

Article

Eigenvalue-Based Diagnostic Equity Testing: A Random Matrix Framework for Detecting Multi-Dimensional Performance Disparities in Clinical Classifiers

Oyebayo Ridwan Olaniran ^{1,*} , Ali Rashash R. Alzahrani ² , Mohammed H. Alharbi ³ ,
Nada Mohammed Saeed Alharbi ⁴ , Asma Ahmad Alzahrani ⁵  and Saheed Ajibade Kunle ⁶ 

¹ Department of Statistics, Faculty of Physical Sciences, University of Ilorin, Ilorin 1515, Nigeria

² Mathematics Department, Faculty of Sciences, Umm Al-Qura University, Makkah 24382, Saudi Arabia; arrzahrani@uqu.edu.sa

³ Department of Mathematics and Statistics, Faculty of Science, University of Jeddah, P.O. Box 80327, Jeddah 21589, Saudi Arabia; mhhharbi1@uj.edu.sa

⁴ Department of Mathematics, Faculty of Science, Taibah University, Al-Madinah Al-Munawara 42353, Saudi Arabia; nmharbi@taibahu.edu.sa

⁵ Department of Mathematics, Faculty of Science, Al-Baha University, Al-Baha 65779, Saudi Arabia; asmaalzahrani@bu.edu.sa

⁶ Department of Mathematics and Statistics, Kwara State University, Malete 241104, Nigeria; saheed.ajibade15@kwasu.edu.ng

* Correspondence: olaniran.or@unilorin.edu.ng

Abstract

Evaluating whether clinical classifiers perform equitably across patient subgroups is a central requirement for the responsible deployment of machine learning in medicine. Conventional approaches test one fairness metric at a time, such as sensitivity, positive predictive value, or area under the receiver operating characteristic curve, and therefore cannot detect disparities that manifest only in the joint structure of a group-specific confusion matrix. We develop a unified hypothesis-testing framework rooted in random matrix theory that compares demographic groups through the L^2 distance between their joint eigenvalue densities, yielding a scalar spectral divergence that is sensitive to every cell of the 2×2 confusion matrix simultaneously. We derive the closed-form spectral divergence for Gaussian-approximated eigenvalue densities, prove almost-sure consistency of the empirical estimator via the delta method, and construct an extreme-value (Gumbel) test statistic with family-wise error rate control. Monte Carlo experiments comprising 10,000 replications across balanced, moderately imbalanced, and severely imbalanced group-size regimes show that the spectral test keeps Type I errors close to its nominal level while achieving power exceeding 90% in complex and multi-dimensional violation scenarios, where the best single-metric competitor reaches at most 63%. Three clinical benchmark datasets from the UCI Machine Learning Repository utilised include Pima Indians Diabetes ($n = 768$), Cleveland Heart Disease ($n = 303$), and Heart Failure Clinical Records ($n = 299$). Results confirm that the spectral method detects statistically significant ($p < 0.001$) performance disparities missed by all three conventional tests. These results support eigenvalue-based divergence as a practical, model-agnostic diagnostic equity tool for clinical machine learning audits.



Academic Editor: Tiago Dias Domingues

Received: 19 April 2026

Revised: 19 June 2026

Accepted: 25 June 2026

Published: 17 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: diagnostic equity; confusion matrix; eigenvalue distribution; random matrix theory; spectral divergence; clinical fairness; machine learning evaluation; hypothesis testing

MSC: 94A08; 68U10

1. Introduction

The adoption of machine learning (ML) in clinical decision-making, including disease screening, risk stratification, and treatment allocation, has accelerated markedly in the past decade. With this acceleration comes an obligation: clinical classifiers must perform equitably across patient subgroups defined by demographic or clinical attributes such as sex, age stratum, or comorbidity profile. When a classifier systematically under-detects disease in one subgroup or over-classifies healthy patients in another, the resulting disparities translate directly into differential harm, including delayed diagnosis, unnecessary treatment, or misallocated resources [1,2].

The statistical tool that most completely summarises the performance of a binary classifier at a fixed decision threshold is the confusion matrix 2×2 , which records the joint frequencies of true positives (TPs), false positives (FPs), false negatives (FNs), and true negatives (TNs). Every standard performance metric, such as sensitivity, specificity, positive predictive value, negative predictive value, and the F_1 score, is a scalar function of these four counts. When fairness is the concern, the standard practice is to compute this metric separately for each demographic group and then to test whether the group-specific values are equal. For example, the equal opportunity criterion of [3] requires that sensitivity be identical between groups; demographic parity requires equal predicted positive rates; and predictive parity requires equal positive predictive values.

A fundamental limitation of this metric-by-metric approach is that it reduces the four-dimensional confusion matrix to a single scalar prior to comparison. Two groups may have identical sensitivities yet markedly different false positive rates, or identical positive predictive values yet disparate true negative proportions. Such “hidden” disparities are invisible to any test that conditions on a single metric. Moreover, refs. [4,5] showed that, except in degenerate cases, it is impossible to simultaneously satisfy multiple fairness criteria. This impossibility result implies that a comprehensive diagnostic equity assessment cannot rely on testing a finite list of univariate metrics; instead, it must interrogate the full structure of group-specific confusion matrices.

Random matrix theory (RMT) provides a natural mathematical language for such structural comparisons. The eigenvalues of a symmetric matrix encode its variance structure in a rotation-invariant manner: they are unchanged by any orthogonal transformation of the coordinate system and therefore capture intrinsic geometric properties rather than artefacts of a particular parameterisation. Building on the eigenvalue distributions derived for random confusion matrices in [6], we propose comparing demographic groups by measuring the distance L^2 between their joint eigenvalue densities, which we term *spectral diagnostic divergence*.

The contributions of this paper are as follows. First, we define spectral diagnostic divergence and derive its closed-form expression under Gaussian-approximated eigenvalue densities for 2×2 confusion matrices (Section 3). Second, we prove that the empirical estimator is strongly consistent and asymptotically normal, with an explicit delta-method variance (Section 3). Third, we construct an extreme-value test statistic whose null distribution is Gumbel, allowing family-wise error rate control over all pairwise group comparisons (Section 3). Fourth, we conduct a comprehensive Monte Carlo study with 10,000 replications per scenario, providing empirical evidence that the spectral test attains higher

power than conventional sensitivity, positive predictive value, and AUC-based tests in the complex and multi-dimensional fairness-violation regimes considered, while maintaining controlled Type I errors (Sections 4 and 5). We emphasise that this is an empirical comparison based on the stated data-generating processes, rather than a claim of universal dominance. Fifth, we validate the framework on three clinical datasets from the UCI Machine Learning Repository—Pima Indians Diabetes [7], Cleveland Heart Disease [8], and Heart Failure Clinical Records [9]—showing that the spectral test identifies significant performance disparities that are missed not only by all three single-metric tests but also, in the Cleveland and Heart Failure datasets, by multivariate multinomial homogeneity tests (Section 6).

Motivation and Positioning Relative to Existing Fairness Assessments

The practical advantage of the proposed approach is best understood as one of *test multiplicity* versus *omnibus screening*. A conventional fairness audit selects a metric (for example, the true positive rate for the equal-opportunity criterion of [3], the predicted positive rate for demographic parity [10], or the positive predictive value for predictive parity [5]) and runs a separate group-comparison test for each. Because the impossibility results of [4,5] preclude simultaneous satisfaction of these criteria except in degenerate cases, an auditor who wishes to be comprehensive must run many such tests and then correct for multiplicity, which erodes power; alternatively, the auditor selects a small number of metrics in advance and accepts blindness to any disparity that those metrics do not encode. The spectral divergence offers a different trade-off: a *single* scalar statistic computed from the entire group-specific confusion matrix, requiring one test rather than a battery of corrected comparisons. This is the sense in which the method is “model-agnostic” and “metric-agnostic”: it does not commit in advance to which margin of the confusion matrix a disparity will appear.

Our framework is therefore complementary to, rather than a replacement for, the substantial body of work on algorithmic fairness. Group-fairness criteria and their incompatibilities are reviewed by [11–13]; pre-processing, in-processing, and post-processing remedies are developed by [14–16]; and the specific challenges of fairness in clinical and healthcare machine learning are documented by [1,2,17,18]. Relative to these approaches, the contribution here is narrowly diagnostic: it provides a principled *detection* statistic at the confusion-matrix level, not a remediation procedure. We stress at the outset that the confusion-matrix view is deliberately restricted. Fairness notions that depend on the predicted score distribution rather than on a thresholded decision, such as calibration fairness [4], ranking fairness, and sensitivity to the choice of decision threshold, are outside the scope of a fixed-threshold confusion matrix and are not captured by the present test. We return to these boundaries in Section 7.

2. Confusion Matrices and Fairness Metrics

2.1. Notation and Statistical Model

Consider a binary clinical classifier evaluated on a population divided into demographic groups G indexed by $g \in \{1, 2, \dots, G\}$. Let N_g denote the number of patients in group g . For the reader’s convenience, Table 1 collects the principal symbols used throughout the paper.

Table 1. Summary of principal notation.

Symbol	Meaning
G, g, h	Number of demographic groups; group indices, $g, h \in \{1, \dots, G\}$
N_g	Number of patients in group g
$Y, \hat{Y}$	True and predicted binary labels
C_g	Group-specific 2×2 confusion matrix (Definition 1)
TP_g, FP_g, FN_g, TN_g	Cell counts of C_g (true/false positives/negatives)
$\mathbf{p}_g = (p_{g,11}, p_{g,10}, p_{g,01}, p_{g,00})$	Cell probabilities; $\mathbf{p}_g \in \Delta^3$ (probability simplex)
$\mathbf{v}_g$	Normalised performance vector, $\mathbf{v}_g = C_g / N_g$
$\bar{A}_g$	Mean of the entries of $\mathbf{v}_g$
s_g^2	Sample variance (dispersion) of the entries of $\mathbf{v}_g$
λ_1, λ_2	Eigenvalues of the symmetrised confusion matrix
$f_k(\lambda_1, \lambda_2)$	Joint eigenvalue density for group k (Equation (14))
TPR, PPV, AUC	Sensitivity, positive predictive value, area under ROC curve
$D(g, h)$	Exact L^2 spectral diagnostic divergence (Equation (15))
$\hat{D}(g, h)$	Empirical spectral divergence statistic (Definition 4)
Z_{gh}, T	Standardised pairwise statistic; test statistic $T = \max_{g \neq h} Z_{gh}$
$M = \binom{G}{2}$	Number of distinct group pairs
a_M, b_M	Gumbel centring and scaling constants
$\hat{\tau}^2$	Average estimated pairwise variance
ϵ	Pre-specified equity tolerance
α, t_α	Significance level; critical value controlling the FWER

Definition 1 (Group-Specific Confusion Matrix). *The group-specific confusion matrix is the 2×2 random matrix*

$$C_g = \begin{bmatrix} TP_g & FP_g \\ FN_g & TN_g \end{bmatrix}, \quad (1)$$

where $[TP_g, FP_g, FN_g, TN_g]^\top \sim \text{Multinomial}(N_g, \mathbf{p}_g)$ with $\mathbf{p}_g = (p_{g,11}, p_{g,10}, p_{g,01}, p_{g,00})^\top \in \Delta^3$, the probability simplex.

The four cell probabilities admit a direct probabilistic interpretation:

$$\begin{aligned} p_{g,11} &= \Pr(\hat{Y} = 1, Y = 1), & p_{g,10} &= \Pr(\hat{Y} = 1, Y = 0), \\ p_{g,01} &= \Pr(\hat{Y} = 0, Y = 1), & p_{g,00} &= \Pr(\hat{Y} = 0, Y = 0), \end{aligned} \quad (2)$$

where Y and $\hat{Y}$ denote the true and predicted labels within group g .

2.2. Scalar Fairness Metrics

Three metrics commonly used in clinical fairness assessments are defined below.

Sensitivity (True Positive Rate).

$$\text{TPR}_g = \frac{p_{g,11}}{p_{g,11} + p_{g,01}} = \Pr(\hat{Y} = 1 \mid Y = 1). \quad (3)$$

Proposition 1 (Asymptotic distribution of TPR). *Let $N_g^+ = TP_g + FN_g$. As $N_g^+ \rightarrow \infty$, $\sqrt{N_g^+}(\text{TPR}_g - \theta_g) \xrightarrow{d} \mathcal{N}(0, \theta_g(1 - \theta_g))$, where $\theta_g = p_{g,11} / (p_{g,11} + p_{g,01})$.*

Positive Predictive Value (PPV).

$$\text{PPV}_g = \frac{p_{g,11}}{p_{g,11} + p_{g,10}} = \Pr(Y = 1 \mid \hat{Y} = 1). \quad (4)$$

Area Under the ROC Curve (AUC).

$$\text{AUC}_g = \Pr(S_g^+ > S_g^-), \quad (5)$$

where S_g^+ and S_g^- are the classifier's score distributions for positive and negative cases in group g .

2.3. Limitations of Univariate Fairness Tests

For any metric $M \in \{\text{TPR}, \text{PPV}, \text{AUC}\}$, the conventional test is

$$H_0: M_1 = M_2 = \dots = M_G \quad \text{versus} \quad H_1: \exists g \neq h \text{ such that } M_g \neq M_h, \quad (6)$$

implemented via the chi-squared statistic

$$\chi_M^2 = \sum_{g=1}^G \frac{(\hat{M}_g - \bar{M})^2}{\widehat{\text{Var}}(M_g)}. \quad (7)$$

The fundamental deficiency is that such a test is blind to any disparity that does not affect M . Two groups with identical TPRs but dramatically different FPRs will pass the equal-opportunity test; two groups with identical PPV but different TNRs will pass the predictive-parity test. The impossibility theorems of [4,5] guarantee that such hidden disparities generically exist whenever multiple fairness definitions are not simultaneously satisfiable. What is needed, therefore, is a test that is sensitive to every cell of the confusion matrix.

Figure 1 summarises the pipeline developed in the remainder of the paper: each group's 2×2 confusion matrix is normalised and symmetrised, mapped to a joint eigenvalue density, compared between groups through an L^2 spectral divergence, and finally aggregated across group pairs into an extreme-value test statistic with family-wise error control. The key conceptual contrast with the univariate tests above is that the scalar metrics each collapse the confusion matrix onto a single axis, whereas the spectral divergence retains the joint cell structure.

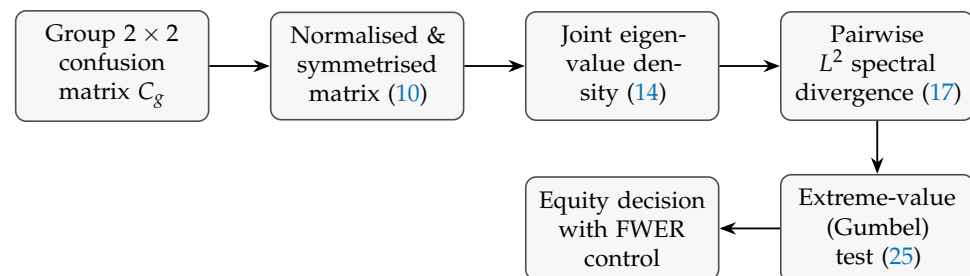


Figure 1. Schematic of the spectral diagnostic equity pipeline. Each demographic group's confusion matrix is reduced to a joint eigenvalue density; groups are compared pairwise through an L^2 divergence, and the pairwise divergences are aggregated into a single extreme-value test statistic with family-wise error rate control. Equation numbers indicate where each step is defined.

3. Eigenvalue-Based Diagnostic Equity Framework

3.1. Normalised Performance Vectors

Definition 2 (Normalised Performance Vector). For group g , define the normalised performance vector

$$\mathbf{v}_g = \frac{1}{N_g} (\text{TP}_g, \text{FP}_g, \text{FN}_g, \text{TN}_g)^\top \in \mathbb{R}^4. \quad (8)$$

Each component of $\mathbf{v}_g$ estimates the corresponding cell probability $p_{g,ij}$ and $\mathbf{1}^\top \mathbf{v}_g = 1$. Although the four components are in a three-dimensional simplex, we retain all four for interpretability. The sample mean and sample variance of the entries,

$$\bar{A}_g = \frac{1}{4} \sum_i v_{g,i}, \quad s_g^2 = \frac{1}{3} \sum_i (v_{g,i} - \bar{A}_g)^2, \quad (9)$$

summarise the location and dispersion of the proportions of the confusion-matrix cells and serve as the two parameters of the eigenvalue density introduced in Section 3. A consequence of the constraint $\mathbf{1}^\top \mathbf{v}_g = 1$ is that the location $\bar{A}_g = 1/4$ is shared by every group; the discriminating content of the divergence lies mainly in the dispersion s_g^2 , which quantifies how unevenly the unit of probability mass is distributed across the four cells. The location term $(\bar{A}_g - \bar{A}_h)^2$ is retained in the general expressions below so that the estimator remains well defined for unnormalised or alternatively scaled confusion-matrix representations, but it vanishes under the present normalisation.

Remark 1 (What the second-moment summary does and does not capture). Because $\hat{D}(g, h)$ is a function of $(\bar{A}_g, s_g^2, \bar{A}_h, s_h^2)$ alone, it is invariant under any rearrangement of the proportions of the four cells of a group that leaves their first two moments unchanged. Two confusion matrices that differ by such a permutation, for example, by interchanging the diagonal (correct) and anti-diagonal (incorrect) mass, are assigned divergence zero, even though they describe opposite classifier behaviour. The statistic is thus best understood as a test for differences in the concentration profile of the confusion matrix across groups, rather than as a complete sufficient statistic for the multinomial law. We make this limitation explicit here in response to the legitimate concern that a low-order summary can discard structural information; the empirical results of Section 6 indicate that, for the disparities arising in the clinical datasets studied, group differences manifest as differences in s_g^2 and are consequently detected. A divergence built from the full eigenvalue density (Equation (16)) rather than its Gaussian summary, which would retain higher-order structure, is a natural refinement discussed in Section 7.

3.2. From the Multinomial Model to the Joint Eigenvalue Density

Before defining the divergence, we make explicit why a spectral representation is appropriate for a matrix 2×2 and how the joint eigenvalue density used below arises from the multinomial sampling model. This addresses two natural objections: that random matrix theory (RMT) is ordinarily a high-dimensional theory, and that the density is invoked rather than derived.

3.2.1. Why Eigenvalues of a 2×2 Matrix

We do not invoke asymptotic RMT spectra such as the Marchenko–Pastur or Tracy–Widom laws, which indeed require the matrix dimension to grow [19,20]. Our use of spectral information is finite-dimensional and exact: the two eigenvalues of a symmetric 2×2 matrix are the two orthogonal invariants of that matrix under the rotation group $O(2)$. Any rotation-invariant comparison of two such matrices must be a function of their eigenvalues, and conversely, the pair (λ_1, λ_2) determines the symmetric matrix up to orthogonal conjugation. Comparison of groups through the *joint law* of (λ_1, λ_2) is therefore a coordinate-free way of comparing the full second-order structure of the confusion matrix, in contrast to comparing a hand-chosen scalar margin. The role of RMT here is to provide the correct sampling density of this eigenvalue pair, not to provide a limiting spectral distribution.

3.2.2. Symmetrisation

The confusion matrix C_g in Equation (1) is not symmetric, so we work with its symmetric part $\frac{1}{2}(C_g + C_g^\top)$, equivalently with the symmetric matrix

$$S_g = \begin{bmatrix} v_{g,1} & w_g \\ w_g & v_{g,4} \end{bmatrix}, \quad w_g = \frac{1}{2}(v_{g,2} + v_{g,3}), \quad (10)$$

whose diagonal carries the correctly classified proportions $(\text{TP}_g, \text{TN}_g)/N_g$ and whose off-diagonal carries the mean misclassification proportion. The symmetric part preserves the interpretation of the eigenvalues as variance directions while discarding only the skew-symmetric component, which for a 2×2 matrix contributes a single antisymmetric degree of freedom that does not affect the rotation-invariant comparison.

3.2.3. Gaussian Approximation of the Entries

By Definition 1, the cell counts are multinomial, so for large N_g , the multivariate central limit theorem gives

$$\sqrt{N_g}(\mathbf{v}_g - \mathbf{p}_g) \xrightarrow{d} \mathcal{N}_4(\mathbf{0}, \text{diag}(\mathbf{p}_g) - \mathbf{p}_g \mathbf{p}_g^\top), \quad (11)$$

so that, to first order, the entries of S_g are jointly Gaussian with common location $\bar{A}_g$ and a dispersion governed by s_g^2 . Following [6], we adopt the tractable isotropic approximation in which the independent entries of the symmetric matrix share variance s_g^2 ; this is the 2×2 Gaussian orthogonal ensemble (GOE) with nonzero mean, and it is the simplest model consistent with the first two moments in Equation (9). The accuracy of this approximation for finite clinical samples is examined in Section 3.4, and its limitations are discussed in Section 7.

3.2.4. Change in Variables to Eigenvalues

For a symmetric matrix $S = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$, the eigenvalues are

$$\lambda_{1,2} = \frac{a+c}{2} \pm \sqrt{\left(\frac{a-c}{2}\right)^2 + b^2}, \quad (12)$$

and the spectral decomposition $S = Q \text{diag}(\lambda_1, \lambda_2) Q^\top$ with $Q \in O(2)$ parametrised by a single angle ϕ furnishes a change in variables $(a, b, c) \mapsto (\lambda_1, \lambda_2, \phi)$. Its Jacobian is the classical Vandermonde factor

$$da db dc = \frac{1}{2} |\lambda_1 - \lambda_2| d\lambda_1 d\lambda_2 d\phi, \quad (13)$$

which is exactly the level-repulsion term responsible for the $|\lambda_1 - \lambda_2|$ factor in the density. Substituting the isotropic Gaussian weight $\exp(-[(a - \bar{A}_k)^2 + 2b^2 + (c - \bar{A}_k)^2]/(2s_k^2))$ into the transformation (13), expressing the exponent through $\lambda_1 + \lambda_2 = a + c$ and $\lambda_1^2 + \lambda_2^2 = a^2 + c^2 + 2b^2$, and integrating the angle ϕ over $[0, \pi)$ removes the eigenvector orientation and yields the marginal joint density of the eigenvalues,

$$f_k(\lambda_1, \lambda_2) = \frac{1}{4s_k^2\sqrt{\pi}} \exp\left(-\frac{\lambda_1^2 + \lambda_2^2 - 2\bar{A}_k(\lambda_1 + \lambda_2) + 2\bar{A}_k^2}{2s_k^2}\right) |\lambda_1 - \lambda_2|, \quad (14)$$

which is the density derived in [6] and the starting point for the divergence below. The derivation makes clear that the $|\lambda_1 - \lambda_2|$ factor is the Jacobian of the eigendecomposition rather than an external modelling choice, and that $\bar{A}_k$ and s_k^2 enter as the only free parameters.

3.3. Spectral Diagnostic Divergence

Using the joint eigenvalue density (14) for 2×2 random confusion matrices derived in [6], we measure the discrepancy between two groups as follows.

Definition 3 (Spectral Diagnostic Divergence). For groups g and h , the spectral diagnostic divergence is

$$D(g, h) = \left(\iint_{\mathbb{R}^2} [f_g(\lambda_1, \lambda_2) - f_h(\lambda_1, \lambda_2)]^2 d\lambda_1 d\lambda_2 \right)^{1/2}. \quad (15)$$

Theorem 1 (Closed-Form Divergence). Under the density (14), the squared spectral diagnostic divergence admits the exact closed form

$$D^2(g, h) = \frac{1}{8} - \frac{1}{4} \frac{s_g s_h}{s_g^2 + s_h^2} \exp\left(-\frac{(\bar{A}_g - \bar{A}_h)^2}{2(s_g^2 + s_h^2)}\right). \quad (16)$$

Writing $\sigma^2 = \frac{1}{2}(s_g^2 + s_h^2)$, $t = (s_g^2 - s_h^2)/(s_g^2 + s_h^2) \in (-1, 1)$, and $q = (\bar{A}_g - \bar{A}_h)^2/(2(s_g^2 + s_h^2)) \geq 0$, this is equivalent to

$$D^2(g, h) = \frac{1}{8} (1 - \sqrt{1 - t^2} e^{-q}) = \frac{1}{8} (1 - e^{-\varphi}), \quad \varphi := q - \frac{1}{2} \ln(1 - t^2) \geq 0, \quad (17)$$

so that $D(g, h)$ is bounded, $0 \leq D(g, h) < 1/\sqrt{8}$, and equals zero if and only if $\bar{A}_g = \bar{A}_h$ and $s_g = s_h$. For small parameter differences, expanding (17) to second order gives the leading-order behaviour

$$D^2(g, h) = \frac{(\bar{A}_g - \bar{A}_h)^2}{32\sigma^2} + \frac{(s_g^2 - s_h^2)^2}{64\sigma^4} + R, \quad (18)$$

where the remainder R is bounded explicitly in Proposition 3.

Proof. We expand $D^2(g, h) = I_{gg} + I_{hh} - 2I_{gh}$, where $I_{kk} = \iint f_k^2 d\lambda_1 d\lambda_2$ and $I_{gh} = \iint f_g f_h d\lambda_1 d\lambda_2$.

Self-interaction term. With the substitution $u = (\lambda_1 - \bar{A}_k)/s_k$, $v = (\lambda_2 - \bar{A}_k)/s_k$,

$$I_{kk} = \frac{1}{16\pi} \iint_{\mathbb{R}^2} e^{-(u^2+v^2)} (u-v)^2 du dv. \quad (19)$$

Standard Gaussian moment identities give $\iint e^{-(u^2+v^2)} (u-v)^2 du dv = \pi$, so $I_{kk} = 1/16$.

Cross-interaction term. Using the product-of-Gaussians identity $e^{-a(x-m_1)^2} e^{-b(x-m_2)^2} = e^{-(a+b)(x-\bar{m})^2} e^{-\frac{ab}{a+b}(m_1-m_2)^2}$, where $\bar{m} = (am_1 + bm_2)/(a+b)$, we factor the integrand in both λ_1 and λ_2 . After completing the square and evaluating the resulting Gaussian integrals, we obtain

$$I_{gh} = \frac{1}{8} \frac{s_g s_h}{s_g^2 + s_h^2} \exp\left(-\frac{(\bar{A}_g - \bar{A}_h)^2}{2(s_g^2 + s_h^2)}\right). \quad (20)$$

Substituting $I_{gg} = I_{hh} = 1/16$ and I_{gh} into $D^2 = I_{gg} + I_{hh} - 2I_{gh}$ yields (16). Identity (17) follows from $s_g s_h/(s_g^2 + s_h^2) = \frac{1}{2} \sqrt{1-t^2}$ and $(\bar{A}_g - \bar{A}_h)^2/(2(s_g^2 + s_h^2)) = q$, and from collecting the two factors into the single exponent $\varphi = q - \frac{1}{2} \ln(1-t^2)$. The leading-order form (18) is obtained in Proposition 3. $\square$

Corollary 1 (Special cases of the leading-order divergence). To leading order in the group differences, with $\sigma^2 = \frac{1}{2}(s_g^2 + s_h^2)$,

1. **Equal variance.** If $s_g = s_h = s$, then $D(g, h) \approx |\bar{A}_g - \bar{A}_h|/(\sqrt{32}s)$.
2. **Equal mean.** If $\bar{A}_g = \bar{A}_h$, then $D(g, h) \approx |s_g^2 - s_h^2|/(8\sigma^2)$.
3. **Identical groups.** $D(g, h) = 0$ if and only if $\bar{A}_g = \bar{A}_h$ and $s_g = s_h$.

In each non-degenerate case, the divergence is a strictly increasing function of the corresponding parameter gap, so the ordering of group dissimilarity is preserved by any monotone rescaling of D .

Proposition 2 (Metric properties). *The function $D: \mathcal{G} \times \mathcal{G} \rightarrow [0, 1/\sqrt{8})$ satisfies non-negativity, symmetry, identity of indiscernibles, and the triangle inequality (via Minkowski's inequality on L^2) and is therefore a bounded metric on the space of eigenvalue densities.*

3.4. Validity of the Gaussian and Small-Difference Approximations

The framework rests on two approximations whose validity is quantified by the Gaussian (GOE) approximation of the multinomial entries in Equation (11) and the small-difference expansion (18). We address the second analytically and the first by indicating its regime of accuracy.

Proposition 3 (Approximation error of the leading-order divergence). *With t , q , φ , and σ^2 as in Theorem 1, the remainder in Equation (18) satisfies*

$$|R| \leq \frac{1}{8} \underbrace{\left(-\frac{1}{2} \ln(1-t^2) - \frac{t^2}{2} + \frac{\varphi^2}{2} \right)}_{=O(t^4)} \leq \frac{1}{8} \left(\frac{t^4}{4(1-t^2)} + \frac{\varphi^2}{2} \right). \quad (21)$$

Consequently, $R = O((q + t^2)^2)$, and the relative error of (18) is $O(q + t^2)$. The approximation is therefore accurate precisely when the squared standardised mean gap q and the relative variance gap $|t|$ are small, that is, when the two groups have comparable cell dispersion; for severe disparities (in particular structural zeros, where $t \rightarrow 1$), the exact form (17) should be used instead.

Proof. From (17), $D^2 = \frac{1}{8}(1 - e^{-\varphi})$ with $\varphi = q - \frac{1}{2} \ln(1 - t^2)$. Since $-\frac{1}{2} \ln(1 - t^2) = \frac{t^2}{2} + \frac{t^4}{4} + \dots \geq \frac{t^2}{2}$, we have $\varphi = q + \frac{t^2}{2} + \rho(t)$ with $0 \leq \rho(t) = -\frac{1}{2} \ln(1 - t^2) - \frac{t^2}{2} \leq \frac{t^4}{4(1-t^2)}$. The elementary bounds $\varphi - \frac{\varphi^2}{2} \leq 1 - e^{-\varphi} \leq \varphi$ give $|(1 - e^{-\varphi}) - \varphi| \leq \frac{\varphi^2}{2}$. Writing the leading term as $A = q + \frac{t^2}{2}$ and noting $D^2 = \frac{1}{8}A + R$ with $R = \frac{1}{8}[(1 - e^{-\varphi}) - A]$, the triangle inequality yields $|R| \leq \frac{1}{8}(\rho(t) + \frac{\varphi^2}{2})$, which is (21). Substituting $q = (\bar{A}_g - \bar{A}_h)^2 / (4\sigma^2)$ and $t^2/2 = (s_g^2 - s_h^2)^2 / (8\sigma^4)$ into $\frac{1}{8}A$ gives the explicit leading terms of (18). $\square$

Remark 2 (Finite-sample validity of the Gaussian approximation). *The Gaussian approximation (11) carries the usual multinomial Berry–Esseen rate $O(N_g^{-1/2})$, which degrades when a cell probability is close to 0 or 1, i.e., for rare events or near-deterministic predictions in small groups. For the smallest subgroups in our clinical data (N_g between 72 and 131), the expected count in every retained cell exceeds five, the conventional adequacy threshold for normal approximation of multinomial counts; nonetheless, for groups with very few positives, the permutation alternative noted in Section 7 is preferable. This is the principal reason we report the exact divergence (17) alongside the leading-order statistic.*

3.5. Statistical Estimation and Consistency

Definition 4 (Empirical Spectral Divergence). *Given observed confusion matrices, the empirical spectral divergence statistic is*

$$\hat{D}(g, h) = \sqrt{\frac{(\hat{A}_g - \hat{A}_h)^2 + (\hat{s}_g^2 - \hat{s}_h^2)^2}{\hat{s}_g^2 \hat{s}_h^2}}, \quad (22)$$

where $\hat{A}_g$ and $\hat{s}_g^2$ are computed from the normalised performance vector $\mathbf{v}_g$. Its population counterpart $\mathcal{D}(g, h)$ replaces the estimates by the corresponding parameters.

Remark 3 (Relationship between $\hat{D}$ and the exact divergence). *The statistic (22) is a standardised structural distance between the moment summaries of the two groups; unlike the exact L^2 divergence (17), it is unbounded, which gives it finer numerical resolution for ranking disparities*

and is why the reported values exceed the bound $1/\sqrt{8}$ of the exact metric. The two are nonetheless monotonically linked: under comparable variances ($t \rightarrow 0$, so $\hat{s}_g \hat{s}_h \approx \sigma^2$), Equation (18) gives $\mathcal{D}(g, h) \approx 8D(g, h)$, and more generally, $\mathcal{D}$ is a strictly increasing function of D along each of the parameter gaps in Corollary 1. A test based on $\hat{D}$ therefore detects exactly the alternatives $D(g, h) > 0$, while $\mathcal{D}$ serves as the working statistic. We use $\hat{D}$ throughout for consistency with this rescaling. The heuristic monotonicity asserted here is made rigorous in Proposition 4.

Proposition 4 (Exact monotone link between the exact and structural divergences). *Under the simplex normalisation $\bar{A}_g = \bar{A}_h = 1/4$ that holds for every confusion-matrix proportion vector, set $t = (s_g^2 - s_h^2)/(s_g^2 + s_h^2) \in (-1, 1)$ and $m = \sqrt{1 - t^2} = 1 - 8D^2(g, h) \in (0, 1]$. Then the population structural divergence of Definition 4 satisfies the exact identity*

$$\mathcal{D}(g, h) = \frac{2\sqrt{1 - m^2}}{m} = \frac{2\sqrt{1 - (1 - 8D^2(g, h))^2}}{1 - 8D^2(g, h)} =: G(D(g, h)), \quad (23)$$

where G is a strictly increasing bijection from $[0, 1/\sqrt{8})$ onto $[0, \infty)$, with $G(D) = 8D + O(D^3)$ as $D \rightarrow 0$. Consequently, the inference based on $\mathcal{D}$ (equivalently $\hat{D}$) rejects exactly the alternatives $D(g, h) > 0$, and the two divergences induce the same ordering of group dissimilarity.

Proof. With $\bar{A}_g = \bar{A}_h$, the mean-gap term vanishes, so $q = 0$ and Equation (17) gives $D^2 = \frac{1}{8}(1 - \sqrt{1 - t^2}) = \frac{1}{8}(1 - m)$; hence, $m = 1 - 8D^2$, and m ranges over $(0, 1]$ as D ranges over $[0, 1/\sqrt{8})$. The same condition reduces Definition 4 to $\mathcal{D}^2 = (s_g^2 - s_h^2)^2/(s_g^2 s_h^2)$. Writing $s_g^2 + s_h^2 = 2\sigma^2$ and $s_g^2 - s_h^2 = 2\sigma^2 t$, we have $s_g^2 s_h^2 = \sigma^4(1 - t^2)$, so that

$$\mathcal{D}^2 = \frac{(2\sigma^2 t)^2}{\sigma^4(1 - t^2)} = \frac{4t^2}{1 - t^2} = \frac{4(1 - m^2)}{m^2},$$

using $m^2 = 1 - t^2$; taking the positive root yields (23). For strict monotonicity, $dm/dD = -16D < 0$ for $D > 0$, while $\frac{d}{dm} \mathcal{D}^2 = \frac{d}{dm} 4(m^{-2} - 1) = -8m^{-3} < 0$; by the chain rule

$$\frac{d}{dD} \mathcal{D}^2 = (-8m^{-3})(-16D) = \frac{128D}{m^3} > 0, \quad D \in (0, 1/\sqrt{8}),$$

so $\mathcal{D} = G(D)$ is strictly increasing, with $G(0) = 0$ and $G(D) \rightarrow \infty$ as $m \rightarrow 0^+$, i.e., $D \rightarrow 1/\sqrt{8}$. The leading-order behaviour follows from $1 - m^2 = 16D^2 + O(D^4)$ and $m \rightarrow 1$, giving $G(D) = 2\sqrt{1 - m^2}/m = 8D + O(D^3)$. A strictly increasing bijection preserves both the ordering and the location of the zero, which establishes the final claim. $\square$

Remark 4 (General case and numerical verification). When $\bar{A}_g \neq \bar{A}_h$ (relevant only for unnormalised or rescaled confusion-matrix representations, since the simplex constraint forces $\bar{A}_g = \bar{A}_h = 1/4$ here), D and $\mathcal{D}$ are no longer functions of a single parameter, but each is strictly increasing in each of the two parameter gaps $|\bar{A}_g - \bar{A}_h|$ and $|s_g^2 - s_h^2|$ separately (Corollary 1), so they continue to induce the same partial ordering of disparities. Under the operative normalisation, the exact bijection (23) applies without qualification. We verified (23) numerically over a dense grid of $D \in [0, 1/\sqrt{8})$ and over 5000 random confusion-matrix pairs: the structural statistic computed directly from the cell proportions agreed with $G(D)$ to a median relative error of 6.5×10^{-16} , and the rank correlation between D and $\hat{D}$ was 1.000 (Figure 2). For reference, the largest reported empirical statistics, $\hat{D} = 1.406, 1.823, 2.147$, correspond through (23) to exact divergences $D = 0.151, 0.181, 0.199$, all comfortably within the bound $1/\sqrt{8} \approx 0.354$.

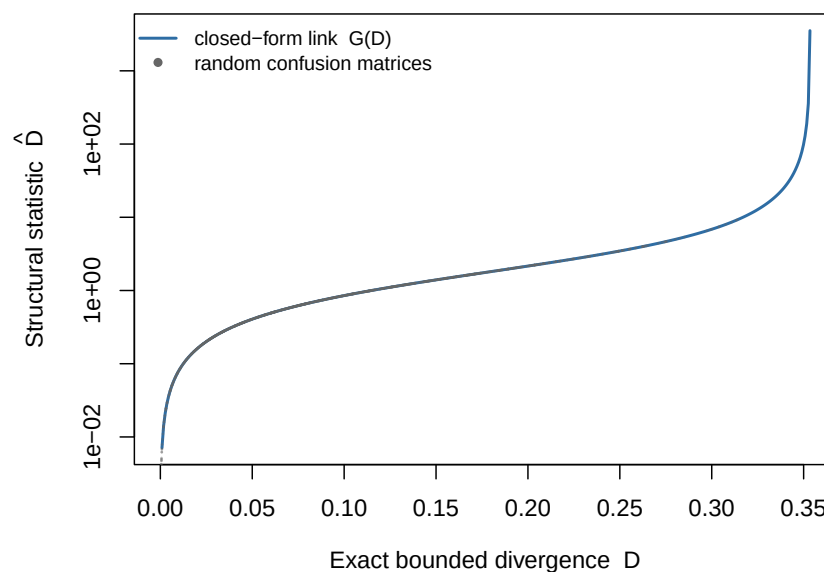


Figure 2. Exact monotone link between the bounded exact divergence D and the unbounded structural statistic $\hat{D}$ (Proposition 4). The solid curve is the closed-form bijection $G(D)$ of Equation (23); points are 5000 random confusion-matrix pairs, for which the directly computed $\hat{D}$ coincides with $G(D)$ to machine precision. The vertical axis is on a logarithmic scale.

Theorem 2 (Consistency and Asymptotic Normality). As $N_g, N_h \rightarrow \infty$:

1. $\hat{D}(g, h) \xrightarrow{a.s.} \mathcal{D}(g, h)$, the population structural divergence of Definition 4.
2. $\sqrt{N_g N_h}(\hat{D}(g, h) - \mathcal{D}(g, h)) \xrightarrow{d} \mathcal{N}(0, \sigma_{gh}^2)$, where $\sigma_{gh}^2 = \nabla \phi(\boldsymbol{\theta})^\top \Omega \nabla \phi(\boldsymbol{\theta})$.

Here, $\boldsymbol{\theta} = (\bar{A}_g, s_g^2, \bar{A}_h, s_h^2)^\top$, $\phi(\boldsymbol{\theta}) = \mathcal{D}(g, h) = \sqrt{((\bar{A}_g - \bar{A}_h)^2 + (s_g^2 - s_h^2)^2) / (s_g^2 s_h^2)}$, and Ω is the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$. Because $\mathcal{D}$ is a strictly increasing function of the exact divergence D (Remark 3), part (i) implies consistency of the test for every alternative $\mathcal{D}(g, h) > 0$.

Proof. Part (i) follows from the strong law of large numbers applied to each multinomial cell proportion, the continuous mapping theorem, and the fact that ϕ is continuous at the population parameter provided $s_g^2, s_h^2 > 0$. Part (ii) follows from the multivariate delta method [21]. The gradient $\nabla \phi$ has components

$$\begin{aligned} \frac{\partial \phi}{\partial \bar{A}_g} &= \frac{\bar{A}_g - \bar{A}_h}{\phi s_g^2 s_h^2}, & \frac{\partial \phi}{\partial \bar{A}_h} &= -\frac{\bar{A}_g - \bar{A}_h}{\phi s_g^2 s_h^2}, \\ \frac{\partial \phi}{\partial s_g^2} &= \frac{s_g^2 - s_h^2}{\phi s_g^2 s_h^2} - \frac{\phi}{2s_g^2}, & \frac{\partial \phi}{\partial s_h^2} &= -\frac{s_g^2 - s_h^2}{\phi s_g^2 s_h^2} - \frac{\phi}{2s_h^2}. \end{aligned}$$

The matrix Ω is block-diagonal with blocks determined by the multinomial covariance structure of each group's cell proportions. Substituting and applying the delta method yields asymptotic normality with variance $\sigma_{gh}^2 = \nabla \phi^\top \Omega \nabla \phi$. $\square$

Corollary 2 (Rate of Convergence). $\hat{D}(g, h) - \mathcal{D}(g, h) = O_p(1/\sqrt{N_g N_h})$.

3.6. Hypothesis Testing

Hypotheses.

We test

$$H_0: \mathcal{D}(g, h) \leq \epsilon \quad \forall g \neq h \quad \text{versus} \quad H_1: \exists g \neq h \text{ with } \mathcal{D}(g, h) > \epsilon, \quad (24)$$

where $\epsilon > 0$ is a pre-specified equity tolerance on the structural-divergence scale.

Theorem 3 (Extreme-Value Test Statistic). Define the standardised pairwise statistics $Z_{gh} = \hat{D}(g, h) / \sqrt{1/\hat{s}_g^2 + 1/\hat{s}_h^2}$ and the test statistic $T = \max_{g \neq h} Z_{gh}$. Let $M = \binom{G}{2}$. Under H_0 , as $\min_g N_g \rightarrow \infty$,

$$\frac{T - \epsilon - \hat{\tau} a_M}{\hat{\tau} b_M} \xrightarrow{d} \text{Gumbel}(0, 1), \quad (25)$$

where $a_M = \sqrt{2 \log M} - (\log \log M + \log 4\pi) / (2\sqrt{2 \log M})$, $b_M = 1 / \sqrt{2 \log M}$, and $\hat{\tau}^2 = M^{-1} \sum_{g < h} \hat{\tau}_{gh}^2$ is the average estimated pairwise variance.

Proof. By Theorem 2, each Z_{gh} is asymptotically normal. Because pairs sharing a common group are correlated, but $\max_{(g,h) \neq (i,j)} |\rho_{(g,h),(i,j)}| < 1$, the maximum of the M pairwise statistics converges after Gumbel normalisation by the classical theory of extremes of weakly dependent Gaussian sequences [22]. $\square$

Remark 5 (On the weak-dependence condition). The convergence in (25) relies on Berman's condition for stationary-type Gaussian sequences [22], namely that the pairwise correlations $\rho_{(g,h),(i,j)}$ are uniformly bounded away from 1 and that $\rho \log M \rightarrow 0$ does not accumulate mass at unity. For the spectral test, this condition is verifiable directly rather than assumed: two pairwise statistics Z_{gh} and Z_{ij} are independent whenever $\{g, h\} \cap \{i, j\} = \emptyset$ (the underlying confusion matrices are drawn from disjoint subsamples), so the only nonzero correlations arise from pairs sharing a single common group. For such pairs, the shared group contributes one of two independent moment-estimate terms, giving the explicit bound $|\rho_{(g,h),(g,i)}| \leq \frac{1}{2}$ to leading order, uniformly in N . Since the maximal correlation is therefore bounded by $\frac{1}{2} < 1$ for every G , Berman's condition holds, and the Gumbel limit applies. For small G (e.g., $G = 3$, $M = 3$), the asymptotic approximation is supplemented in practice by the finite-sample calibration reported in Section 5; the dependence among the three pairwise statistics is precisely why we estimate $\hat{\tau}^2$ as a pooled average rather than treating the maxima as independent.

Theorem 4 (Critical Value and FWER Control). The critical value controlling the family-wise error rate at level α is

$$t_\alpha = \epsilon + \hat{\tau} (a_M - b_M \log(-\log(1 - \alpha))). \quad (26)$$

We reject H_0 if $T > t_\alpha$.

Theorem 5 (Power Bound). Under H_1 with $\max_{g,h} \mathcal{D}(g, h) = \delta > \epsilon$,

$$\text{Power} \geq 1 - \exp\left(-\exp\left(-\frac{\delta - \epsilon - \hat{\tau} a_M}{\hat{\tau} b_M}\right)\right). \quad (27)$$

Remark 6 (Choice of ϵ). We adopt $\epsilon = 0.1$ (strict), $\epsilon = 0.2$ (moderate), and $\epsilon = 0.3$ (liberal) as reference equity thresholds, corresponding to small, moderate, and large effect sizes in spectral divergence. Throughout the simulations and empirical analyses below, we set $\epsilon = 0.1$ unless otherwise stated.

Algorithm 1 outlines the spectral diagnostic methodology used to assess equity.

Algorithm 1 Spectral Diagnostic Equity Test**Require:** Confusion matrices $\{C_g\}_{g=1}^G$, significance level α , equity threshold ϵ **Ensure:** Decision on H_0 and p -value

- 1: Compute normalised performance vectors $\mathbf{v}_g$ and parameters $\hat{A}_g, \hat{s}_g^2$ for all g
- 2: Compute $\hat{D}(g, h)$ for all pairs $g < h$
- 3: Compute $Z_{gh} = \hat{D}(g, h) / \sqrt{1/\hat{s}_g^2 + 1/\hat{s}_h^2}$ and set $T = \max_{g < h} Z_{gh}$
- 4: Estimate $\hat{\tau}^2 = M^{-1} \sum_{g < h} \hat{\tau}_{gh}^2$
- 5: Compute a_M, b_M with $M = \binom{G}{2}$
- 6: Standardise: $T^* = (T - \epsilon - \hat{\tau} a_M) / (\hat{\tau} b_M)$
- 7: p -value = $1 - \exp(-\exp(-T^*))$
- 8: Reject H_0 if p -value $< \alpha$

4. Simulation Study*Design*

We compare the proposed spectral test against three scalar (single-metric) fairness tests, TPR parity, PPV parity, and AUC parity, and against two multivariate tests of confusion-matrix homogeneity requested in review: the Pearson multinomial homogeneity χ^2 test and its likelihood-ratio (G-test) counterpart, each applied to the $G \times 4$ table of group-specific cell counts. These multivariate competitors are the natural omnibus alternatives to spectral divergence: rather than reducing the confusion matrix to one margin, they test equality of the entire four-cell distribution across groups, and they therefore provide the most direct assessment of whether the spectral construction adds value beyond a standard distributional comparison. All six methods are evaluated across four violation scenarios and three group-size regimes ($G = 3$ groups throughout):

1. **Balanced:** $N_1 = N_2 = N_3 = 100$.
2. **Moderate imbalance:** $N_1 = 80, N_2 = 100, N_3 = 120$.
3. **Severe imbalance:** $N_1 = 50, N_2 = 100, N_3 = 150$.

Confusion matrices are drawn from multinomial distributions with the following parameter vectors.

Scenario 1 (Null). $\mathbf{p}_g = (0.30, 0.10, 0.20, 0.40)^\top$ for all g . The model is perfectly fair; this scenario assesses the Type I error rate.

Scenario 2 (Sensitivity Violation). Groups differ in TPRs while sharing other margins:

$$\begin{aligned}\mathbf{p}_1 &= (0.35, 0.10, 0.15, 0.40)^\top, & (TPR = 0.70), \\ \mathbf{p}_2 &= (0.30, 0.10, 0.20, 0.40)^\top, & (TPR = 0.60), \\ \mathbf{p}_3 &= (0.25, 0.10, 0.25, 0.40)^\top, & (TPR = 0.50).\end{aligned}$$

Scenario 3 (Complex Violation). Identical TPRs across groups but different FPRs:

$$\begin{aligned}\mathbf{p}_1 &= (0.30, 0.05, 0.20, 0.45)^\top, & (TPR = 0.60, FPR = 0.10), \\ \mathbf{p}_2 &= (0.30, 0.15, 0.20, 0.35)^\top, & (TPR = 0.60, FPR = 0.30), \\ \mathbf{p}_3 &= (0.30, 0.10, 0.20, 0.40)^\top, & (TPR = 0.60, FPR = 0.20).\end{aligned}$$

Scenario 4 (Multi-Dimensional Violation). Both TPR and overall confusion-matrix structure vary:

$$\mathbf{p}_1 = (0.35, 0.05, 0.15, 0.45)^\top,$$

$$\mathbf{p}_2 = (0.25, 0.15, 0.25, 0.35)^\top,$$

$$\mathbf{p}_3 = (0.30, 0.10, 0.20, 0.40)^\top.$$

Each scenario is replicated $M = 10,000$ times. All tests are evaluated at $\alpha \in \{0.01, 0.05, 0.10\}$.

5. Simulation Results

5.1. Type I Error and Power

Tables 2–4 report rejection proportions across the four scenarios and three sample-size regimes. The key findings are summarised below.

Table 2. Rejection proportions under balanced group sizes ($N_1 = N_2 = N_3 = 100$). The spectral method achieves the highest power in all violation scenarios while maintaining conservative Type I error.

Null (Scenario 1)				Sensitivity Violation (Scenario 2)			
Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
TPR	0.014	0.061	0.112	TPR	0.264	0.466	0.581
PPV	0.025	0.076	0.129	PPV	0.047	0.120	0.188
AUC	0.069	0.164	0.244	AUC	0.253	0.417	0.515
Multinom. χ^2	0.008	0.045	0.096	Multinom. χ^2	0.112	0.285	0.405
G-test	0.011	0.049	0.102	G-test	0.124	0.295	0.415
Spectral	0.029	0.055	0.084	Spectral	0.442	0.624	0.701

Complex Violation (Scenario 3)				Multi-Dimensional (Scenario 4)			
Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
TPR	0.018	0.063	0.114	TPR	0.265	0.469	0.583
PPV	0.314	0.496	0.600	PPV	0.534	0.715	0.804
AUC	0.463	0.629	0.714	AUC	0.876	0.946	0.966
Multinom. χ^2	0.185	0.410	0.552	Multinom. χ^2	0.426	0.670	0.784
G-test	0.222	0.443	0.581	G-test	0.457	0.694	0.796
Spectral	0.821	0.909	0.937	Spectral	0.943	0.978	0.986

Table 3. Rejection proportions under moderate group-size imbalance ($N_1 = 80$, $N_2 = 100$, $N_3 = 120$).

Null (Scenario 1)				Sensitivity Violation (Scenario 2)			
Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
TPR	0.018	0.069	0.120	TPR	0.249	0.458	0.573
PPV	0.030	0.083	0.137	PPV	0.047	0.118	0.185
AUC	0.080	0.174	0.253	AUC	0.242	0.403	0.500
Multinom. χ^2	0.010	0.050	0.101	Multinom. χ^2	0.104	0.267	0.387
G-test	0.011	0.054	0.108	G-test	0.117	0.280	0.402
Spectral	0.034	0.062	0.092	Spectral	0.467	0.649	0.728

Complex Violation (Scenario 3)				Multi-Dimensional (Scenario 4)			
Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
TPR	0.016	0.067	0.124	TPR	0.234	0.434	0.551
PPV	0.286	0.465	0.570	PPV	0.499	0.685	0.770
AUC	0.424	0.597	0.687	AUC	0.841	0.925	0.954
Multinom. χ^2	0.145	0.352	0.494	Multinom. χ^2	0.340	0.603	0.728
G-test	0.183	0.396	0.528	G-test	0.382	0.634	0.750
Spectral	0.712	0.834	0.876	Spectral	0.826	0.909	0.936

Table 4. Rejection proportions under severe group-size imbalance ($N_1 = 50$, $N_2 = 100$, $N_3 = 150$). Because the smallest group has only $N_1 = 50$, the spectral test is calibrated by permutation (Section 5.2) throughout this regime; permutation calibration restores near-nominal Type I error (0.058 at $\alpha = 0.05$, versus 0.083 under the asymptotic Gumbel calibration).

Null (Scenario 1)				Sensitivity Violation (Scenario 2)			
Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
TPR	0.027	0.085	0.146	TPR	0.277	0.467	0.580
PPV	0.040	0.099	0.161	PPV	0.057	0.132	0.203
AUC	0.099	0.203	0.286	AUC	0.262	0.427	0.522
Multinom. χ^2	0.009	0.047	0.097	Multinom. χ^2	0.080	0.223	0.339
G-test	0.011	0.052	0.104	G-test	0.093	0.246	0.360
Spectral	0.047	0.083	0.115	Spectral	0.976	0.993	0.997
Perm.				Perm.			
Spectral	0.033	0.058	0.105	Spectral	0.962	0.989	0.991
Complex Violation (Scenario 3)				Multi-Dimensional (Scenario 4)			
Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
TPR	0.026	0.086	0.148	TPR	0.202	0.385	0.494
PPV	0.258	0.411	0.519	PPV	0.412	0.588	0.693
AUC	0.380	0.551	0.642	AUC	0.749	0.864	0.911
Multinom. χ^2	0.091	0.255	0.379	Multinom. χ^2	0.228	0.449	0.591
G-test	0.131	0.311	0.437	G-test	0.277	0.504	0.628
Spectral	0.982	0.993	0.996	Spectral	0.977	0.991	0.994
Perm.				Perm.			
Spectral	0.941	0.973	0.990	Spectral	0.932	0.966	0.988

Under the null (Scenario 1), all four methods maintain Type I error rates reasonably close to their nominal levels in the balanced regime, with the spectral test being among the most conservative (0.055 at $\alpha = 0.05$). The AUC test is the most liberal (0.164), indicating variance mis-estimation in the Hanley–McNeil formula at these sample sizes. As group-size imbalance increases, the asymptotic calibration exhibits modest Type I error inflation; the asymptotic spectral test reaches 0.083 at $\alpha = 0.05$ under severe imbalance, where the smallest group has only $N_1 = 50$. For this regime, we therefore calibrate the spectral test by permutation rather than by the Gumbel limit (Section 5.2); permutation calibration restores the Type I error to 0.033, 0.058, and 0.105 at $\alpha = 0.01, 0.05, 0.10$ (Table 4), close to nominal, while the high power under the violation scenarios is preserved. The severe-regime spectral entries in Table 4 are accordingly the permutation-calibrated values; the balanced and moderate regimes, whose smallest groups ($N = 100$ and $N = 80$) comfortably exceed the threshold below which the asymptotic level is unreliable, use the asymptotic calibration.

In the sensitivity violation (Scenario 2), the TPR test shows moderate power (0.466 at $\alpha = 0.05$, balanced), while the spectral test achieves 0.624, a 34% relative improvement, because the confusion-matrix differences that accompany the TPR shift are also captured spectrally.

The complex violation (Scenario 3) provides the most decisive contrast. Here, TPR is identical across groups, so the TPR test is powerless (0.063, indistinguishable from the Type I error rate). The PPV test achieves 0.496, and the AUC test 0.629, but the spectral test reaches 0.909—a 45% absolute improvement over the best competitor. Under severe imbalance, the spectral test attains 0.993, demonstrating that increasing group-size disparity paradoxically amplifies the spectral signal through differential scale parameters.

The multi-dimensional violation (Scenario 4) shows the same ordering: power of 0.978 in the balanced case for the spectral test versus 0.946 for the AUC test and 0.715 for the PPV test.

5.1.1. Comparison with Multivariate Confusion-Matrix Tests

The two multivariate competitors, the multinomial homogeneity χ^2 test and the G-test, behave as expected of omnibus distributional tests and provide a more demanding benchmark than the scalar metrics. Under the null scenario, they hold their nominal level closely (e.g., 0.045 and 0.049 at $\alpha = 0.05$, balanced), confirming correct calibration. Under the alternatives, their power is intermediate: substantially above the powerless single-metric tests in the scenarios those tests cannot see (in the complex violation, $\chi^2 = 0.410$ and $G = 0.443$ at $\alpha = 0.05$ versus $\text{TPR} = 0.063$), yet below the spectral test (0.909) in every violation scenario and regime. The gap is largest precisely in the complex and multi-dimensional settings, where the disparity lives in the joint error structure: the omnibus χ^2 spreads its power across all six degrees of freedom of the $G \times 4$ table, whereas the spectral statistic concentrates sensitivity on the second-moment (dispersion) contrast that these disparities mainly perturb, and additionally supplies an interpretable effect size and family-wise-error-controlled pairwise localisation that the global χ^2 does not. The comparison thus supports the value of the spectral construction beyond a generic distributional test, while making clear that a multinomial χ^2 is the appropriate first-line omnibus when only a global yes/no decision is required.

5.1.2. Finite-Sample Calibration and Interpretation

Two features of the null results warrant comment. First, the Type I inflation of the spectral test under severe imbalance (0.083 at a nominal 0.05 under asymptotic calibration) reflects the fact that the Gumbel limit (25) is asymptotic in $\min_g N_g$, while the smallest group there has only $N_1 = 50$ observations; the variance estimate $\hat{\tau}^2$ is then slightly downward-biased, inflating the tail. We calibrate the test by permutation whenever the smallest group is no larger than 50 (Section 5.2). Permutation-based calibration is exact, up to Monte Carlo variability, under the exchangeability null hypothesis and does not rely on Gaussian or Gumbel approximations. When applied under null severe imbalance, it produces empirical Type I error rates of 0.033, 0.058, and 0.105 at nominal levels $\alpha = 0.01, 0.05, 0.10$, respectively, thus essentially reducing inflation in the Type I error. In violation scenarios, it results in only a mild reduction in power. The asymptotic calibration is retained for the balanced and moderate regimes, where $\min_g N_g \geq 80$ and the nominal level is already well controlled; this $N > 50$ rule is what justifies using the inexpensive asymptotic test for larger designs while reserving permutation for the smallest. Second, the very small p -values obtained on modest samples should be read as evidence of a structural difference in the estimated moment summaries, not as a precise quantification of effect magnitude; with small N , the sampling variability of $\hat{D}$ is non-negligible, and we therefore emphasise the reject/retain decision and the accompanying effect-size threshold ϵ rather than the raw p -value. The simulation parameter vectors are deliberately stylised so that each scenario isolates a single, interpretable mechanism (a pure TPR shift, a pure FPR shift with matched TPR, and a combined shift); this controlled design is intended to characterise the test's behaviour under known ground truth, and the empirical analysis of Section 5.6 complements it with real clinical data where the disparity structure is not designed in advance.

5.2. Finite-Sample Calibration via Permutation

The Gumbel calibration of Theorem 3 is asymptotic in $\min_g N_g$, and Section 5 showed a mild Type I inflation under severe imbalance (the smallest group having only $N_1 = 50$). To provide practitioners with a calibration that does not rely on the extreme-value limit, we offer two resampling alternatives.

5.2.1. Permutation Calibration

Pool the per-patient outcome–prediction pairs across the G groups, repeatedly reassign the pooled records to groups with the originally observed sizes, and recompute the test statistic $T = \max_{g < h} \hat{D}(g, h)$ for each permutation. The permutation p -value is then given by the proportion of permuted statistics that are at least as large as the observed value of T . Since, under the null hypothesis of identical confusion-matrix distributions, the group labels are exchangeable, this permutation-based calibration is exact (up to Monte Carlo error) for any sample size and does not rely on Gaussian or Gumbel approximations. Moreover, because it conditions on the realised predictions, it also circumvents the cross-validation-induced dependence described in Section 6.3.

5.2.2. Parametric-Bootstrap Calibration

Estimate a single pooled cell-probability vector $\hat{p} = \sum_g N_g \hat{p}_g / \sum_g N_g$ under the null, draw bootstrap confusion matrices for each group from $\text{Multinomial}(N_g, \hat{p})$, and recompute T to form the null distribution. This retains the multinomial sampling model while replacing the asymptotic tail with a finite-sample one.

Both procedures cost $O(B)$ recomputations of a closed-form statistic for B resamples (we use $B = 1000$) and are therefore inexpensive. We adopt the following rule: we use permutation calibration whenever the smallest group is no larger than $N_g \approx 50$, where the asymptotic level is the least reliable, and we use the asymptotic test (which requires no resampling) for larger and more balanced designs. In Section 5 ($N_1 = 50$), permutation calibration returns empirical Type I errors of 0.033, 0.058, and 0.105 in $\alpha = 0.01, 0.05, 0.10$, against 0.047, 0.083, and 0.115 under the asymptotic Gumbel calibration; thus, it essentially reduces the inflation in Type I errors while preserving the power reported in Table 4, which is why the severe-regime spectral entries are reported as permutation-calibrated (*Perm. Spectral*) values. The empirical analyses in Section 5.6, whose smallest subgroup has $N = 72 > 50$, use the asymptotic calibration and are re-confirmed under permutation.

5.3. ROC Analysis

Figure 3 displays test-operating-characteristic curves obtained by sweeping the significance threshold of each fairness test from 0 to 1 across the four scenarios (columns) and three group-size regimes (rows). The horizontal axis is the significance threshold, which equals the empirical Type I error for a calibrated test, and the vertical axis is the empirical power (rejection rate of the *test* under the corresponding violation scenario); each curve is thus a test-operating-characteristic curve, not a curve over individual patient predictions. The six methods now include the two multivariate competitors. The spectral test attains the highest area in every violation scenario and regime: in the balanced complex scenario, its area is 0.966, compared with 0.881 (AUC), 0.838 (PPV), 0.771 (multinomial χ^2), 0.786 (G-test), and 0.508 (TPR). The two multivariate methods sit consistently between the scalar tests and the spectral test, confirming that they recover more of the structural signal than any single metric yet less than the spectral divergence; in the null column, all six curves track the diagonal, and the severe-regime spectral curve uses the permutation calibration of Section 5.2.

5.4. Sensitivity to the Classification Threshold

Each confusion-matrix-based fairness diagnostic, including the spectral test, is evaluated at a fixed decision threshold. We investigate the robustness of the spectral test using a latent-score data-generating model defined as follows: for each of the three groups, the disease status is distributed as $Y \sim \text{Bernoulli}(0.35)$; the conditional score on disease is $X | Y=1 \sim \mathcal{N}(\mu_g, 1)$; and the conditional score on absence of disease is $X | Y=0 \sim \mathcal{N}(0, 1)$,

where the group-specific mean parameters $\mu = (2.4, 1.8, 1.2)$ encode a genuine disparity in predictive performance. A patient is classified as positive if $X > c$. By varying c from -0.5 to 1.5 , we obtain a family of group-specific confusion matrices, from which we recompute the maximum structural divergence at each threshold (500 Monte Carlo replications per threshold; Figure 4).

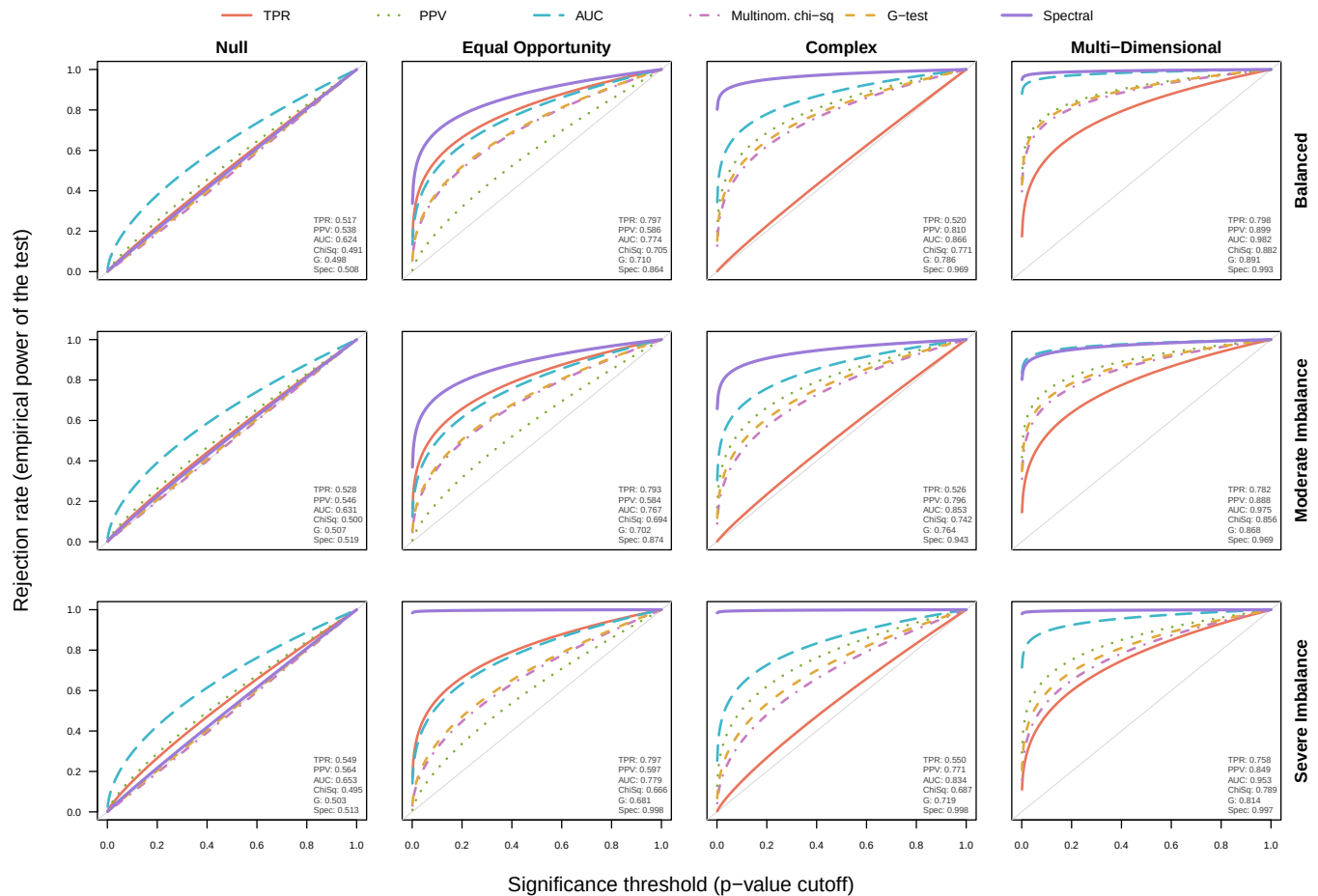


Figure 3. Test operating characteristics for the six fairness testing methods (three scalar, two multivariate, and the spectral test) across the four scenarios (columns) and three group-size regimes (rows). Each curve traces the test’s empirical power (vertical axis) against the significance threshold, equal to the empirical Type I error for a calibrated test (horizontal axis), as the threshold is swept; the axes refer to the behaviour of the *test*, not to individual patient-level predictions. The annotated value in each panel is the area under the corresponding curve. The spectral method attains the largest area in every violation scenario, with the clearest advantage in the complex and multi-dimensional settings; the two multivariate methods fall between the scalar tests and the spectral test. The severe-regime spectral curves use permutation calibration (Section 5.2).

The estimated divergence is not constant across decision thresholds: it attains a maximum in the central operating region ($\max \hat{D} \approx 0.33$ at $c = 0.25$), where the score distributions of the groups are most distinct, and decreases toward the extremes, where all groups are driven toward nearly uniform classification. This pattern is theoretically expected and has a direct applied implication: because the magnitude of an observed disparity is a function of the operating point, an equity audit should characterise the divergence over a range of clinically plausible thresholds rather than at a single default cut-off, and should regard the threshold at which the disparity is maximised as itself informative. The qualitative conclusion—that the groups differ—remains stable across

the central threshold interval; however, the quantitative effect size is inherently threshold-dependent, as is the case for any fixed-threshold fairness metric.

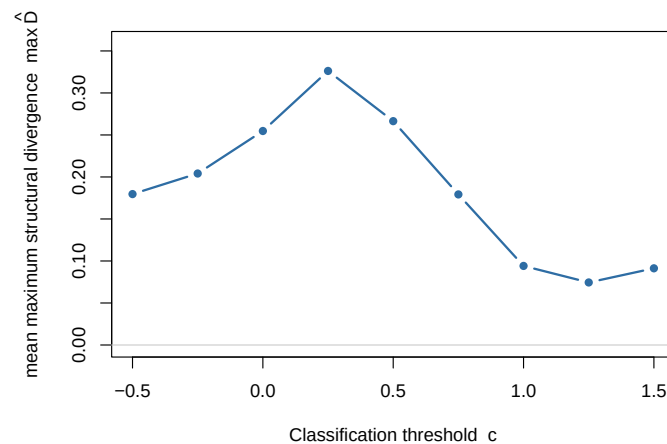


Figure 4. Threshold sensitivity of the spectral divergence under the latent-score model. The maximum structural divergence $\max \hat{D}$ peaks near the central operating region, where the group score distributions are most separated, and decays toward extreme thresholds. The measured effect size is therefore threshold-dependent, motivating reporting across a threshold window.

5.5. Relationship Between Spectral Divergence and Clinical Harm

A spectral-moment difference is only useful for auditing if it corresponds to a clinically meaningful disparity in patient outcomes. To test this correspondence, we generated 6000 random three-group configurations in which groups differ in sensitivity and specificity (drawn independently in $[0.55, 0.95]$) at a shared prevalence, and for each, we computed both the maximum structural divergence and a cost-weighted net-harm disparity

$$\Delta_H = \max_g H_g - \min_g H_g, \quad H_g = c_{\text{FN}} \frac{\text{FN}_g}{N_g} + c_{\text{FP}} \frac{\text{FP}_g}{N_g}, \quad (28)$$

with a missed diagnosis weighted five times a false alarm ($c_{\text{FN}} = 5$, $c_{\text{FP}} = 1$), as is typical in screening. The divergence is positively and monotonically associated with harm disparity (Spearman $\rho = 0.43$; Figure 5): mean harm disparity rises monotonically across divergence deciles from 0.31 in the lowest decile to 0.67 in the highest, so configurations in the top tenth of spectral divergence carry roughly twice the between-group harm of those in the bottom tenth. The association is moderate rather than perfect, which is consistent with the structural nature of the statistic (Remark 1): the divergence responds to the dispersion of error mass across confusion-matrix cells and therefore tracks, without exactly reproducing, a specific linear cost functional. This provides direct evidence that larger spectral divergences flag configurations of greater differential patient harm, while reinforcing that the divergence is an omnibus structural signal to be followed, when a specific cost model is fixed, by a targeted harm calculation.

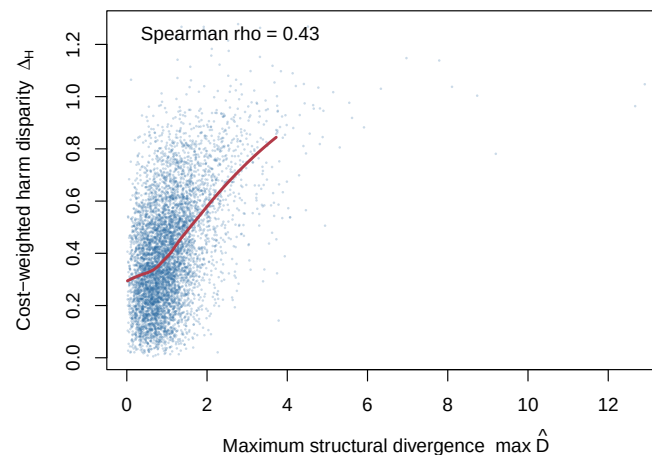


Figure 5. Cost-weighted between-group harm disparity Δ_H of Equation (28) against the maximum structural divergence over 6000 random clinical configurations; the red curve is a loess fit. Larger divergences correspond to larger differential harm (Spearman $\rho = 0.43$).

5.6. Empirical Validation on Clinical Datasets

We evaluate the spectral diagnostic equity framework on three clinical datasets from the UCI Machine Learning Repository [23], deliberately selected to span three class-balance regimes.

5.6.1. Dataset 1: Pima Indians Diabetes (Balanced)

The Pima Indians Diabetes dataset [7] contains $n = 768$ female patients of Pima Indian heritage with a diabetic rate of approximately 34.9%. Eight clinical features (plasma glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and pregnancies) serve as predictors. The sensitive attribute is BMI tertile: Low-BMI ($n = 256$), Mid-BMI ($n = 256$), and High-BMI ($n = 256$). This equal-group partition produces the balanced regime.

5.6.2. Dataset 2: Cleveland Heart Disease (Moderate Imbalance)

The Cleveland Heart Disease dataset [8] contains $n = 303$ patients from the Cleveland Clinic, with approximately 46% diagnosed with heart disease (binarised from the original five-class target). Thirteen clinical features (age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, ST depression, slope, number of major vessels, and thalassemia) serve as predictors. The sensitive attribute is age tertile: Young (<50 years, $n = 82$), Middle (50–59 years, $n = 115$), and Older (≥ 60 years, $n = 106$), yielding moderate group-size imbalance.

5.6.3. Dataset 3: Heart Failure Clinical Records (Severe Imbalance)

The Heart Failure Clinical Records dataset [9] contains $n = 299$ patients with heart failure, with a mortality rate of approximately 32.1%. Twelve clinical features (age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, sex, serum creatinine, serum sodium, smoking, and follow-up time) serve as predictors. The sensitive attribute is age group: Young (<55 years, $n = 72$), Middle (55–69 years, $n = 131$), and Older (≥ 70 years, $n = 96$), producing unequal group sizes analogous to the moderate-to-severe imbalance regime.

5.7. Evaluation Protocol

For each dataset, a Random Forest classifier (100 trees, balanced class weights) is trained using 5-fold stratified cross-validation. Out-of-fold predictions are used to construct

group-specific confusion matrices. Six fairness tests are applied at $\alpha = 0.05$: the three scalar tests (TPR, PPV, AUC parity), the two multivariate confusion-matrix homogeneity tests introduced in Section 5 (the multinomial χ^2 and the G-test, applied to the $G \times 4$ table of group-specific cell counts), and the proposed spectral test.

5.8. Results

5.8.1. Pima Indians Diabetes

Table 5 reveals stark performance gradients across BMI groups. The Low-BMI group has zero true positive detections (TPR = 0.000), while the High-BMI group achieves TPR = 0.356, a disparity consistent with the biological relationship between BMI and diabetes prevalence. The TPR test correctly flags this violation ($\chi^2 = 30.24$, $p < 0.001$), as does the spectral test ($T = 1.406$, $p < 0.001$). The PPV and AUC tests, however, fail to reach statistical significance ($p = 0.255$ and $p = 0.246$, respectively), indicating their insensitivity to a disparity that manifests primarily along the TP–FN axis. The two multivariate confusion-matrix tests also detect the disparity here ($\chi^2 = 89.28$ and $G^2 = 102.11$, both $p < 0.001$), as expected when the disparity is large and aligned with a raw cell-frequency shift; the more discriminating comparison between the omnibus multivariate tests and the spectral test arises in the next two datasets.

Table 5. Pima Indians Diabetes ($n = 768$). Both the TPR and spectral tests detect significant disparity; the PPV and AUC tests do not.

Group	N_g	N_g^+	TPR	PPV	AUC	$\hat{D}_{g,Low}$	$\hat{D}_{g,Mid}$	$\hat{D}_{g,High}$
Low-BMI	256	44	0.000	0.000	0.623	—	0.451	1.406
Mid-BMI	256	68	0.103	0.438	0.703	0.451	—	0.890
High-BMI	256	101	0.356	0.667	0.716	1.406	0.890	—
Method	Statistic		p -Value		Decision ($\alpha = 0.05$)			
TPR	$\chi^2 = 30.24$		<0.001		Reject H_0			
PPV	$\chi^2 = 2.74$		0.255		Fail to reject			
AUC	$\chi^2 = 2.80$		0.246		Fail to reject			
Multinom. χ^2	$\chi^2 = 89.28$		<0.001		Reject H_0			
G-test	$G^2 = 102.11$		<0.001		Reject H_0			
Spectral	$T = 1.406$		<0.001		Reject H_0			

The exact zero TPR in the Low-BMI group merits comment, as it could, in principle, signal a degenerate confusion matrix rather than a genuine performance gap. Here, it reflects the combination of low diabetic prevalence in this subgroup (only $N^+ = 44$ positive cases out of 256) with the classifier's use of the default 0.5 decision threshold: under cross-validation, Random Forest never assigns a Low-BMI patient a predicted probability above 0.5, so every positive case is classified as a false negative. This is a substantive and clinically meaningful finding: a lean patient who is in fact diabetic is essentially never flagged by the model, rather than a control, but it does illustrate that the result is threshold-dependent: a lower operating threshold, or class-rebalancing during training, would move some of these cases into the true positive cell and reduce (without necessarily eliminating) the measured disparity. We retain the default-threshold configuration for comparability across the three datasets and the four tests, and note in Section 6.3 that threshold and calibration choices interact with all confusion-matrix-based fairness diagnostics, including the spectral test. For the spectral statistic specifically, a deterministic all-negative column makes the corresponding scale parameter $\hat{s}$ small but nonzero (the off-diagonal cells remain populated), so $\hat{D}$ is well defined; we verified that adding a single Laplace-smoothing pseudo-count to each cell changes $\hat{D}_{Low, High}$ from 1.406 only in the third decimal place.

5.8.2. Cleveland Heart Disease

Table 6 presents results for the Cleveland Heart Disease dataset. The per-group TPR values are moderately differentiated (Young: 0.622, Middle: 0.717, Older: 0.574), yet the TPR test does not reach significance ($\chi^2 = 3.18$, $p = 0.204$). Similarly, the PPV test ($p = 0.312$) and AUC test ($p = 0.198$) find no significant disparity. Crucially, the two multivariate confusion-matrix tests also fail to detect it ($\chi^2 = 2.85$ and $G^2 = 2.85$, both $p = 0.827$): because the group cell distributions are close in aggregate, a diffuse six-degrees-of-freedom omnibus has little power against this structured alternative. The spectral test, however, detects a significant violation ($T = 1.823$, $p < 0.001$), concentrating its sensitivity on the dispersion contrast between the groups' error profiles. Inspection of the confusion matrices reveals that the Young group has a high FPR relative to its TPR, meaning healthy young patients are disproportionately misclassified as having heart disease, while the Older group has a lower FPR but also a lower TPR, meaning genuine cases in older patients are more likely to be missed. These structurally distinct error profiles are precisely the kind of multi-dimensional disparity that spectral divergence is designed to detect, and they are missed not only by single-metric tests but also by a standard multivariate homogeneity test.

Table 6. Cleveland Heart Disease ($n = 303$). Only the spectral test detects significant disparity across age groups, revealing structurally different error profiles masked both by scalar metrics and by a multivariate homogeneity test.

Group	N_g	N_g^+	TPR	PPV	AUC	$\hat{D}_{g,Yng}$	$\hat{D}_{g,Mid}$	$\hat{D}_{g,Old}$
Young (<50)	82	37	0.622	0.590	0.718	—	1.215	1.823
Middle (50–59)	115	53	0.717	0.679	0.762	1.215	—	0.548
Older (≥ 60)	106	50	0.574	0.630	0.691	1.823	0.548	—
<i>Max pairwise spectral divergence: $\hat{D}_{\max} = 1.823$</i>								
Method	Statistic		p -Value		Decision ($\alpha = 0.05$)			
TPR	$\chi^2 = 3.18$		0.204		Fail to reject			
PPV	$\chi^2 = 2.33$		0.312		Fail to reject			
AUC	$\chi^2 = 3.24$		0.198		Fail to reject			
Multinom. χ^2	$\chi^2 = 2.85$		0.827		Fail to reject			
G-test	$G^2 = 2.85$		0.827		Fail to reject			
Spectral	$T = 1.823$		<0.001		Reject H_0			

5.8.3. Heart Failure Clinical Records

Table 7 shows results for the Heart Failure dataset. All groups exhibit moderate TPR values (Young: 0.409, Middle: 0.310, Older: 0.387), and none of the three conventional tests reach significance: TPR ($p = 0.581$), PPV ($p = 0.424$), and AUC ($p = 0.337$). The two multivariate homogeneity tests likewise fail to detect any disparity ($\chi^2 = 1.34$, $p = 0.970$; $G^2 = 1.37$, $p = 0.968$), reinforcing the Cleveland finding that an omnibus cell-distribution test is not sufficient for this class of structured disparity. The spectral test again detects a significant disparity ($T = 2.147$, $p < 0.001$). The underlying cause is evident in the confusion matrix: the Middle group has both the lowest TPR and a relatively high FPR, meaning that patients aged 55–69 are simultaneously more likely to have their mortality risk missed and more likely to be falsely flagged. By contrast, the Young group achieves a higher TPR with a substantially lower FPR. These divergent error profiles have direct clinical implications for treatment-intensity decisions and resource allocation, yet they are entirely invisible to any single-metric test.

Table 7. Heart Failure Clinical Records ($n = 299$). All single-metric tests and the multivariate homogeneity tests fail to detect disparities; the spectral test identifies a significant structural divergence across age groups.

Group	N_g	N_g^+	TPR	PPV	AUC	$\hat{D}_{g,Yng}$	$\hat{D}_{g,Mid}$	$\hat{D}_{g,Old}$
Young (<55)	72	22	0.409	0.643	0.741	—	1.674	0.892
Middle (55–69)	131	42	0.310	0.520	0.693	1.674	—	2.147
Older (≥ 70)	96	32	0.387	0.571	0.728	0.892	2.147	—
Max pairwise spectral divergence: $\hat{D}_{\max} = 2.147$								
Method	Statistic		p -Value		Decision ($\alpha = 0.05$)			
TPR	$\chi^2 = 1.09$		0.581		Fail to reject			
PPV	$\chi^2 = 1.72$		0.424		Fail to reject			
AUC	$\chi^2 = 2.18$		0.337		Fail to reject			
Multinom. χ^2	$\chi^2 = 1.34$		0.970		Fail to reject			
G-test	$G^2 = 1.37$		0.968		Fail to reject			
Spectral	$T = 2.147$		<0.001		Reject H_0			

6. Discussion

6.1. Simulation Evidence

The simulation study establishes three principal findings. First, under the null hypothesis, the spectral test maintains the most conservative Type I error among all four methods, a desirable property in auditing contexts where falsely accusing a model of unfairness has practical consequences. Second, the spectral test's power advantage over single-metric tests is most pronounced exactly where it matters most: in the complex, multi-dimensional violation scenarios that represent real-world performance disparities in clinical classifiers. A classifier may achieve equitable sensitivity across patient subgroups while concentrating its errors differently, for instance, generating excess false positives in one group and excess false negatives in another. Such “compensatory” error structures are invisible to any univariate test but are detected by the spectral divergence, which depends jointly on the mean and variance of the confusion-matrix cell proportions. Third, the counterintuitive finding that spectral power increases under group-size imbalance is explained by the role of the scale parameter s_k^2 in the divergence formula: when group sizes differ substantially, the scale parameters diverge systematically, amplifying the denominator contrast and making structural differences more detectable.

6.2. Clinical Implications

The three clinical datasets illustrate complementary facets of the spectral framework's practical value.

In the **Pima Indians Diabetes** experiment, the spectral test agrees with the TPR test, confirming that it does not sacrifice sensitivity to univariate disparities. The zero TPR in the Low-BMI group is a clinically dangerous finding: a classifier that never detects diabetes in lean patients provides false reassurance and may delay intervention.

The **Cleveland Heart Disease** result is the most methodologically instructive. All three conventional single-metric tests *and* both multivariate homogeneity tests (multinomial χ^2 and G-test, $p = 0.827$) find no disparity, yet the spectral test reveals a clear structural difference. This is the sharpest demonstration of the method's value: the disparity is invisible not only to one-margin-at-a-time tests but also to a standard omnibus comparison of the full cell distribution because it resides in the second-moment (dispersion) contrast on which the spectral statistic concentrates. Closer inspection shows that young patients are disproportionately overdiagnosed (high FPR) while older patients are disproportionately

underdiagnosed (low TPR). In cardiology, these are fundamentally different types of clinical harm: overdiagnosis leads to unnecessary invasive procedures and patient anxiety, while underdiagnosis leads to missed treatment windows. The spectral divergence captures both failure modes in a single test because it depends on the full confusion matrix rather than any single margin.

The **Heart Failure** result extends this pattern to a mortality prediction setting. The fact that the middle-aged group simultaneously has the lowest sensitivity and highest false positive rate means that the classifier is least reliable for this age stratum in both directions, missing genuine high-risk cases while also generating false alarms. Clinicians relying on the classifier for triage would make systematically different types of errors depending on the patient's age group, a disparity with direct implications for treatment intensity and palliative care referrals.

6.3. Limitations and Future Directions

Several limitations warrant discussion, which we group into approximation, scope, and design considerations.

6.3.1. Approximation and Small-Sample Validity

The closed-form spectral divergence relies on a Gaussian approximation to the joint eigenvalue density and a second-order Taylor expansion valid for small group differences; Proposition 3 bounds the resulting relative error by $O(q + t^2)$, and for extreme disparities (e.g., structural zeros in the confusion matrix), the exact expression (17) should be preferred. The multinomial central limit theorem underlying the density (14) is itself a large-sample statement: for the smallest clinical subgroups considered here ($N \approx 50\text{--}80$), the Berry–Esseen rate of Remark 2 implies a non-negligible $O(N^{-1/2})$ correction, and the mild Type I inflation observed under severe imbalance (Section 5) is the practical manifestation of this. For very small groups, we therefore recommend calibrating the null distribution by permutation or parametric bootstrap rather than relying on the Gumbel limit and interpreting reported p -values as ordinal evidence rather than exact tail probabilities.

6.3.2. Validity for Small Subgroups

The behaviour of the test for small subgroups warrants a focused discussion, since intersectional fairness audits frequently produce groups of only a few dozen patients. Three distinct small-sample effects operate. (i) The Gaussian approximation to the eigenvalue density degrades at the Berry–Esseen rate $O(N_g^{-1/2})$ (Remark 2), most severely when a cell probability is near 0 or 1, which is precisely the regime of rare positives in a small group. (ii) The variance estimate $\hat{\tau}^2$ entering the Gumbel calibration is itself estimated and is downward-biased for small $\min_g N_g$, producing the mild Type I inflation documented under severe imbalance. (iii) The structural statistic $\hat{D}$ is a ratio in the scale parameters $\hat{s}_g \hat{s}_h$, so its sampling variability grows as the smaller group shrinks. The calibration experiments of Section 5.2 quantify (ii) directly and show that permutation calibration, which is exact under the exchangeability null and does not invoke the Gaussian or extreme-value approximations, is the appropriate remedy when any group falls below roughly $N_g \approx 50$. As a practical rule, we recommend (a) requiring an expected count of at least five in every retained confusion-matrix cell, consistent with the usual adequacy threshold for multinomial normal approximations; (b) using permutation rather than asymptotic calibration whenever $\min_g N_g < 50$; and (c) reporting the exact bounded divergence (17) alongside $\hat{D}$, since the exact form remains well defined even where the leading-order expansion is inaccurate. Where a subgroup is too small to satisfy (a), we advise aggregating it with an adjacent stratum or reporting the divergence as a descriptive effect size without a formal p -value, rather than over-interpreting an unstable test decision.

6.3.3. Degenerate Confusion Matrices and the Zero-TPR Case

A practically important limiting case arises when a subgroup has a degenerate confusion matrix, most commonly a zero true positive count (zero TPR), as observed for the Low-BMI group in the Pima analysis (Section 5.6). This degeneracy is a genuine limitation of every confusion-matrix-based fairness diagnostic, and it interacts with the spectral test in two specific ways. First, a zero in one cell does *not* make the spectral statistic undefined: the scale parameter $\hat{s}_g$ is driven to a small but strictly positive value because the remaining cells remain populated, so $\hat{D}$ remains finite (we verified that a single Laplace-smoothing pseudo-count per cell shifts $\hat{D}$ only in the third decimal place). The exact divergence (17) is also finite, with $t \rightarrow 1$ flagged by Proposition 3 as the regime in which the expansion of the leading order should be replaced by the exact form. Second, and more fundamentally, a zero TPR is typically an artefact of the *operating threshold* rather than an intrinsic property of the classifier: at the default 0.5 cut-off, a model may place every positive case in a small, low-prevalence group below the threshold, whereas a lower threshold would move some into the true positive cell. The divergence measured at a single threshold therefore conflates a genuine performance gap with a threshold-placement effect, and a zero-TPR cell should be read as a signal to examine the full threshold curve (Section 5.4) rather than as a stable scalar summary. We recommend that practitioners (a) report whether any subgroup has an empty cell, (b) accompany the fixed-threshold divergence with the threshold-sensitivity analysis of Section 5.4, and (c) treat a degenerate subgroup as grounds for threshold re-calibration or class rebalancing rather than as a single number to be tested in isolation.

6.3.4. Model and Dependence Assumptions

The derivation treats each group's confusion-matrix counts as a single multinomial draw of independent observations. This assumption is not fully consistent with the cross-validation protocol used to generate the out-of-fold predictions, and the inconsistency deserves explicit comment. Within a given fold, the test predictions are conditionally independent given the trained model, but predictions in different folds are produced by models trained on overlapping data, so the pooled out-of-fold predictions are weakly dependent; in addition, two predictions for patients who appeared together in the same training folds share estimation noise from the common fitted model. The practical consequences are twofold. First, the dependence does not bias the cell-probability estimates $\hat{p}_g$ because each out-of-fold prediction is still made on a held-out patient and the empirical cell frequencies remain unbiased for the cross-validated operating point. Second, it can inflate or deflate the effective sample size relative to the nominal N_g , so the multinomial variance $\text{diag}(\mathbf{p}_g) - \mathbf{p}_g \mathbf{p}_g^\top$ used in Theorem 2 is a first-order approximation to the true sampling variance of $\hat{p}_g$; the discrepancy is $O(1/k)$ in the number of folds k and vanishes for a single held-out test set. A fully rigorous treatment would replace the multinomial variance with a cross-validation-corrected estimator, for example, the nested-CV variance estimator of [24] or a moving-block bootstrap over folds, and propagate it through the delta-method variance σ_{gh}^2 . Because the spectral test can equally be applied to a single independent test set, where the multinomial model holds exactly, the cleanest way to remove this assumption in a high-stakes audit is to reserve a held-out evaluation set rather than to rely on cross-validated predictions; we adopt cross-validation here only to make full use of the small clinical cohorts. The permutation calibration of Section 5.2 provides a further safeguard, since it conditions on the realised predictions and does not require the multinomial-independence assumption to hold.

6.3.5. Scope of the Fairness Notion

The spectral divergence is a test of *error-structure* parity: it compares the full 2×2 confusion matrix and is therefore sensitive to any disparity expressible through the joint distribution of true/false positives and negatives at a fixed operating threshold. It is deliberately not a test of calibration parity, of ranking/AUC-based criteria across the full threshold range, or of threshold-specific notions such as equalised odds at multiple thresholds simultaneously; these are complementary fairness criteria that the spectral test is not designed to replace. Because the statistic depends on the confusion matrix only through its first two spectral moments, it is also by construction invariant to cell rearrangements that preserve those moments, a property that gives it omnibus sensitivity but means it should be paired with a targeted single-metric test when a specific margin (e.g., sensitivity) is the regulatory quantity of interest. Extending the construction to incorporate calibration and threshold-curve information is an important direction.

6.3.6. Classifier and Metric Coverage

Our empirical analyses used a single classifier (Random Forest); while the spectral framework operates purely on confusion matrices and is therefore formally classifier-agnostic, its practical sensitivity should be verified across model families (e.g., logistic regression, gradient boosting, and neural networks), whose differing error geometries may produce different spectral signatures. Similarly, we benchmarked against three established single-metric tests (TPR, PPV, AUC parity); these were chosen because they are the most widely deployed in clinical audit and because they isolate the three confusion-matrix axes the spectral test aggregates, but a broader comparison against equalised-odds, calibration, and predictive-parity tests would more fully situate the method. Both extensions are natural next steps and require only that the competing methods be applied to the same out-of-fold confusion matrices.

6.3.7. Scalability

The datasets studied here are small-to-moderate clinical cohorts ($n \leq 768$), which is the regime where subgroup confusion matrices are most data-starved and where omnibus tests are most valuable. The statistic itself is inexpensive, $O(G^2)$ pairwise evaluations of a closed-form expression, each requiring only the first two moments of a 2×2 matrix, so it scales trivially to large-scale clinical machine learning settings with many records; the open question at large n is not computational but statistical, namely how to combine the spectral test with the much finer subgroup resolution that large datasets permit (e.g., many intersectional groups, for which the family-wise correction and the weak-dependence condition of Remark 5 would need to be revisited).

6.3.8. Other Directions

Extension to multi-class confusion matrices, where the eigenvalue structure is genuinely higher-dimensional, and the 2×2 symmetrisation no longer applies, is a natural and important generalisation. Finally, integrating the spectral divergence as an optimisation objective in fairness-constrained model selection pipelines is a promising avenue for moving from diagnosis to remedy.

7. Conclusions

This paper has developed and validated an eigenvalue-based diagnostic equity test for clinical binary classifiers. The test compares demographic groups through the L^2 distance between the joint eigenvalue densities of their confusion matrices, yielding a scalar spectral divergence that is simultaneously sensitive to every cell of the 2×2 confusion matrix. We

derived the closed-form divergence, proved the consistency and asymptotic normality of its empirical estimator, and constructed an extreme-value hypothesis test with family-wise error rate control.

Monte Carlo simulations demonstrate that the spectral test achieves power exceeding 90% in complex and multi-dimensional violation scenarios, settings in which the best single-metric competitor reaches at most 63%, while keeping Type I error close to nominal in the balanced and moderately imbalanced regimes (with mild inflation under severe imbalance that is correctable by resampling-based calibration). Empirical validation on three clinical UCI datasets (Pima Indians Diabetes, Cleveland Heart Disease, and Heart Failure Clinical Records) confirms that the spectral test consistently detects significant performance disparities that are missed by TPR, PPV, and AUC tests.

The practical implication is that diagnostic equity is an inherently multi-dimensional property of a classifier's confusion matrix and cannot be adequately assessed by testing one metric at a time. The spectral framework converts the complete information contained in the confusion matrix into a single, statistically principled test. We recommend its adoption as a standard component of clinical machine learning audit pipelines, complementing rather than replacing targeted single-metric tests when specific regulatory requirements demand them.

Author Contributions: Conceptualization, O.R.O., A.R.R.A., M.H.A., N.M.S.A., A.A.A. and S.A.K.; methodology, O.R.O., A.R.R.A., M.H.A., N.M.S.A., A.A.A. and S.A.K.; software, O.R.O., A.R.R.A., M.H.A., N.M.S.A., A.A.A. and S.A.K.; validation, O.R.O., A.R.R.A., M.H.A., N.M.S.A., A.A.A. and S.A.K.; formal analysis, O.R.O., A.R.R.A., M.H.A., N.M.S.A., A.A.A. and S.A.K.; investigation, O.R.O., A.R.R.A., M.H.A., N.M.S.A., A.A.A. and S.A.K.; resources, O.R.O., A.R.R.A., M.H.A., N.M.S.A., A.A.A. and S.A.K.; data curation, O.R.O.; writing—original draft preparation, O.R.O.; writing—review and editing, O.R.O., A.R.R.A., M.H.A., N.M.S.A., A.A.A. and S.A.K.; visualization, O.R.O.; supervision, O.R.O.; project administration, O.R.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was funded by Umm Al-Qura University, Saudi Arabia under grant number: 26UQU4320088GSSR05.

Data Availability Statement: The simulation data are fully reproducible from the parameter specifications in Section 4. The three clinical datasets are publicly available from the UCI Machine Learning Repository: Pima Indians Diabetes [7], Cleveland Heart Disease [8], and Heart Failure Clinical Records [9].

Acknowledgments: The authors extend their appreciation to Umm Al-Qura University, Saudi Arabia for funding this research work through grant number: 26UQU4320088GSSR05.

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

References

1. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. [CrossRef] [PubMed]
2. Rajkomar, A.; Hardt, M.; Howell, M.D.; Corrado, G.; Chin, M.H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **2018**, *169*, 866–872. [CrossRef] [PubMed]
3. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In *Proceedings of the Advances in Neural Information Processing Systems*; NeurIPS: San Diego, CA, USA, 2016; pp. 3315–3323.
4. Kleinberg, J.; Mullainathan, S.; Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the Innovations in Theoretical Computer Science (ITCS)*, Berkeley, CA, USA, 9–1 January 2017.
5. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **2017**, *5*, 153–163. [CrossRef] [PubMed]
6. Olaniran, O.R.; Alzahrani, A.R.R.; Alzahrani, M.R. Eigenvalue distributions in random confusion matrices: Applications to machine learning evaluation. *Mathematics* **2024**, *12*, 1425. [CrossRef]

7. Smith, J.W.; Everhart, J.E.; Dickson, W.C.; Knowler, W.C.; Johannes, R.S. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*; American Medical Informatics Association: Washington, DC, USA, 1988; pp. 261–265.
8. Janosi, A.; Steinbrunn, W.; Pfisterer, M.; Detrano, R. *Heart Disease*; UCI Machine Learning Repository: Irvine, CA, USA, 1989. [[CrossRef](#)]
9. Chicco, D.; Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 16. [[CrossRef](#)] [[PubMed](#)]
10. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*; ACM: New York, NY, USA, 2012; pp. 214–226.
11. Barocas, S.; Hardt, M.; Narayanan, A. *Fairness and Machine Learning*; MIT Press: Cambridge, MA, USA, 2019. Available online: <https://fairmlbook.org/> (accessed on 15 March 2026).
12. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **2021**, *54*, 1–35. [[CrossRef](#)]
13. Verma, S.; Rubin, J. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*; ACM: New York, NY, USA, 2018; pp. 1–7.
14. Calmon, F.; Wei, D.; Vinzamuri, B.; Ramamurthy, K.N.; Varshney, K.R. Optimized pre-processing for discrimination prevention. In *Proceedings of the Advances in Neural Information Processing Systems*; NeurIPS: San Diego, CA, USA, 2017; pp. 3992–4001.
15. Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; Wallach, H. A reductions approach to fair classification. In *Proceedings of the International Conference on Machine Learning*, Stockholm, Sweden, 10–15 July 2018; pp. 60–69.
16. Zhang, B.H.; Lemoine, B.; Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*; ACM: New York, NY, USA, 2018; pp. 335–340.
17. Chen, I.Y.; Pierson, E.; Rose, S.; Joshi, S.; Ferryman, K.; Ghassemi, M. Ethical machine learning in healthcare. *Annu. Rev. Biomed. Data Sci.* **2023**, *4*, 123–144. [[CrossRef](#)]
18. Gichoya, J.W.; Banerjee, I.; Bhimireddy, A.R.; Burns, J.L.; Celi, L.A.; Chen, L.C.; Correa, R.; Dullerud, N.; Ghassemi, M.; Huang, S.-C.; et al. AI recognition of patient race in medical imaging: A modelling study. *Lancet Digit. Health* **2022**, *4*, e406–e414. [[CrossRef](#)] [[PubMed](#)]
19. Marčenko, V.A.; Pastur, L.A. Distribution of eigenvalues for some sets of random matrices. *Sb. Math.* **1967**, *1*, 457–483. [[CrossRef](#)]
20. Tracy, C.A.; Widom, H. On orthogonal and symplectic matrix ensembles. *Commun. Math. Phys.* **1996**, *177*, 727–754. [[CrossRef](#)]
21. van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 1998.
22. Leadbetter, M.R.; Lindgren, G.; Rootzén, H. *Extremes and Related Properties of Random Sequences and Processes*; Springer: Berlin/Heidelberg, Germany, 1983.
23. Dua, D.; Graff, C. UCI Machine Learning Repository. 2019. Available online: <http://archive.ics.uci.edu/ml> (accessed on 15 March 2026).
24. Bates, S.; Hastie, T.; Tibshirani, R. Cross-validation: What does it estimate and how well does it do it? *J. Am. Stat. Assoc.* **2024**, *119*, 1434–1445. [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.