Review Article

Journal of Epidemiology and Medical Statistics

Overlapping Coefficient in Biomedical Statistics: Methodology and Comparison with ROC-AUC

Mohammad O. Edous[1*], Bashar O. Edous[1] and Omar M. Eidous[2]

[1]*Faculty of Medicine, Yarmouk University, Irbid, Jordan*

[2]*Department of Statistics, Yarmouk University, Irbid, Jordan*

[*]**Corresponding authors**

Mohammad O. Edous, Faculty of Medicine, Yarmouk University, Irbid, Jordan.

Received: June 17, 2026; **Accepted:** June 29, 2026; **Published:** July 08, 2026

Abstract

The Weitzman overlapping coefficient (OVL) is a probabilistic measure of similarity between two distributions defined as the common probability mass shared by their density functions. In biomedical statistics, OVL has become increasingly attractive because of its direct interpretability and its ability to quantify distributional similarity between healthy and diseased populations. Although substantial progress has been made in parametric, nonparametric, and numerical estimation methods, the relationship between OVL and the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) has received comparatively little attention. This paper presents a unified overview of OVL estimation methods and investigates its theoretical and empirical relationship with ROC-AUC. A probabilistic representation of OVL is developed and compared with the classical probabilistic representation of ROC-AUC. Under the equal-variance normal model, explicit analytical expressions for both measures are reviewed, revealing a monotonic inverse relationship between them. A simulation study demonstrates how increasing distributional separation leads to decreasing overlap and increasing classification performance. The results indicate that OVL and ROC-AUC provide complementary information and should be interpreted jointly in biomedical applications.

Keywords: Overlapping coefficient, ROC-AUC, biomedical statistics, kernel density estimation, diagnostic accuracy, simulation study.

Introduction

The comparison of probability distributions is a fundamental problem in biomedical statistics. Measures of distributional similarity are frequently employed in biomarker evaluation, disease classification, treatment comparison, epidemiological studies, and clinical decision-making.

Among the available measures, the overlapping coefficient (OVL), originally proposed by Weitzman [1], has emerged as an intuitive measure of similarity between distributions. Unlike divergence measures such as the Kullback–Leibler divergence or Bhattacharyya distance, OVL has a direct probabilistic interpretation as the common probability mass shared by two populations. For continuous densities f and g, the overlapping coefficient (OVL) between f and g is defined by

$$OVL(f,g)=\int_{-\infty}^{\infty}\min\{f(x),g(x)\}dx$$

The coefficient satisfies:

$$0\le OVL(f,g)\le 1$$

Values near one indicate substantial overlap, whereas values near zero indicate strong separation.

During the last decade, considerable methodological developments have been made regarding estimation and inference of overlap coefficients. These include kernel-based methods numerical integration approaches asymptotic theory parametric procedures and multivariate extensions [5-8, 11-14].

In biomedical research, however, the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) remains the dominant measure for evaluating diagnostic performance. ROC-AUC is defined as

$AUC=P(X_D>X_H)$

where X_D and X_H denote diagnostic measurements from diseased and healthy populations, respectively.

Although both OVL and ROC-AUC are frequently interpreted as measures of discrimination, they quantify fundamentally

Citation: Mohammad O. Edous, Bashar O. Edous, Omar M. Eidous. Overlapping Coefficient in Biomedical Statistics: Methodology and Comparison with ROC-AUC. J Epidemiol Med Stat. 2026. 1(1): 1-5. DOI: doi.org/10.61440/JEMS.2026.v1.02

different characteristics of the underlying distributions.
The primary objective of this paper is therefore fourfold:

- Develop a unified estimation framework for OVL.
- Establish general consistency results.
- Derive a probabilistic representation of OVL analogous to ROC-AUC.
- Investigate the theoretical relationship between OVL and ROC-AUC.

The Overlapping Coefficient

For two continuous densities f and g,

$$OVL(f,g)=\int_{-\infty}^{\infty}\min\{f(x),g(x)\}dx$$

Using

$$\min(a,b)=\frac{a+b-|a-b|}{2}$$

we obtain

$$OVL(f,g)=1-\frac{1}{2}\int_{-\infty}^{\infty}|f(x)-g(x)|dx$$

Thus OVL is directly related to the total variation distance:

$$TV(f,g)=\frac{1}{2}\int_{-\infty}^{\infty}|f(x)-g(x)|dx$$

Consequently,

$$OVL(f,g)=1-TV(f,g)$$

This identity provides an important geometric interpretation of overlap.

OVL Estimation

Let $X_1,X_2,\ldots,X_{n1}$ be a random sample of size $n_1$from f and let $Y_1,Y_2,\ldots,Y_{n2}$ be another random sample of size n_2 from g, where the two samples are independent. A general estimator of OVL is:

$$\widehat{OVL}=\int_{-\infty}^{\infty}\min\left\{\hat{f}(x),\hat{g}(x)\right\}dx=1-TV\left(\hat{f},\hat{g}\right)$$

where $\hat{f}(x)$ and $\hat{g}(x)$ are the density estimators (parametric or nonparametric) of f(x) and g(x) respectively

Although the estimator of overlap coefficient is defined through the simple expression, its computation is often challenging in practice. The difficulty arises from the need to identify the regions where $\hat{f}(x)\le\hat{g}(x)$ and $\hat{f}(x)>\hat{g}(x)$, which generally requires determining the intersection points of the two density functions. For many parametric models these intersection points do not admit closed-form solutions, while for nonparametric density estimators the problem becomes even more complicated because the estimated densities are represented as sums of kernel functions. To overcome this difficulty, Eidous and Talafha [11] proposed a computational procedure for locating the density crossing points and evaluating the overlap coefficient accurately, thereby avoiding the need for explicit analytical solutions.

The nonparametric kernel density estimators of f(x) and g(x) are given by:

$$\hat{f}(x)=\frac{1}{n_1h_1}\sum_{i=1}^{n_1}K\left(\frac{x-X_i}{h_1}\right),\quad -\infty<x<\infty$$

and

$$\hat{g}(x)=\frac{1}{n_2h_2}\sum_{j=1}^{n_2}K\left(\frac{x-Y_j}{h_2}\right),\quad -\infty<x<\infty$$

The kernel function K is taken to be Gaussian kernel, i.e. $K(u)=\frac{1}{\sqrt{2\pi}}e^{\frac{-1}{2}u^2}$ and the smoothing parameters are computed by using the following rule recommended by Silverman [10],
$h_i=0.9\,An_i^{-1/5}$,i=1,2

where A=min{standard deviation,interquartile range/1.34} is computed based on the first sample $X_1,X_2,\ldots,X_{n1}$ to obtain h_1 and the second sample $Y_1,Y_2,\ldots,Y_{n2}$ to obtain h_2. The estimator $\widehat{OVL}_{Ker}$ of OVL(f,g) is given by [11],

$$\widehat{OVL}_{Ker}=\frac{1}{2}\left\{\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\frac{\min\left\{\hat{f}(X_k),\hat{g}(X_k)\right\}}{\hat{f}(X_k)}\right]+\frac{1}{n_2}\sum_{k=1}^{n_2}\left[\frac{\min\left\{\hat{f}(Y_k),\hat{g}(Y_k)\right\}}{\hat{g}(Y_k)}\right]\right\}$$

These methods of estimating OVL coefficients are consistent with earlier parametric, nonparametric kernel-based developments, and numerical approaches for OVL estimation [11–14].

ROC-AUC Estimation

The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) is defined as [15]
AUC=P(X>Y)=E I(Y<X),
where $X=X_D$ denotes the diagnostic measurement from the diseased population and $Y=X_H$ denotes the diagnostic measurement from the healthy population. The empirical estimator of AUC is [16]

$$\widehat{AUC}=\frac{1}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}I(Y_j<X_i).$$

This estimator is unbiased since

$$E\left(\widehat{AUC}\right)=\frac{1}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}EI(Y_j<X_i)$$

$$=\frac{1}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}P(Y<X)$$

$$=P(Y<X)=AUC$$

ROC-AUC Relationship with OVL

Using the identity

$$\min(a,b)=aI(a<b)+bI(b<a),$$

we obtain

$$OVL(f,g)=\int_{-\infty}^{\infty}f(x)\,I(f(x)\le g(x))dx+\int_{-\infty}^{\infty}g(x)\,I(g(x)<f(x))dx.$$

This can be rewritten in expectation form as

$$OVL(f,g)=E\,I(f(X)\le g(X))+E\,I(g(Y)<f(Y)),$$

Hence,

$$OVL(f,g)=P(f(X)\le g(X))+P(g(Y)<f(Y)).$$

This representation provides a probabilistic interpretation of OVL.
Unlike AUC, however, OVL is not based on pairwise comparisons between observations from the two populations. Instead, it compares the relative magnitudes of the two density functions at sampled locations.

The comparison suggests a deeper duality:

- ROC-AUC is a distributional separation measure in the sample space.
- OVL is a geometric overlap measure in the density space.

Both are rank-type probabilistic functional, but OVL operates on density ordering rather than observation ordering. This structural similarity explains why both quantities are widely used as measures of discrimination, yet they respond differently under scaling, tail behavior, and multimodality.

Normal-Theory Relationship Between OVL and ROC-AUC

Let $X \sim N(\mu_D, \sigma^2)$ and $Y \sim N(\mu_H, \sigma^2)$. Under these normal assumptions with equal variances, the AUC admits the closed-form expression:

$$AUC = \Phi\left(\frac{\delta}{\sqrt{2}}\right)$$

The overlapping coefficient (OVL) under the same model is given by:

$$\text{OVL} = 2\left(-\frac{\delta}{2}\right)$$

where

$$\delta = \frac{\mu_D - \mu_H}{\sigma},$$

and Φ denotes the cumulative distribution function of the standard normal distribution. This function can be approximated or bounded using several approaches (see, for example, [17-24].

These expressions show that both measures are deterministic functions of the standardized mean difference δ. Consequently, under the normal location model, ROC–AUC and OVL are mathematically linked and vary monotonically in opposite directions. However, despite this structural connection, they represent fundamentally different statistical concepts: ROC–AUC is based on probabilistic ranking performance, whereas OVL quantifies distributional overlap and similarity.

Theorem 1: Under the equal-variance normal model,

$$\text{OVL} = 2\left(-\frac{\Phi^{-1}(\text{AUC})}{\sqrt{2}}\right)$$

Proof: Since

$$AUC = \Phi\left(\frac{\delta}{\sqrt{2}}\right)$$

it follows that

$$\delta = \sqrt{2}\Phi^{-1}(\text{AUC}).$$

Substituting into the OVL formula yields

$$\text{OVL} = 2\left(-\frac{\Phi^{-1}(\text{AUC})}{\sqrt{2}}\right)$$

Simulation Study and Empirical Comparison

Simulation Design

A simulation study is conducted under a Gaussian biomedical separation model:

$$X_D \sim N(\mu_D, 1), X_H \sim N(0,1)$$

where:

$$\mu_D \in \{0, 0.5, 1, 1.5, 2, 3\}$$

with sample sizes (n_1, n_2) = (20,50), (100,100) and R=1000 replications.

OVL is estimated using kernel density estimation and $\widehat{\text{OVL}}_{\text{Ker}}$, while ROC-AUC is estimated using the empirical estimator $\widehat{\text{AUC}}$ as given in Section (4). The simulation results included the mean and mean square error (MSE) of each estimator are summarized in Table 1.

Table 1: True values of the overlapping coefficient (OVL) and the area under the curve (AUC), together with the Monte Carlo estimates of the mean and mean squared error (MSE) for the kernel-based estimator $\widehat{\text{OVL}}_{\text{Ker}}$, and the estimator $\widehat{\text{AUC}}$, under two sample size configurations (n_1, n_2) = (20,50), (100,100).

(n_1, n_2) = (20,50)

μ_D	True OVL	Mean of $\widehat{\text{OVL}}$ Ker	MSE of $\widehat{\text{OVL}}$ Ker	True AUC	Mean of $\widehat{\text{AUC}}$	Mean of $\widehat{\text{AUC}}$
0.0	1.0000	0.8316	0.0324	0.5000	0.4994	0.0058
0.5	0.8026	0.7611	0.0091	0.6382	0.6342	0.0052
1.0	0.6171	0.6146	0.0097	0.7603	0.7582	0.0040
1.5	0.4533	0.4722	0.0085	0.8556	0.8550	0.0024
2.0	0.3173	0.3489	0.0080	0.9214	0.9225	0.0012
3.0	0.1336	0.1598	0.0042	0.9831	0.9826	0.0002

(n_1, n_2) = (100,100)

μ_D	True OVL	Mean of $\widehat{\text{OVL}}$ Ker	MSE of $\widehat{\text{OVL}}$ Ker	True AUC	Mean of $\widehat{\text{AUC}}$	Mean of $\widehat{\text{AUC}}$
0.0	1.0000	0.9000	0.011	0.5000	0.5006	0.0016
0.5	0.8026	0.7968	0.0029	0.6382	0.6365	0.0014
1.0	0.6171	0.6263	0.0032	0.7603	0.7623	0.0011
1.5	0.4533	0.4728	0.0030	0.8556	0.8548	0.0007
2.0	0.3173	0.3401	0.0025	0.9214	0.9216	0.0003
3.0	0.1336	0.1531	0.0014	0.9831	0.9833	0.0001

Simulation Results

The simulation results in Table 1 provide a concise assessment of the finite-sample behavior of the proposed estimators and the relationship between OVL and AUC.

Kernel estimator of OVL

The kernel estimator shows reasonable accuracy across all settings. Small bias appears when the true overlap is close to one, which is expected due to boundary and smoothing effects in kernel estimation. For example, when μ_D=0,

$$OVL = 1, \widehat{OVL}_{Ker} = 0.8316, 0.9000$$

for sample sizes (20,50) and (100,100), respectively. The bias decreases as sample size increases.

The MSE also decreases clearly with sample size:
0.0091→0.0029 (μ_D=0.5),
0.0042→0.0014 (μ_D=3).

AUC estimator

The estimator AUĈ performs very well in all scenarios. Bias is negligible, and the estimates closely match the theoretical values. For instance, when μ_D=1,
AUC=0.7603,AUĈ=0.7582, 0.7623.

When μ_D=3,
AUC=0.9831,AUĈ=0.9826, 0.9833.

MSE values are very small and decrease with sample size, consistent with U-statistic properties.

elationship between OVL and AUC

A clear monotone inverse relationship is observed: as separation increases, OVL decreases and AUC increases. For example, when μ_D moves from 0 to 3, OVL moves from 1 to 0.1336, while AUC moves from 0.5 to 0.9831.
Under the normal model,

$$AUC = \Phi\left(\frac{\delta}{\sqrt{2}}\right), OVL = 2\Phi\left(-\frac{\delta}{2}\right),$$

which implies

$$\ddot{O}VL = 2\left(-\frac{\ddot{O}^{-1}(AUC)}{\sqrt{2}}\right)$$

Interpretation

Although related, OVL and AUC msure different features:
AUC=P(X_D>X_H),

$$OVL = \int_{-\infty}^{\infty} \min\{f(x), g(x)\}\, dx.$$

For example, when μ_D=3,
AUC=0.9831,OVL=0.1336,
showing strong discrimination but non-negligible overlap. When μ_D=0.5,
AUC=0.6382,OVL=0.8026,
indicating weak separation.

Overall, the simulation confirms that the kernel estimator is consistent, the AUC estimator is highly stable, and both measures provide complementary information about population separation.

Conclusion

This study developed a unified framework for the overlapping coefficient in biomedical statistics and evaluated its behavior relative to ROC-AUC. The results confirm a stable inverse relationship between both measures under Gaussian (normal) models and demonstrate their complementary roles. OVL is established as a robust measure of distributional overlap, while ROC-AUC remains a classification-oriented metric. Their joint interpretation provides a more complete understanding of biomedical data structures.

Use of artificial intelligence tools: The authors used ChatGPT to assist in the preparation of this manuscript. The author takes full responsibility for the content of the manuscript.

References

1. Weitzman MS. Measures of overlap of income distributions of white and negro families in the United States. 1970. Technical Report 22, U.S. Department of Commerce.
2. Zou L, Guo H, Berzuini C. Overlapping sample Mendelian randomization with multiple exposures: A Bayesian approach. BMC Medical Research Methodology. 2020. 20: 295.
3. Pastore M, Calcagnì A. Measuring distribution similarities between samples: A distribution-free overlapping index. Frontiers in Psychology. 2019. 10: 1089.
4. Al-Saidy O, Samawi HM, Al-Saleh MF. Inference on overlap coefficients under the Weibull distribution: Equal shape parameter. ESAIM: Probability and Statistics. 2005. 9: 206-219.
5. Alodat MT, Al Fayez M, Eidous OM. On the asymptotic distribution of Matusita's overlapping measure. Communications in Statistics – Theory and Methods. 2022. 51:6963-6977.
6. Eidous OM, Daradkeh SK. On inference of Weitzman overlapping coefficient Δ(X, Y) in the case of two normal distributions. International Journal of Theoretical and Applied Mathematics. 2024. 10: 14-22.
7. Eidous OM, Abu Al-Hayja'a, MM. Numerical integration approximations to estimate the Weitzman overlapping measure: Weibull distributions. Yugoslav Journal of Operations Research. 2023. 33: 699-712.
8. Eidous OM, Abu Al-Hayja'a, MM. Estimation of overlapping measures using numerical approximations: Weibull distributions. Jordan Journal of Mathematics and Statistics. 2023. 16: 741-761.
9. Eidous OM, Ananbeh EA. Kernel method for estimating overlapping coefficient using numerical integration methods. Applied Mathematics and Computation. 2024. 462: 128339.
10. Silverman BW. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London. 1986.
11. Eidous OM, Al-Talafha SA. Kernel method for overlapping coefficients estimation. Communications in Statistics – Simulation and Computation. 2022. 51: 5139-5156.
12. Eidous OM, Ananbeh EA. Kernel method for estimating Matusita overlapping coefficient using numerical approximations. Annals of Data Science. 2025. 12: 1265-1283.
13. Eidous OM, Alsheyyab MM. Generalization of the overlapping coefficient for k ≥ 2 normal distributions. Mathematical Modelling of Natural Phenomena. 2025. 20.
14. Maqableh HY, Eidous OM. On inference of Weitzman overlapping coefficient in two Weibull distributions. Journal of Health Statistics Reports. 2024. 3: 1-6.

15. Colak E, Mutlu F, Bal C, Oner S, Ozdamar K et al. Comparison of semiparametric, parametric, and nonparametric ROC analysis for continuous diagnostic tests using a simulation study and acute coronary syndrome data. Computational and Mathematical Methods in Medicine. 2012. 698320.
16. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology. 1975. 12: 387-415.
17. Eidous O. Improvements of Pólya upper bound for cumulative standard normal distribution and related functions. arXiv preprint. 2022. 2205.03485.
18. Eidous O. New tightness lower and upper bounds for the standard normal distribution function and related functions. Mathematical Methods in the Applied Sciences. 2023. 46: 15011-15019.
19. Ananbeh E, Eidous O. New simple bounds for the standard normal distribution function. Communications in Statistics – Simulation and Computation. 2025. 54: 2762-2769.
20. Edous M, Eidous O. A simple approximation for normal distribution function. Mathematics and Statistics. 2018. 6: 47-49.
21. Eidous O. Lower and upper bounds for the Q-function and its related functions. Communications in Statistics – Simulation and Computation. 2025. 1-8.
22. Eidous O, Abu Hawwas J. An accurate approximation for the standard normal distribution function. Journal of Information and Optimization Sciences. 2021. 42: 17-27.
23. Eidous O, Al-Rawwash M. Approximation for standard normal distribution function and its invertible. Journal of Algorithms & Computational Technology. 2025. 19: 1-5.
24. Eidous O, Ananbeh E. Approximations for cumulative distribution function of standard normal. Journal of Statistics and Management Systems. 25: 541-547.