021st edition. Cham, Switzerland: Springer Nature; 2021.
67. Branscum AJ, Johnson WO, Hanson TE, Gardner IA. Bayesian semiparametric ROC curve estimation and disease diagnosis. Stat Med. 2008;27:2474–96.

68. Davison AC, Hinkley DV. Cambridge series in statistical and probabilistic mathematics: Bootstrap methods and their application series number 1. Cambridge, England: Cambridge University Press; 2013. https://doi.org/10.1017/cbo9780511802843.
69. Zheng J, Frey HC. Quantification of variability and uncertainty using mixture distributions: evaluation of sample size, mixing weights, and separation between components. Risk Anal. 2004;24:553–71.
70. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn. 2001;45:171–86.
71. Thiele C, Hirschfeld G. cutpointr: Improved Estimation and Validation of Optimal Cutpoints in *R*. J Stat Softw. 2021;98:1–27.
72. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565–74.
73. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18.

Appendices

A.1. List of abbreviations

DAM: diagnostic accuracy measure

GUM: Guide to the Expression of Uncertainty in Measurement

OGTT: oral glucose tolerance test

ADA: American Diabetes Association

ROC: receiver operating characteristic

A.2. Notation

A.2.1. Populations

$\bar{D}$: 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 : 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

μ_P: mean of the measurements of a test in the population P

σ_P: standard deviation of the measurements of a test in 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 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. Input

A.3.1. 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.3.2. Additional input options

A.3.2.1. Plots

Users can select between an extended and limited plot range.

A.3.2.2. Tables

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

A.3.3. 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 are finely manipulated by pressing the *alt* or *opt* key while dragging the mouse. Pressing the *shift* or *ctrl* keys can even more finely manipulate them.

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.4. Software availability and requirements

Program name: *DiagAccU*

Version: 1.0.0

Project home page: https://www.hcsl.com/Tools/DiagnosticAccuracy/ (accessed on January 30, 2026)

Program source: *DiagAccU.nb*

Available to download as a ZIP archive at:
https://www.hcsl.com/Tools/DiagnosticAccuracy/DiagAccU.zip (accessed on January 30, 2026)

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 January 30, 2026) 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: February 1, 2026