BMC Research Notes

RESEARCH NOTE

Open Access

Comparison of methods of optimal cut-point selection for biomarkers in diagnostic medicine: a simulation study with application of clinical data in health informatics

Mojtaba Hassanzad[1], Karimollah Hajian-Tilaki[2,3*], Zinatossadat Bouzari[4] and Shahla Yazdani[4]

Abstract

Objectives Several methods of cut-point selection for biomarkers have been suggested in biomedical research but the superiority of them over others was not studied comprehensively under different pairs of distributions, degree of overlap, and the ratio of sample sizes. This simulation study was aimed to compare five popular methods with application of clinical examples.

Results The data of simulation was generated from the 12 configurations of binormal, bigamma, and biexponential pairs with different sample sizes The results showed that the four popular methods of Youden, Euclidean, Product, and Index of Union (IU) yielded identical optimal cut-point under binormal model with homoscedastic. While, with high AUC, the Youden may produce less bias and MSE, but for moderate and low AUC, Euclidean has less bias and MSE than other methods. The IU yielded more precise findings than the Youden for moderate and low AUC in binormal pairs, but its performance was lower with skewed distributions. In contrast, the cut-points produced by diagnostic odds ratio (DOR) were extremely high with low sensitivity and high MSE and bias. The results of clinical data showed that when AUC > 0.95, the five methods may produce identical cut-point, but DOR yields an extremely high value of cut-point for AUC < 0.95.

Keywords Youden, Euclidean, Product, Index of Union, Diagnostic odds ratio, Optimal cut-point, ROC analysis

Introduction

One of the important applications of ROC curve is to determine the optimal cut-off point for quantitative biomarkers [1, 2]. However, there is no single method for determining the optimal cut-point. Several methods of cut-point selection have been developed based on ROC curve analysis [3–5]. A reasonable subset of the most famous of them are Youden, Euclidean, Product, Index of union, and diagnostic odds ratio (DOR). Some of them are widely used in medical research for biomarkers in diagnosis and predicting outcomes. Each of these methods are defined using unique definition based on object function criteria in ROC space. The clinicians need to better understand the accuracy and precision of

*Correspondence:
Karimollah Hajian-Tilaki
drhajian@yahoo.com
[1] Student Research Center, Research Institute, Babol University of Medical Sciences, Babol, Iran
[2] Department of Biostatistics and Epidemiology, School of Public Health, Babol University of Medical Sciences, Babol, Iran
[3] Social Determinant Health Research Center, Research Institute, Babol University of Medical Sciences, Babol, Iran
[4] Department of Obstetrics and Gynecology, Ayatollah Rohani Hospital, Babol University of Medical Sciences, Babol, Iran

the proposed methods in clinical practice. Consistency and inconsistency of results of cut-point are possible in some conditions of screening test results [6]. This may depend on underlying distributions of test results in diseased and non-diseased and degree of separation of pairs of distributions. However, limited data is available for this matter and the question is which of the proposed methods determines the optimal cut-point precisely and more accurately? A few studies have been conducted based on the population-based distributions of diagnostic test data [6], as well as simulation study from a limited pair of distributions [7–12]. In some studies, inconsistency in determining the cut-off point was shown between some methods [13]. In other simulated studies, limited cases of certain distributions have been mainly addressed, and the impact of the size of the diagnostic accuracy and the inequality of the variances and the inequality of the sample size and the degree of severe skewness in the estimation of bias and MSE have not been widely evaluated.

There are several clinical examples that motivate the topic of biomarkers in early diagnosis of diseases and health-related outcomes in modern medicine. For example, premature rupture of membrane (PROM) refers to the rupture of the ammoniatic sac before labor begins that has been reported in 3–18% of pregnancies [14]. PROM increases the risk of perinatal mortality and accounts for approximately 18–20% of perinatal fetal deaths in the United States [15] and it is the cause of approximately one third of all premature births in America [16]. Its accurate diagnosis is important because failure to recognize it can lead to obstetric complications such as chorioamnionitis, premature birth, maternal and fetal infections, and prolapsed umbilical cord [17]. On the other hand, improper diagnosis of PROM can lead to unnecessary interventions such as hospitalization [18]. Some diagnostic methods such as nitrazine, pooling and Fern test, measuring vaginal diamino oxidase, prolactin, a-fetoprotein, insulin-like growth factor-binding protein 1, fetal fibronectin and placental 1 α-macroglobulin are currently available [19]. Tests such as nitrazine and pooling are expensive, and less used as screening tests. They are considered as our gold-standard. Laboratory biomarkers such as Beta-human chorionic gonadotropin (B-HCG), urea (BUN) and creatinine (Cr) are used as PROM screening tests [20, 21]. There was no a clear clarification of the methods in their cut-point selection. Therefore, the aim of this study is twofold. Beyond the simulation of data from different configurations of pairs of distributions and comparing the different methods of cut-point election, another aim is the clinical application in "Illustrations and applications with clinical examples of data" section.

Methods

Simulation study

Data was generated by R software in pairs of diseased and non-diseased distributions of bi-normal, bi-gamma, bi-exponential with certain parameters shown in Fig. 1 (12 panels A to L) with sample size and degree of accuracy with equal and unequal variance in 1000 runs. In each pair, the certain parameters were deliberately established that the area under the curve (AUC) is in the range of: low (AUC = 60), medium (AUC = 75), high (AUC = 90), which is the degree of overlap between pairs of distributions. The samples were produced in equal sizes of 50/50, 100/100, and 200/200, and unequal sizes of 50/100, 50/150, and 50/200 in the diseased and nondiseased population respectively that is a disease prevalence of 0.33, 0.25, and 0.20 respectively.

Statistical methods for the optimal cut-point

We focused on a subset of the five most popular methods, including Youden's J statistics, Euclidean distance, Product method, Index of Union (IU), and diagnostic odds' ratio (DOR). The full statistical descriptions have been illustrated elsewhere in detail [5]. In brief: (1) $C\text{-}_{Youden} = Max\ (Se\ (c) + Sp(c) - 1)$ that maximizes the percent of net classification that is clinically interesting [4]. (2) $C\text{-}_{Euclidean} = Min\{Sqrt[(1 - Se(c)]^2 + [1 = Sp(c)]^2\}$ that minimizes the Euclidian distance between the point on ROC curve to right corner (1, 0) in ROC space [10]. (3) Liu's method that maximizes the product of Se (c) and Sp (c) which is also known as Product methods [3]. (4) $C\text{-}_{Union} = Min\ \ |Se(c) - AUC| + |Sp(c) - AUC|$. This criterion minimizes the difference between Se and Sp and also the difference of the sum of Se and Sp by 2 times that AUC [7]. (5) $C\text{-}_{DOR}$ that maximizes the ratio of the positive likelihood to the negative likelihood. The latter index as a ratio metric has more fluctuations and its shape is convex under some distributional assumptions of diagnostic test results [5, 6, 22].

Determining the true optimal cut-points

First, we calculated the true values of cut-points with five methods under the parameters of different pairs of distributions that were presented in Fig. 1 by analytical calculating sensitivity (Se) and specificity (SP), and AUC for all possible cut-off values of decision scale using the Excel 21.0. The optimal cut-points were selected by maximizing or minimizing the related metrics depending on the methods used.

Appraisal of five different methods of cut-point selection

The performance of the estimates of cut-points were assessed by bias, relative bias (RB), mean square errors (MSE), the coverage rate of confidence interval (CI)

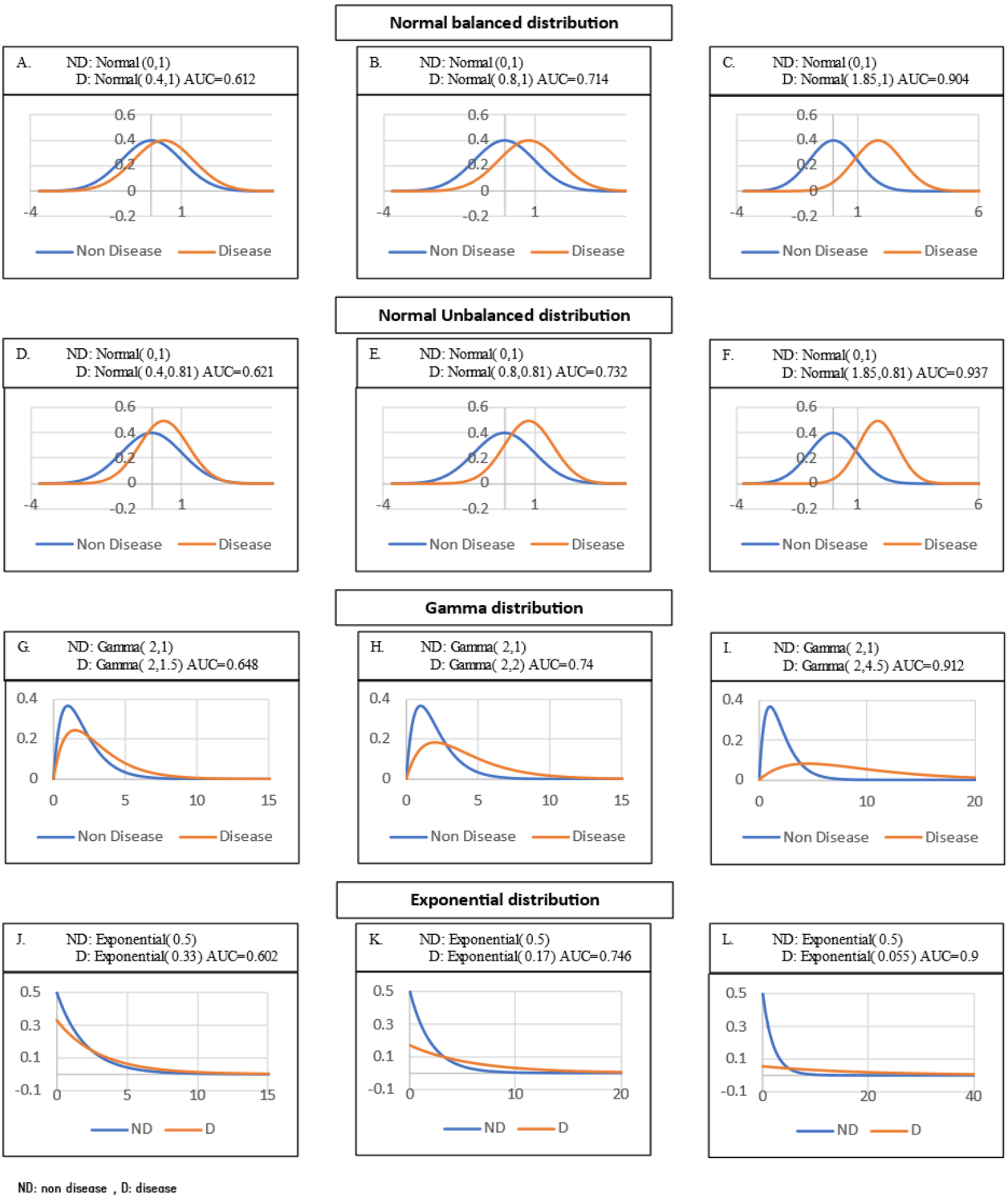

Fig. 1 Density plot of pairs of various distributions of nondiseased and diseased groups with different parameters

for true parameter of cut point, and mean length of CI. The average of cut-points was estimated in 1000 runs of datasets. Then the bias, relative bias and MSE were estimated. Their estimates were calculated by their empirical estimators in 1000 runs of data generated. For example, the empirical estimator of bias is determined by the average of estimated cut-points in 1000 runs minus the true value for each method respectively etc. In addition, in order to examine the percentage of coverage of true parameter of cut point, we applied the bootstrap resampling technique. To calculate the bootstrap estimate of cut- point of $\hat{c}$ and its standard deviation (SD), a random sampling with replacement was drawn to generate 200 bootstrap samples in in all configurations of distributions. Furthermore, to generate a 95% CI for the optimal cut point, the percentile method was applied considering the 2.5th and 97.5th percentiles of the bootstrap distribution of $\hat{c}$.

In cross-validation of the findings, first we performed our R code program with the parameters of distributions used by Unal [7] to generate data and then the outputs of our program were compared with those reported by Unal. If our outputs differ from those reported, we reexamined the R code program, repeating until the outputs become similar to those reported.

Ethical considerations

The study protocol was approved by the ethical Board of Babol University of Medical Sciences, Babol, Iran (Ethical code: IR.MUBABOL.HRI.REC.1402.308). The informed consent was obtained from all participants reported in "Illustrations and applications with clinical examples of data" section.

Results of simulation

Bi-normal model

Table 1 presents when data is generated from a homoscedastic binormal model. The least bias was found by Euclidean method and followed by Product, IU, Youden and DOR while the least MSE was observed by IU methods for low and moderate AUC, but Youden and Euclidean for high AUC. In all configurations of binormal model, both bias and MSE, as one expected, declined with higher sample size except for DOR. For a given sample size, the lowest bias and MSE were found by high AUC. The highest bias, relative bias and MSE are related to DOR as well, which is almost unacceptable. The bottom of Table 1 shows the results when data was generated with unbalanced sample size. The IU method has the lowest bias and is followed by product, Euclidean, Youden, and DOR. For unequal sample sizes. Unless the Youden, the MSE and bias of Euclidean, Product, and IU are lower for equal sample sizes than unequal for all configurations of degree of overlap. However, surprisingly, for Youden, the bias of equal sample size appeared to be higher than unequal sample sizes but not for a high AUC. The Youden index produced the less precise estimates of cut-points in particular for low and moderate AUC than the three other methods with relatively higher bias and MSE but not for high AUC. Overall, the highest MSE and bias were found by DOR in all configurations.

Table 2 indicates the results when data was generated by binormal model with non-homoscedastic. The IU method resulted the lowest MSE for low and moderate AUC but the Euclidean index yielded the lowest MSE for high AUC. While, the product method has the lower bias but a similar MSE with IU and Euclidean. The IU method produced the least MSE and bias that were followed by Euclidean and Product method. The results of coverage rate with homoscedastic binormal data, indicated that the four popular methods had a similar coverage rate of CI for true parameter of cut point ranging from 94 to 99% for equal and unequal sample size depending on AUC and sample size used but DOR had very poor coverage rate which none of CI did cover the true value of cut point at all. However, the IU methods had the smaller mean length of CI but the DOR had the highest value of mean length of CI (see Appendix in Table 1.b and Table 2.b).

Bi-gamma distributions

Table 3 presents the findings when data was generated by very skew pairs of distributions of Gamma with equal and unequal sample size. For balanced sample size, the least bias was found by Euclidean and it was followed by Product method, the Youden index, IU, and DOR but among the four popular methods, the greatest MSE was found by the Youden and the least by the Euclidean. For unequal sample size, similarly the greatest bias and MSE were attributed to IU and Youden respectively among the four methods while the Youden index had the least biased. The Euclidean is more precise than Youden index but the Youden was less biased. Similar to other pairs of distributions, the worst appraisal was found by DOR with extremely high MSE, bias and relative bias. The coverage rate of bootstrap CI for true values of cut point ranging from 95 to 98% have been observed using the three methods of Youden, Product, and Euclidean. However, the coverage rate of CI was declined from 77 to 89% for IU method. Meanwhile none of bootstrap CI did cover the true cut point by DOR method (see Table 3.b in Appendix).

Table 1 The MSE, bias and elative bias of different methods of optimal cut-point selection when data generated under binormal distributions with homoscedastic and equal/unequal sample size from diseased (D) and non-diseased (ND) according to degree of overlap or accuracy

Equal sample size[1]

Degree of overlap	Sample size	Youden index			Euclidian index			Product method			Index of Union			Diagnostic odds ratio		
	D = ND	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B
Low	50	0.2440	−0.0527	−0.2635	0.0532	−0.0022	−0.0109	0.0742	−0.0042	−0.0210	0.0177	−0.0134	−0.0668	86.782	−9.1075	−1.1147
	100	0.2039	−0.0483	−0.2414	0.0318	−0.0030	−0.0149	0.0412	−0.0040	−0.0198	0.0100	−0.0087	−0.0434	93.958	−9.4772	−1.1600
	200	0.1376	−0.0223	−0.1115	0.0198	0.0039	0.0196	0.0278	0.0076	0.0380	0.0066	−0.0028	−0.0141	99.779	−9.7638	−1.1951
Moderate	50	0.1448	−0.0481	−0.1204	0.0416	−0.0061	−0.0151	0.0633	−0.0075	−0.0189	0.0235	−0.0216	−0.0539	97.668	−9.7809	−1.1972
	100	0.1044	−0.0475	−0.1187	0.0274	−0.0066	−0.0165	0.0419	−0.0092	−0.0230	0.0158	−0.0154	−0.0386	103.43	−10.062	−1.2317
	200	0.0659	−0.0237	−0.0592	0.0159	−0.0043	−0.0108	0.0258	−0.0047	−0.0118	0.0123	−0.0109	−0.0272	109.53	−10.367	−1.2690
High	50	0.0739	−0.0395	−0.0429	0.0367	−0.0155	−0.0166	0.0545	−0.0166	−0.0179	0.0449	−0.0303	−0.0326	104.10	−10.193	−1.2476
	100	0.0463	−0.0204	−0.0221	0.0203	0.0001	0.0001	0.0330	−0.0031	−0.0033	0.0321	−0.0135	−0.0145	110.14	−10.487	−1.2837
	200	0.0302	−0.0109	−0.0118	0.0126	−0.0003	−0.0003	0.0222	−0.0003	−0.0003	0.0258	−0.0135	−0.0146	115.20	−10.724	−1.3127

Unequal sample size[1]

	D ≠ ND		MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B
Low	50	100	0.2383	−0.0161	−0.0804	0.0430	0.0228	0.1141	0.0585	0.0288	0.1438	0.0146	0.0010	0.0052	100.37	−9.8880	−1.2103
	50	150	0.2207	−0.0204	−0.1021	0.0388	0.0201	0.1004	0.0519	0.0232	0.1158	0.0134	0.0112	0.0559	106.14	−10.199	−1.2484
	50	200	0.2162	0.0138	0.0688	0.0406	0.0325	0.1627	0.0539	0.0261	0.1303	0.0121	0.0125	0.0624	108.95	−10.346	−1.2663
Moderate	50	100	0.1240	−0.0031	−0.0077	0.0404	0.0227	0.0568	0.0547	0.0240	0.0600	0.0210	0.0064	0.0161	106.39	−10.249	−1.2545
	50	150	0.1192	−0.0025	−0.0062	0.0345	0.0260	0.0650	0.0537	0.0289	0.0723	0.0196	0.0121	0.0303	110.97	−10.491	−1.2842
	50	200	0.1092	0.0125	0.0311	0.0337	0.0268	0.0669	0.0487	0.0259	0.0647	0.0183	0.0127	0.0317	113.80	−10.639	−1.3022
High	50	100	0.0624	−0.0236	−0.0257	0.0295	0.0033	0.0035	0.0451	−0.0039	−0.0042	0.0375	−0.0114	−0.0123	110.24	−10.493	−1.2843
	50	150	0.0571	0.0230	0.0250	0.0286	0.0323	0.0347	0.0459	0.0316	0.0340	0.0361	0.0225	0.0241	113.39	−10.642	−1.3027
	50	200	0.0591	0.0393	0.0427	0.0286	0.0408	0.0438	0.0476	0.0408	0.0439	0.0371	0.0250	0.0269	115.34	−10.735	−1.3140

1.$X_D \sim N(\mu_D, 1)$, $X_{ND} \sim N(0,1)$, and μ_D was taken as 0.4, 0.8, and 1.85, respectively

MSE: Mean square error, R.B.: Relative bias

Table 2 The MSE, bias and relative bias of different methods of optimal cut-point selection when data generated under binormal distributions with heteroscedastic variance and equal/unequal sample size from diseased (D) and non-diseased (ND) distributions according to degree of overlap/accuracy

Equal sample size[1]

Degree of overlap	Sample size	Youden index			Euclidian index			Product method			Index of Union			Diagnostic odds ratio		
	D = ND	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B
Low	50	0.1922	0.0662	−0.4731	0.0419	0.0141	0.1006	0.0572	0.0105	0.0959	0.0169	0.0503	0.3352	9.6550	−2.9039	−2.5473
	100	0.1270	0.0025	−0.0178	0.0242	0.0030	0.0215	0.0345	0.0085	0.0773	0.0108	0.0521	0.3471	11.816	−3.3278	−2.9191
	200	0.0775	0.0075	−0.0533	0.0148	0.0028	0.0198	0.0217	0.0017	0.0152	0.0079	0.0482	0.3213	13.601	−3.5951	−3.1536
Moderate	50	0.1043	−0.0287	−0.1197	0.0319	−0.0025	−0.0065	0.0482	0.0009	0.0027	0.0300	0.1065	0.3551	17.799	−4.1671	−1.9028
	100	0.0658	−0.0230	−0.0958	0.0212	0.0002	0.0004	0.0334	−0.0068	−0.0194	0.0231	0.0989	0.3296	20.387	−4.4862	−2.0485
	200	0.0427	−0.0097	−0.0402	0.0125	−0.0061	−0.0160	0.0187	−0.0074	−0.0211	0.0164	0.0877	0.2922	22.621	−4.7287	−2.1592
High	50	0.0577	−0.0313	−0.0337	0.0296	−0.0152	−0.0153	0.0436	0.0009	0.0010	0.0345	0.0224	0.0241	52.676	−7.2479	−1.3912
	100	0.0350	−0.0232	−0.0250	0.0162	−0.0062	−0.0063	0.0266	0.0044	0.0047	0.0239	0.0170	0.0183	56.903	−7.5361	−1.4465
	200	0.0232	−0.0129	−0.0139	0.0101	−0.0000	−0.0000	0.0187	0.0082	0.0086	0.0189	0.0069	0.0074	60.626	−7.7801	−1.4933

Unequal sample size[1]

	D ≠ ND		MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B
Low	50	100	0.1617	0.0504	−0.3603	0.0328	0.0234	0.1672	0.0452	0.0276	0.2511	0.0157	0.0618	0.4120	11.805	−3.3304	−2.9214
	50	150	0.1527	0.0779	−0.5567	0.0318	0.0344	0.2459	0.0415	0.0387	0.3515	0.0160	0.0732	0.4878	12.970	−3.5326	−3.0988
	50	200	0.1385	0.0769	−0.5493	0.0315	0.0418	0.2983	0.0437	0.0402	0.3653	0.0148	0.0710	0.4733	13.707	−3.6564	−3.2074
Moderate	50	100	0.0843	−0.0025	−0.0104	0.0298	0.0160	0.0420	0.0430	0.0091	0.0259	0.0321	0.1208	0.4028	20.419	−4.4900	−2.0502
	50	150	0.0814	0.0249	0.1037	0.0268	0.0269	0.0708	0.0390	0.0227	0.0647	0.0331	0.1305	0.4351	21.834	−4.6543	−2.1253
	50	200	0.0794	0.0556	0.2318	0.0249	0.0308	0.0809	0.0388	0.0367	0.1049	0.0332	0.1344	0.4479	22.692	−4.7500	−2.1690
High	50	100	0.0482	−0.0107	−0.0115	0.0227	0.0111	0.0112	0.0398	0.0048	0.0051	0.0317	0.0424	0.0456	56.925	−7.5369	−1.4466
	50	150	0.0469	0.0253	0.0272	0.0240	0.0280	0.0282	0.0402	0.0324	0.0341	0.0363	0.0652	0.0701	59.199	−7.6875	−1.4755
	50	200	0.0450	0.0265	0.0285	0.0216	0.0345	0.0349	0.0365	0.0379	0.0399	0.0347	0.0709	0.0763	60.553	−7.7751	−1.4923

1.$X_D \sim N(\mu_D, 0.81)$, $X_{ND} \sim N(0,1)$, and μ_D was taken as 0.4, 0.8, and 1.85, respectively

MSE: Mean square error, R.B.: Relative bias

Table 3 The MSE, bias and relative bias of different methods of optimal cut-point selection when data generated under bigamma distributions with equal/unequal sample size from diseased (D) and non-diseased (ND) according to degree of overlap/accuracy

Equal sample size[1]

Degree of Overlap	Sample size	Youden index			Euclidian index			Product method			Index of Union			Diagnostic odds ratio		
	D = ND	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B
Low	50	0.5859	−0.1071	−0.0441	0.1371	−0.0004	−0.0002	0.2103	0.0196	0.0091	0.0770	−0.1705	−0.0772	1471.9	−38.257	−0.9507
	100	0.4081	−0.0743	−0.0306	0.0777	0.0176	0.0083	0.1084	0.0187	0.0087	0.0537	−0.1516	−0.0686	1449.5	−37.921	−0.9424
	200	0.2725	−0.0479	−0.0197	0.0469	0.0171	0.0080	0.0670	0.0169	0.0078	0.0410	−0.1423	−0.0644	1420.2	−37.482	−0.9315
Moderate	50	0.5225	−0.1103	−0.0398	0.1534	0.0018	0.0007	0.2263	0.0172	0.0068	0.1590	−0.2720	−0.1034	1506.7	−38.726	−0.9624
	100	0.3140	−0.0545	−0.0197	0.0872	0.0155	0.0063	0.1513	0.0432	0.0171	0.1180	−0.2455	−0.0934	1497.4	−38.581	−0.9588
	200	0.2175	−0.0235	−0.0085	0.0498	0.0092	0.0037	0.0851	0.0307	0.0121	0.0831	−0.2027	−0.0771	1487.7	−38.423	−0.9549
High	50	0.5156	−0.1224	−0.0316	0.2672	0.0140	0.0040	0.4176	−0.0054	−0.0054	0.4365	−0.4156	−0.1074	1577.0	−39.684	−0.9862
	100	0.3264	−0.0723	−0.0187	0.1556	0.0295	0.0085	0.2726	0.0278	0.0074	0.2632	−0.3006	−0.0777	1575.8	−39.657	−0.9855
	200	0.1990	−0.0693	−0.0179	0.0825	0.0208	0.0060	0.1500	−0.0012	0.0003	0.1527	−0.2242	−0.0579	1574.1	−39.626	−0.9848

Unequal sample size[1]

	D ≠ ND		MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B	MSE	Bias	R.B
Low	50	100	0.5428	−0.0252	−0.0104	0.1047	0.0392	0.0185	0.1623	0.0457	0.0211	0.0618	−0.1407	−0.0637	1519.8	−38.889	−0.9664
	50	150	0.5093	−0.0299	−0.0123	0.1030	0.0722	0.0340	0.1627	0.0741	0.0343	0.0528	−0.1292	−0.0585	1535.6	−39.097	−0.9716
	50	200	0.5077	0.0444	0.0183	0.1117	0.0617	0.0291	0.1749	0.0882	0.0408	0.0513	−0.1295	−0.0586	1555.6	−39.373	−0.9785
Moderate	50	100	0.4163	−0.0318	−0.0115	0.1334	0.0589	0.0239	0.2073	0.0656	0.0259	0.1285	−0.2287	−0.0870	1548.5	−39.281	−0.9762
	50	150	0.3968	0.0510	0.0184	0.1190	0.0745	0.0303	0.1983	0.0965	0.0382	0.1051	−0.2000	−0.0761	1560.1	−39.432	−0.9799
	50	200	0.4239	0.0658	0.0238	0.1117	0.0662	0.0269	0.1856	0.0733	0.0290	0.1051	−0.2023	−0.0769	1575.6	−39.646	−0.9852
High	50	100	0.4459	−0.0482	−0.0124	0.2211	0.0749	0.0215	0.3740	0.0492	0.0132	0.3226	−0.3175	−0.0820	1590.4	−39.859	−0.9905
	50	150	0.4828	−0.0031	−0.0008	0.2002	0.1011	0.0291	0.3835	0.0805	0.0216	0.2805	−0.2716	−0.0702	1597.8	−39.958	−0.9930
	50	200	0.3951	0.0234	0.0060	0.2119	0.1217	0.0350	0.3345	0.0861	0.0231	0.2615	−0.2527	−0.0653	1602.7	−40.023	−0.9946

1.$X_D \sim G(2, \beta_D)$, $X_{ND} \sim G(2,1)$, and β_D was taken as 1.5, 2, and 4.5, respectively

MSE: Mean square error, R.B.: Relative bias

Bi-exponential distributions

Table 4 shows the appraisal findings with data of an extremely skew distributions of exponential pairs. The lowest bias was found by the Euclidean that followed by Product, IU, Youden and DOR respectively. The Euclidean has the lowest MSE that were followed by product, IU and Youden index. The Youden had the high MSE but IU had low MSE at low AUC. For unequal sample size, the lowest bias and MSE were observed by the Euclidean method. Similar to other scenarios, an extremely high MSE, and bias were found by DOR for all combinations. Overall, the extremely deviation of binormality, the bias, and MSE of all methods substantially increased. In this case, the least bias and MSE were observed by Euclidean. Moreover, the three methods of Youden, Euclidian, and Product had the high coverage of CI ranging from 95 to 98% while the IU produced the lower coverage rate ranging from 79 to 91% depending on the AUC and sample size used. Meanwhile, the poor performance of DOR has been observed in terms of coverage rate and mean length of CI (see Table 4.b in Appendix).

Illustrations and applications with clinical examples of data

Data

In a case control study of pregnant women in the third trimester of pregnancy suspected of having PROM were included in the study. These pregnant women were referred to the emergency obstetrics and gynecology clinic of Ayatollah Rouhani hospital in Babol, the north of Iran [19]. Based on the gold standard test status, 60 cases with PROM and 60 healthy individuals without PROM were diagnosed. Briefly, first, the informed consent was obtained from all patients. The full description of inclusion and exclusion criteria were described elsewhere [19]. Pregnant women who were diagnosed as negative in one of two gold standard tests of pooling or nitrazine were excluded from the study as suspicious subjects. Mothers who tested positive for both of these two tests were diagnosed with definite PROM (n = 60), and those who tested negative in both tests were considered as true negative (n = 60). The three biomarkers of BHCG, BUN, and Cr, by enzymatic photometry and Jafee methods and the results were recorded in PROM diagnostic database.

Results of cut-point selection of biomarkers

Figure 2 depicted the density function of three biomarkers in pregnant women with and without PROM in panel A, B, and C. The results Wilcoxon rank test showed that the values of biomarkers are significantly higher in PROM than without PROM (P = 0.001) and a higher SD of biomarkers were observed in PROM patients. Figure 3 shows the nonparametric ROC curve for BHCG, BUN, and Cr for diagnosis of PROM with a high diagnostic accuracy. In Fig. 3, the highest AUC (AUC = 0.992, 95%CI: 0.963, 0.998) was found by BHCG and followed by BUN (AUC = 0.975, 95%CI 0.929, 0.991), and Cr (AUC = 0.954, 95%CI 0.904, 0.978). In Table 5, the results show that for BHCG, and BUN, the five methods produced the identical cut-points (BHCG (44 IU/L), and BUN (1.07 mmol/L) while for CR, the DOR resulted in an extremely higher value of cut-points (40.66 μmol/L) with low sensitivity but the cut-point selection of the four other methods are identical (21.22 μmol/L). Figure 4 shows the changes in five metrics of cut-point selection over various cut-off values by different methods that have been shown with different colors in three panels for BHCG, BUN, and Cr.

Discussion

Our findings show that the IU method has the lowest bias, relative bias and MSE than other methods when data are generated from binormal model but not for a highly skew distribution of bigamma and biexponential pairs. The part of results related to the pairs of binormal model are in accordance with those reported by Liker Unal [7]. However, we found the poor performance of IU methods when data is generated from bigamma and biexponential pairs that was highly skewed. The IU may have a clinical interpretation in diagnostic appraisal. It simultaneously minimizes the difference between Se and Sp and also minimize the difference of either Se or Sp with AUC. This property might be clinically interesting in terms of diagnostic accuracy for cut-point selection.

On the other hand, the most popular method of Youden index that has a greater clinical interpretation in terms of net classification, the corresponded cut point is less precise especially for low and moderate AUC even under binormal data with equal sample sizes but nor for high AUC, and its bias and MSE are almost higher than Euclidean and product methods. The diagnostic performance of these two latter methods outrages than others with highly skewed distributions of diagnostic test results. While the product method maximizes the product of Se and Sp that might be interested clinically. Based on our findings, the more precise estimate of cut-point is estimated by Euclidean. These results are also in accordance with other reports [7, 8, 10]. Despite the higher precision and less biased of the Euclidean in some scenarios, it has less clinically interpretations.

Among the five methods of cut-point selection in this study, the worst method was DOR in term of extremely high bias and MSE and very low performance of coverage of CI for true cut point in all configurations of distributions studied. In a population-based distributions under different scenarios, it has been reported by Hajian-Tilaki

Table 4 The MSE, bias and relative bias of different methods of optimal cut-point selection when data generated under biexponential distributions with equal/unequal sample size from diseased (D) and non-diseased (ND) according to degree of overlap/accuracy

Equal sample size[1]																	
Degree of overlap	**Sample size**	**Youden index**			**Euclidian index**			**Product method**			**Index of Union**			**Diagnostic odds ratio**			
	D = ND	**MSE**	**Bias**	**R.B**	**MSE**	**Bias**	**R.B**	**MSE**	**Bias**	**R.B**	**MSE**	**Bias**	**R.B**	**MSE**	**Bias**	**R.B**	
Low	50	1.3970	−0.3006	−0.1232	0.2183	−0.0007	−0.0004	0.3020	0.0184	0.0100	0.0946	−0.1495	−0.0813	2192.5	−46.599	−0.9320	
	100	1.1067	−0.1676	−0.0687	0.1558	0.0183	0.0101	0.2040	0.0380	0.0206	0.0584	−0.1334	−0.0725	2157.9	−46.167	−0.9233	
	200	0.8694	−0.0393	−0.0161	0.0830	0.0136	0.0075	0.1245	0.0257	0.0140	0.0433	−0.1313	−0.0714	2099.1	−45.432	−0.9086	
Moderate	50	1.2422	−0.1779	−0.0544	0.3517	0.0263	0.0103	0.5716	0.0498	0.0182	0.3349	−0.3786	−0.1382	2262.8	−47.362	−0.9472	
	100	0.8318	−0.1231	−0.0376	0.2096	0.0282	0.0110	0.3660	0.0574	0.0210	0.2350	−0.3191	−0.1165	2239.3	−47.057	−0.9411	
	200	0.5211	−0.0969	−0.0296	0.1324	0.0279	0.0109	0.2294	0.0483	0.0176	0.1519	−0.2507	−0.0915	2239.5	−47.020	−0.9404	
High	50	1.8496	−0.2331	−0.0469	0.8368	0.0728	0.0181	1.4758	−0.0018	−0.0004	1.1215	−0.6995	−0.1521	2401.6	−48.911	−0.9782	
	100	1.0507	−0.2031	−0.0409	0.4273	0.0232	0.0058	0.9021	−0.0029	−0.0006	0.6957	−0.5225	−0.1136	2398.1	−48.851	−0.9770	
	200	0.7329	−0.1623	−0.0327	0.2325	0.0212	0.0053	0.5498	−0.0294	−0.0064	0.3676	−0.3916	−0.0851	2403.6	−48.902	−0.9780	
Unequal sample size[1]																	
	D ≠ ND		**MSE**	**Bias**	**R.B**	**MSE**	**Bias**	**R.B**	**MSE**	**Bias**	**R.B**	**MSE**	**Bias**	**R.B**	**MSE**	**Bias**	**R.B**
Low	50	100	1.2307	−0.2583	−0.1059	0.1872	0.0363	0.0201	0.2866	0.0630	0.0343	0.0677	−0.1283	−0.0697	2302.7	−47.799	−0.9560
	50	150	1.3978	−0.0562	−0.0230	0.2039	0.0809	0.0138	0.3071	0.1205	0.0655	0.0633	−0.1041	−0.0566	2335.3	−48.154	−0.9631
	50	200	1.3240	−0.0907	−0.0372	0.1852	0.0580	0.0320	0.2724	0.0846	0.0460	0.0591	−0.1108	−0.0602	2354.0	−48.360	−0.9672
Moderate	50	100	1.1099	−0.0389	−0.0119	0.3138	0.0742	0.0290	0.5186	0.1006	0.0367	0.2738	−0.3377	−0.1232	2369.0	−48.529	−0.9706
	50	150	1.0450	0.0012	0.0004	0.3105	0.1373	0.0536	0.5531	0.1647	0.0601	0.2281	−0.2742	−0.1001	2384.9	−48.686	−0.9737
	50	200	1.1334	0.0518	0.0159	0.2600	0.0963	0.0376	0.4370	0.1119	0.0408	0.2043	−0.2667	−0.0973	2415.2	−49.032	−0.9806
High	50	100	1.4932	−0.0966	−0.0194	0.7014	0.1167	0.0290	1.3132	0.0918	0.0199	0.8665	−0.5620	−0.1222	2449.8	−49.438	−0.9888
	50	150	1.4370	−0.0088	−0.0018	0.6562	0.2136	0.0531	1.2091	0.1662	0.0360	0.7004	−0.4604	−0.1001	2451.6	−49.450	−0.9890
	50	200	1.4623	0.0945	0.0190	0.6928	0.2503	0.0623	1.2345	0.2275	0.0492	0.6146	−0.4032	−0.0876	2468.2	−49.639	−0.9928

$X_D \sim E(\lambda_D)$, $X_{ND} \sim E(0.5)$, and λ_D was taken as 0.33, 0.17, and 0.055, respectively

MSE: Mean square error, R.B.: Relative bias

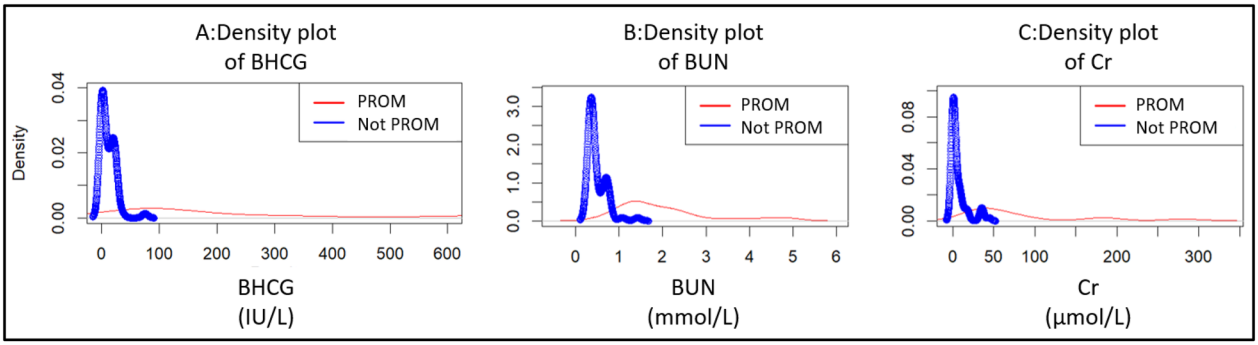

Fig. 2 The density plot of three biomarkers of BHCG, BUN, and Cr in healthy and PROM groups

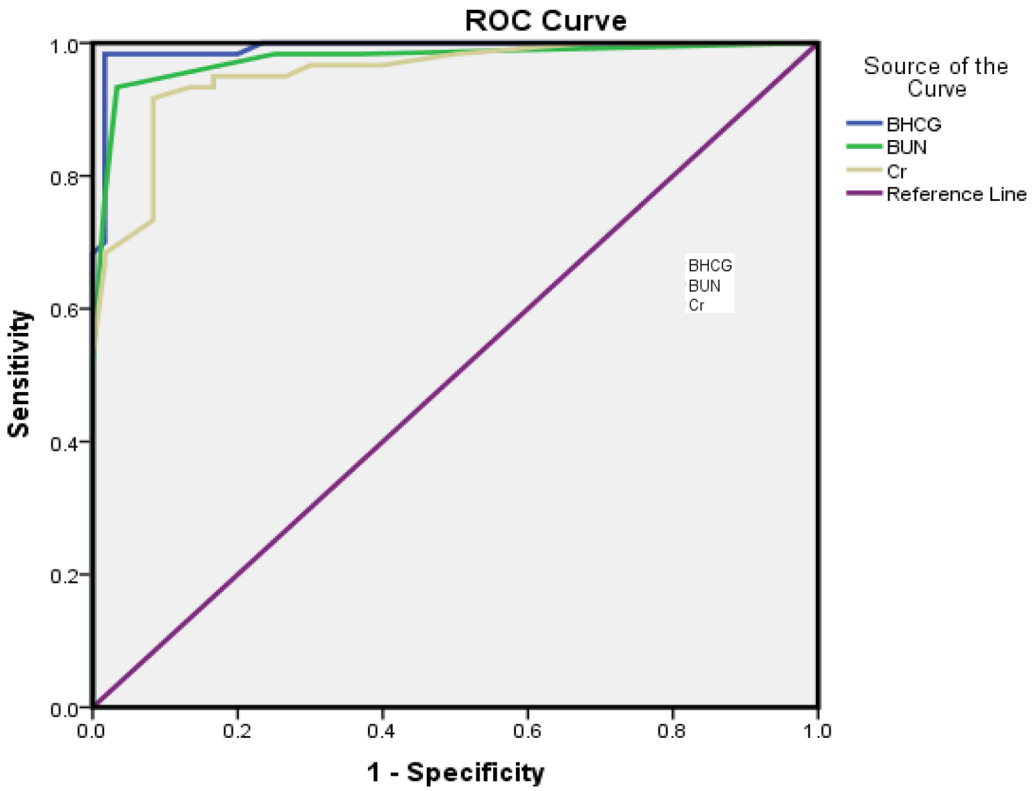

Fig. 3 Empirical ROC curve of three biomarkers of BHCG, BUN, and Cr in diagnosis of PROM

[6] that DOR produced unexpected high cut-point with poor Se because the convex pattern of DOR as ratio metrics [22]. Even under the bilogistic model DOR metric might be noninformative or have a linear trend that has not produced a proper optimal cut-point [6].

As one expected, in our simulation, the bias and MSE of all methods except for DOR, declined with increasing sample sizes and the higher degree of accuracy. From statistical perspective, the amount of data in term of sample sizes provides the more precise estimate and also less biased estimates of cut point. Meanwhile the high degree of separation pairs of distribution (or high AUC) leads to less room for sampling variability in ROC space. Therefore, the more precise estimates of cut-points are estimable in this scenario as our finding demonstrated. In particular, the results of current study showed the higher precise estimates and less biased with a high AUC by the Youden index.

Table 5 The diagnostic performance of cut-point of three biomarkers in diagnosis of PROM by different methods

Biomarker	Methods	Cut-point	Se (95%CI)	Sp (95%CI)	PPV (95%CI)	NPV (95%CI)	LR+ (95%CI)	LR− (95%CI)
BHCG (IU/L)	Youden Euclidian Product IU DOR	44.00	0.98 (0.91, 0.99)	0.98 (0.91, 0.99)	0.98 (0.91, 0.99)	0.98 (0.91, 0.99)	59.00 (8.43, 411.32)	0.02 (0.01, 0.12)
BUN (urea) (mmol/L)	Youden Euclidian Product IU DOR	1..07	0.93 (0.84, 0.98)	0.97 (0.88, 0.99)	0.96 (0.88, 0.99)	0.93 (0.84, 0.98)	28.00 (7.16, 109.67)	0.07 (0.03, 0.18)
Cr (µmol/L)	Youden Euclidian Product IU	21. 22	0.92 (0.82, 0.97)	0.92 (0.82, 0.97)	0.92 (0.82, 0.97)	0.92 (0.82, 0.97)	11.00 (4.74, 25.56)	0.09 (0.04, 0.21)
	DOR	40.6 6	0.68 (0.55, 0.80)	0.98 (0.91, 0.99)	0.98 (0.87, 0.99)	0.76 (0.65, 0.85)	41.00 (5.81, 287.93)	0.32 (0.22, 0.47)

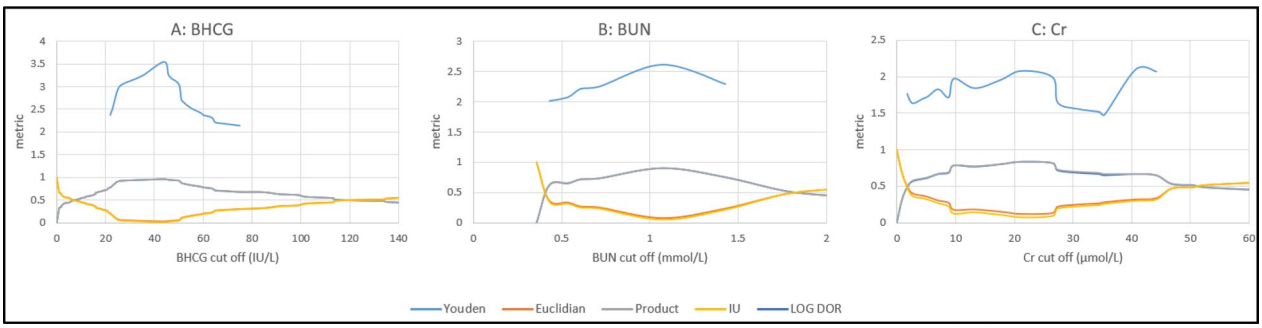

Fig. 4 The changes of five metrics of cut-off selection versus various cut off values of biomarkers BHCG, BUN, and Cr in diagnosis of PROM in panels of **A**, **B** and **C**, respectively

Moreover, we found, the inconsistency in determining true cut-points by different methods in particular with highly skew pairs of distributions. With binormal pairs and homoscedastic variance, the consistency of true cut-points values is possible but not for unbalanced variance and by the DOR. However, in analysis of our clinical example of data of biomarkers for diagnosis of PROM, surprisingly, identical results of cut-point were observed by five different methods for BHCG and BUN with AUC = 0.992 and AUC = 0.975 respectively. These identical results of all investigated methods can be explained by very high diagnostic performance of these two biomarkers. In this scenario, with extremely high diagnostic accuracy, there are less room in ROC space for variation of cut-points by different methods. In contrast, the inconsistency of cut-point by DOR with other methods in our simulation was present because the highest AUC in our simulation was considered as AUC = 0.90 but in our clinical example of detection PROM, the AUC for these two biomarkers were greater than 0.95. While for Cr that its diagnostic performance was lower than BHCG and BUN, the estimated cut-point of Cr by DOR was substantially higher than other methods with low performance of Se. Overall, the four competitive methods yielded identical results of cut-point for Cr as well but not DOR.

To our best knowledge, the design and results of the current simulation study are novel in terms of different configurations of distributions of diagnostic test results. So far, the other published simulation studies have not included the more extreme skew distributions with different degrees of overlapping pairs with five different methods simultaneously as we studied. Further simulation studies with other pairs of distributions may need to explore the performance of the different methods in other conditions.

Conclusion

Despite the clinical interest property of the Youden index, it may not produce a more precise estimate of the optimal cut-point for severe departure from binormality, in particular for low and moderate AUC. The greatest deviation from binormality, the bias and

MSE increased substantially in all methods. In a case, data generated from very skewed distributions of bigamma and biexponential, the lowest bias and MSE resulted from the Euclidean index and the highest yielded by DOR and IU, and Youden respectively. The precision and bias in estimating cut points by different methods may depend on the underling distributions of test results and AUC s, and the sample size used. However, the DOR has an extremely poor performance with very high bias and MSE, and very low coverage rate.

Limitations

The various methods for determining optimal cut-points optimize different objective functions and they have their own true cut-points. In many cases, the choice of the objective function is understood to depend on the specific purpose of the study either may focus on more weighting sensitivity or specificity, or the cost of false positives and false negatives or to maximize the sensitivity at a given value of the specificity. The objective function that has been defined as a criterion for cut-point selection has been criticized in the literature [23]. Moreover, our simulation was limited to the prevalence of 0.20, 0.25, 0.33, and 0.50 for diseased based on the ratios of sample sizes were considered in the study. In practice, the prevalence might be less than 0.20. However, the classical accuracy-based methods of cut-point selection are not influenced by the prevalence of disease, whereas its diagnostic performances as positive predicted value and negative predicted value are affected.

Abbreviations

ROC	Receiver characteristic curve
AUC	Area under the curve
Se	Sensitivity
Sp	Specificity
PPV	Positive predicted value
NPV	Negative predicted value
LR+	Positive likelihood ratio
LR−	Negative likelihood ratio
DOR	Diagnostic odds ratio
IU	Index of Union
MSE	Mean square error
PROM	Premature rupture of membrane
B-HCG	Beta-human chorionic gonadotropin
BUN	Blood urea nitrogen
Cr	Creatinine
D	Diseased
ND	Non-diseased

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13104-025-07245-9.

Supplementary Material 1.

Acknowledgements
We acknowledge the Deputy of Research and Technology of Babol University of Medical Sciences for their supports.

Author contributions
M.H: Conception in design, critical review, data analysis, interpretation, and drafting of manuscript. K.H: Conception in design, critical literature review, data analysis, manuscript drafting, and supervision. Z.B: Conception in design, clinical data collection, interpretation, drafting manuscript. S.Y: Conception in design, clinical data collection, interpretation, drafting of manuscript. All authors read and approved the final version of the manuscript.

Funding
Not applicable.

Availability of data and materials
Our data can not be shared openly to protect study participant privacy but it is available from corresponding authors upon reasonable request.

Declarations

Ethics approval and consent to participate
The clinical data has been used as an illustration of different methods of cut-point selection, has conformed to the standard of the World Medical Association, as embodied in the Declaration of Helsinki. The related protocol was approved by the local ethics committee of Babol University of Medical Sciences (ethical code: IR. MUBABOL.HRI.REC.1402.308) and the informed consent was obtained from all participants reported in "Illustrations and applications with clinical examples of data" section.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 19 December 2024 Accepted: 7 April 2025
Published online: 23 April 2025

References

1. Hajian-Tilaki KO, Hanley JA. Comparison of three methods for estimating the standard error of the area under the curve in ROC analysis of quantitative data. Acad Radiol. 2002;9(11):1278–85.
2. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med. 2013;4(2):627–35.
3. Liu X. Classification accuracy and cut point selection. Stat Med. 2012;31(23):2676–86.
4. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–5.
5. Hassanzad M, Hajian-Tilaki K. Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review. BMC Med Res Methodol. 2024;24(1):84.
6. Hajian-Tilaki K. The choice of methods in determining the optimal cut-off value for quantitative diagnostic test evaluation. Stat Methods Med Res. 2018;27(8):2374–83.
7. Unal I. Defining an optimal cut-point value in ROC analysis: an alternative approach. Comput Math Methods Med. 2017;2017:3762651.
8. Rota M, Antolini L. Finding the optimal cut-point for Gaussian and Gamma distributed biomarkers. Comput Stat Data Anal. 2014;69:1–14.

9. Mariska MG, Leeflang KGM, Moons JB, et al. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: Mechanisms, magnitude. Clin Chem. 2008;54:4. https://doi.org/10.1373/clinchem.2007.096032.
10. Gerke O, Zapf A. Convergence behavior of optimal cut-off points derived from receiver 0perating characteristics curve analysis: a simulation study. Mathematics. 2022;10:4206. https://doi.org/10.3390/math10224206.
11. López-Ratón M, Rodríguez-Álvarez MX, Cadarso-Suárez C, Gude-Sampedro F. Optimal Cutpoints An R package for selecting optimal cutpoints in diagnostic tests 2014. J Stat Softw. 2014;61(8):1–36. https://doi.org/10.18637/jss.v061.i08.
12. Kelson M, Dunstan FD, Lloyd K, Fone DL. Evaluating cutpoints for the MHI-5 and MCS using the GHQ-12: a comparison of five different methods. BMC Psychiatry. 2008;8(1):10. https://doi.org/10.1186/1471-244X-8-10.
13. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol. 2006;163(7):670–5.
14. Modena AB, Kaihura C, Fieni S. Prelabour rupture of the membranes: recent evidence. Acta Biomed. 2004;75(Suppl 1):5–10.
15. Caughey AB, Robinson JN, Norwitz ER. Contemporary diagnosis and management of preterm premature rupture of membranes. Rev Obstet Gynecol. 2008;1(1):11–22.
16. Simhan HN, Canavan TP. Preterm premature rupture of membranes: diagnosis, evaluation and management strategies. BJOG. 2005;112(Suppl 1):32–7.
17. Medina TM, Hill DA. Preterm premature rupture of membranes: diagnosis and management. Am Fam Physician. 2006;73(4):659–64.
18. Kim YH, Park YW, Kwon HS, Kwon JY, Kim BJ. Vaginal fluid beta-human chorionic gonadotropin level in the diagnosis of premature rupture of membranes. Acta Obstet Gynecol Scand. 2005;84(8):802–5.
19. Bouzari Z, Shahhosseini R, Mohammadnetaj M, Barat S, Yazdani S, Hajian-Tilaki K. Vaginal discharge concentrations of β-human chorionic gonadotropin, creatinine, and urea for the diagnosis of premature rupture of membranes. Int J Gynaecol Obstet. 2018;141(1):97–101.
20. Kariman N, Afrakhte M, Hedayati M, Fallahian M, Alavi MH. Diagnosis of premature rupture of membranes by assessment of urea and creatinine in vaginal washing fluid. Iran J Reprod Med. 2013;11(2):93–100.
21. Aguswahyudi D, Suwiyoga K, Negara IKS, Megadhana IW, Mahendra INB, Mayura IGPM, et al. The beta Human Chorionic Gonadotropin (hCG) levels in the vaginal washing fluid as a predictor of Premature Rupture of Membrane (PROM). Intisari Sains Medis. 2022;13(1):75–8.
22. Böhning D, Holling H, Patilea V. A limitation of the diagnostic-odds ratio in determining an optimal cut-off value for a continuous diagnostic test. Stat Methods Med Res. 2011;20(5):541–50.
23. Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. 2nd ed. Hoboken: John Wiley & Sons Inc; 2011. https://doi.org/10.1002/9780470906514.

Publisher's Note