Article

Receiver Operating Characteristic Prediction for Classification: Performances in Cross-Validation by Example

Andra Ciocan ^{1,2}, Nadim Al Hajjar ^{2,3,*}, Florin Graur ^{2,3}, Valentin C. Oprea ⁴, Răzvan A. Ciocan ⁵ and Sorana D. Bolboacă ^{1,*}

¹ Department of Medical Informatics and Biostatistics, “Iuliu Hațieganu” University of Medicine and Pharmacy Cluj-Napoca, Louis Pasteur Street, No. 6, 400349 Cluj-Napoca, Romania; Andra.Ciocan@umfcluj.ro

² “Prof. Dr. Octavian Fodor” Regional Institute of Gastroenterology and Hepatology Cluj-Napoca, Croitorilor Street, No. 19–21, 400162 Cluj-Napoca, Romania; florin.graur@umfcluj.ro

³ Department of Surgery, “Iuliu Hațieganu” University of Medicine and Pharmacy Cluj-Napoca, Croitorilor Street, No. 19–21, 400162 Cluj-Napoca, Romania

⁴ “Dr. Constantin Papilian” Military Emergency Hospital Cluj-Napoca, General Traian Moșoiu Street, No. 22, 400132 Cluj-Napoca, Romania; valentin.oprea@smucluj.ro

⁵ Department of Medical Skills—Human Sciences, “Iuliu Hațieganu” University of Medicine and Pharmacy Cluj-Napoca, Marinescu Street, No. 23, 400337 Cluj-Napoca, Romania; Razvan.Ciocan@umfcluj.ro

* Correspondence: nadim.alhajjar@umfcluj.ro (N.A.H.); sbolboaca@umfcluj.ro (S.D.B.)

Received: 14 September 2020; Accepted: 7 October 2020; Published: 10 October 2020

Abstract: The stability of receiver operating characteristic in context of random split used in development and validation sets, as compared to the full models for three inflammatory ratios (neutrophil-to-lymphocyte (NLR), derived neutrophil-to-lymphocyte (dNLR) and platelet-to-lymphocyte (PLR) ratio) evaluated as predictors for metastasis in patients with colorectal cancer, was investigated. Data belonging to patients admitted with the diagnosis of colorectal cancer from January 2014 until September 2019 in a single hospital were used. There were 1688 patients eligible for the study, 418 in the metastatic stage. All investigated inflammatory ratios proved to be significant classification models on both the full models and on cross-validations (AUCs > 0.05). High variability of the cut-off values was observed in the unrestricted and restricted split (full models: 4.255 for NLR, 2.745 for dNLR and 255.56 for PLR; random splits: cut-off from 3.215 to 5.905 for NLR, from 2.625 to 3.575 for dNLR and from 134.67 to 335.9 for PLR), but with no effect on the models characteristics or performances. The investigated biomarkers proved limited value as predictors for metastasis (AUCs < 0.8), with largely sensitivity and specificity (from 33.3% to 79.2% for the full model and 29.1% to 82.7% in the restricted splits). Our results showed that a simple random split of observations, weighting or not the patients with and without metastasis, in a ROC analysis assures the performances similar to the full model, if at least 70% of the available population is included in the study.

Keywords: receiver operating characteristic; area under the curve; models performances; biomarkers; metastasis

1. Introduction

The receiver operating characteristic (ROC) has been introduced in medicine as a methodological tool used for the investigation of signal detection [1–3]. The method uses the probability of correct detection (as a binary outcome) against the probability of a false positive outcome. The ROC detection

method has been widely adopted in medicine and it is used to confirm (*rule-in*) or exclude (*rule-out*) the presence of a disease, as instruments of triage, monitorization, prognosis or screening [4]. The ROC analysis is particularly used in laboratory medicine to identify the threshold able to discriminate pathological versus normal values of biomarkers [5–8], images in radiology [9–12] or accounts in bioinformatics [13–15]. The methodology of ROC analysis is extensively described in the scientific literature, with emphasize on the need of a gold standard method in order to certify the presence/absence of the disease, to understand the effects of verification bias (application of the test only on those with known disease status [16,17]) or to verify indices and metrics of accuracy (also known as classification performances) [4,9,18–21]. A gold standard or reference test is defined as the best available method able to classify the subjects as presenting or not a certain disease [22]. Most of the time, the available reference method is not necessarily a perfect method of classification, so we have an imperfect standard [23] and the effects of the measurement errors have already been evaluated [24,25]. Latent class models (LCM), which combined multiple test results to estimate the accuracy of a test in the absence of a reference standard have also been proposed and their utility proved [26–28].

The concept of big data provides the input data for uni- and multivariate association towards precision medicine [29]. New methods and learning algorithms emerged for analysis, medical data interpretation [30] and the main challenges are validation and replication. The classification model is identified in a training or development set and validated on a validation set [31]. The cross-validation, repeatedly randomised division of a dataset into development and validation sets, is applied to test the efficiency of the model on validation of external data [31–33]. Several cross-validation approaches, such as leave-one-out (LOO—one observation is left out of the development set) [34], leave-p-out (p observations are left out of the development set) [35], k -fold (the sample is split in k equal size sub-samples and one sub-sample is being repeated k -times, the one used for validation) [36], bootstrap method [37] or holdout method (also known as random splitting—observations are randomly split into two sets, development and validation set) [38] are used.

Colorectal cancer is the third most common type of cancer in men and second in women. The death rate from colorectal cancer is 8.5%. Surgical resection remains the only curative option for these patients, although a number of them develop metastases in the first five years after surgery [39]. Tumour size, tumour location, venous invasion, degree of differentiation proved valuable abilities to correctly predict local recurrences, metastasis and survival rate of colorectal cancer patients [40]. New biomarkers, such as chemokine ligand 7 (CXCL7) serum concentration have been tested for diagnostic accuracy [41], but the costs of such new biomarkers are generally higher than the conventional ones and health insurance systems do not support them. Consequently, a low-cost and time-bound predictor would be desirable. Pine et al. showed that tumour progression and prognosis are bound to tumour characteristics, as well as to the inflammatory response [42]. The inflammatory response is reflected by the level of neutrophils, lymphocytes, platelets, albumin and C-reactive protein, their measurement being current practice. Neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), lymphocyte-to-monocyte ratio (LMR) along with others, proved predictors for survival of patients with colorectal cancer [43–45]. Xia et al. reported NLR (cut-off = 2.8, AUC (Area Under the Curve) = 0.711, $p = 0.007$), LMR (cut-off = 3.9, AUC = 0.679, $p = 0.023$) and PNI (Prognostic Nutritional Index, cut-off = 47.1, AUC = 0.746, $p = 0.002$) as predictors for post-operative complications in patients with T1-T2 rectal cancer [46]. NLR (cut-off = 2.15, AUC = 0.790, 95% CI (0.736 to 0.884), CI = confidence interval) and PLR (cut-off = 123, AUC = 0.846, 95% CI (0.801 to 0.891)) presented performance as diagnostic markers for colorectal cancer as reported by Stojkovic Lalošević [47]. Different cut-off values of the biomarkers are reported in the range from moderate to low performances [43–50]. In the light of the state-of-the-art, our goals were two-folded, first to evaluate the performances of ROC models using three ratios associated with metastasis on unrestricted and restricted random cross-validation and secondly, to demonstrate the impact of these two randomizations approaches on the performances of the classification models.

2. Materials and Methods

The study was conducted according to the principles of the Declaration of Helsinki. It was approved by the Ethical Committee of “Iuliu Hațieganu” University of Medicine and Pharmacy Cluj-Napoca (approval no. 492/21.11.2019) and the Ethical Committee of “Prof. Dr. Octavian Fodor” Regional Institute of Gastroenterology and Hepatology Cluj-Napoca (approval no. 17517/20.12.2019). All participants signed the written informed consent before enrolment in the study.

2.1. Dataset

The dataset was represented by a retrospective collection of absolute neutrophil, lymphocyte, leucocyte, platelet counts of patients admitted at Third Surgical Clinic, “Prof. Dr. Octavian Fodor” Regional Institute of Gastroenterology and Hepatology Cluj-Napoca with the diagnosis of colorectal cancer from January 2014 until September 2019. The presence of metastasis was the outcome variable in this study. The metastatic stage was documented by imaging explorations, namely computer tomography (CT), magnetic resonance imaging (MRI) or contrast enhanced ultrasound. Raw data belonging to the patients with histopathological diagnosis of colorectal cancer were evaluated. Records with incomplete data (e.g., laboratory results, TNM classification of the disease—T size or extent of the primary tumor, N degree of spread to regional lymph nodes, and M presence of distant metastasis) were excluded.

The absolute counts were the input data for calculation of the following ratios, used as predictors of metastasis in this study:

$$\text{NLR} = (\text{absolute neutrophil count})/(\text{absolute lymphocyte count}) \quad (1)$$

$$\text{dNLR} = (\text{absolute neutrophil count})/[(\text{absolute leucocyte count}) - (\text{absolute neutrophil count})] \quad (2)$$

$$\text{PLR} = (\text{absolute platelet count})/(\text{absolute lymphocyte count}) \quad (3)$$

where NLR neutrophil-to-lymphocyte ratio, dNLR = derived neutrophil-to-lymphocyte ratio and PLR = platelet-to-lymphocyte ratio.

2.2. Methods

The receiver operating characteristic (ROC) analysis was conducted using three scenarios to test the classification abilities of the predictive models incorporating the studied ratios, which are presented in Equations (1)–(3) regarding the presence of metastasis (Figure 1):

- First scenario (❶): the whole sample was used to generate the classification model, the control model.
- Second scenario (❷): thirteen random sets were generated (1688 patients) by specifying the desired percentage of subjects and used to identify the classification models. No restrictions were imposed regarding the percentage of patients with and without metastasis and for this reason the generated sets were named unrestricted (*URun_x*, where $1 \leq x \leq 13$). The following percentages were used: 70% for *URun01*, 65% for *URun02*, 60% for *URun03*, 55% for *URun04*, 50% for *URun05*, 45% for *URun06*, 40% for *URun07*, 35% for *URun08*, 30% for *URun09*, 25% for *URun10*, 20% for *URun11*, 15% for *URun12*, 10% for *URun13*.
- Third scenario (❸): five sets, each with a development and validation group were randomly generated by weighting the percentage of patients with and without metastasis. These sets were name restricted (*RRun_y*, $1 \leq y \leq 5$; with 70% in development set & 30% in validation set). The model was generated using the development set (Se, Sp, Acc-Accuracy) and tested on validation set (PPV—Positive Predictive Value and NPV—Negative Predictive Value).

Figure 1. Flowchart of the applied methodology. NLR = neutrophil-to-lymphocyte ratio; dNLR = derived neutrophil-to-leucocyte ratio; PLR = platelet-to-lymphocyte ratio; Se = sensitivity; Sp = specificity; Acc = accuracy; +LR = positive likelihood ratio; -LR = negative likelihood ratio; PPV = positive predictive value; NPV = negative predictive value; MCR = miss-classification rate; DOR = diagnostic odds ratio; NND = number needed to diagnose; MCC = Matthews correlation coefficient; AUC = area under the curve; [95%CI] = lower bound to upper bound of the AUC 95% confidence interval; StdErr = standard error.

The performances of the models were evaluated whenever the AUC proved statistically significant and the Gini index (GI) was higher than 0 (GI equal with 0 indicates a random model). Ten metrics (see Figure 1) were used to prove the accuracy of a classification model. The sensitivity and specificity, important metrics widely used to show models accuracy [4,5,11,14,51,52]. The F1-score ($2 \times ((Se \times PPV) / (Se + PPV))$) was calculated by combining the performances in the development set with those in the validation set in restricted runs. Matthews correlation coefficient ($MCC = (TP \times TN - FP \times FN) / \sqrt{[(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)]}$) confers a more suitable balance to all four confusion matrix categories (TP—true positive, TN—true negative, FP—false positive, FN—false negative) and it is considered a more specific score [53], compared to the model accuracy or F1-Score, which are overoptimistic estimators [54].

The ROC analysis was conducted with SPSS v. 26 (trial version) under the non-parametric assumptions. The cut-off value for each biomarker in each run was used to create a derivate dichotomial variable (presence of metastasis for values equal to or higher than the cut-off). Youden's index ($J = \max$), which maximizes the distance from the diagonal (random classification model) [55] was used to identify the cut-off values. The tangent method ($d = \sqrt{((1 - Se)^2 + (1 - Sp)^2)}$, $d = \min$) [56] was also used to identify the cut-off values in the control model. The observed confusion matrix was generated and the true positive (TP), false positive (FP), false negative (FN) and true negative (TN) values were used to calculate the performances of the models using the following online resource: <https://statpages.info/ctab2x2.html> (accessed 3 August 2020). Clinical utility index (CUI) was calculated for the control model using the online resource available at <http://www.clinicalutility.co.uk/> (accessed 30 September 2020). A fair utility for case-finding (+CUI) or screening (-CUI) is seen if $0.49 \leq CUI < 0.69$, a good utility if $0.69 \leq CUI < 0.81$ and an excellent utility if $CUI \geq 0.81$ [57].

3. Results

3.1. First Scenario: Full Predictive Models

Significant contribution in identifying the presence of metastasis (AUC significantly different by 0.5) was identified for all investigated ratios, when the whole sample was used (Table 1).

Table 1. Characteristics of the full (control) models.

	AUC [95%CI]	StdErr	p-Value	Gini Index	Cut-Off Value *	
					J	d
NLR	0.647 [0.616 to 0.679]	0.0161	<0.0001	0.295	4.255	3.885
dNLR	0.634 [0.602 to 0.666]	0.0162	<0.0001	0.268	2.745	2.655
PLR	0.568 [0.535 to 0.601]	0.0168	<0.0001	0.136	255.560	162.98

NLR = neutrophil-to-lymphocyte ratio; dNLR = derived neutrophil-to-leucocyte ratio; PLR = platelet-to-lymphocyte ratio; AUC = area under the curve; [95%CI] = lower bound to upper bound of the AUC 95% confidence interval; StdErr = standard error; * J = max (Youden index), where Youden index = (Se + Sp - 1), Se = sensitivity; Sp = specificity; d = min.

The performances metrics of the control models showed modest or low classification abilities of the investigated ratios (Table 2), NLR being the most performant biomarker for metastasis.

Table 2. Performances of the (control) models (reported in Table 1).

	NLR	dNLR	PLR
<i>cut-off threshold by J = max</i>			
TP	220	228	139
TN	902	860	1006
FP	368	410	264
FN	198	190	279
Se	52.6 [47.8 to 57.4]	54.5 [49.8 to 59.3]	33.3 [28.7 to 37.8]
Sp	71.0 [68.5 to 73.5]	67.7 [65.1 to 70.3]	79.2 [77.0 to 81.4]
Acc	66.5 [64.2 to 68.7]	64.5 [52.2 to 66.7]	67.8 [65.6 to 70.1]
PPV	37.4 [34.4 to 40.3]	35.7 [33.0 to 38.5]	24.8 [22.8 to 26.9]
NPV	82 [80.4 to 83.6]	81.9 [80.2 to 83.6]	15.1 [11.9 to 18.9]
+LR	1.8 [1.6 to 2.1]	1.7 [1.5 to 1.9]	1.6 [1.3 to 1.9]
-LR	0.7 [0.6 to 0.7]	0.7 [0.6 to 0.7]	0.8 [0.8 to 0.9]
MCR	33.5 [31.5 to 35.6]	35.5 [33.5 to 37.6]	32.2 [30.2 to 34.1]
DOR	2.7 [2.2 to 3.4]	2.5 [2 to 3.2]	1.9 [1.5 to 2.4]
NND	4.2 [3.4 to 5.5]	4.5 [3.6 to 6]	8 [5.7 to 13.5]
+CUI	0.197 [0.144 to 0.250]	0.195 [0.413 to 0.247]	0.115 [0.059 to 0.171]
-CUI	0.582 [0.561 to 0.604]	0.555 [0.532 to 0.577]	0.620 [0.601 to 0.644]
F1-Score	51.4	49.7	18.8
MCC	0.214	0.198	0.126
<i>cut-off threshold by d = min</i>			
TP	244	235	229
TN	810	836	691
FP	460	434	579
FN	174	183	189
Se	58.4 [54.1 to 62.5]	56.2 [52.0 to 60.4]	54.8 [50.5 to 59]
Sp	63.8 [62.4 to 65.1]	65.8 [64.4 to 67.2]	54.4 [53 to 55.8]
Acc	62.4 [60.3 to 64.5]	63.4 [61.4 to 65.5]	54.5 [52.4 to 56.6]
PPV	34.7 [32.1 to 37.1]	35.1 [32.5 to 37.7]	28.3 [26.1 to 30.5]
NPV	82.3 [80.5 to 84.1]	82 [80.3 to 83.7]	78.5 [76.5 to 80.5]
+LR	1.6 [1.4 to 1.8]	1.6 [1.5 to 1.8]	1.2 [1.1 to 1.3]
-LR	0.7 [0.6 to 0.7]	0.7 [0.6 to 0.7]	0.8 [0.7 to 0.9]
MCR	37.6 [35.5 to 39.7]	36.6 [34.5 to 38.6]	45.5 [43.4 to 47.6]
DOR	2.469 [1.958 to 3.115]	2.474 [1.962 to 3.119]	1.4 [1.2 to 1.8]
NND	4.51 [3.62 to 6.05]	4.54 [3.63 to 6.09]	10.9 [6.8 to 28.5]
+CUI	0.202 [0.152 to 0.253]	0.197 [0.416 to 0.249]	0.155 [0.107 to 0.203]
-CUI	0.525 [0.502 to 0.548]	0.540 [0.517 to 0.563]	0.427 [0.402 to 0.453]
F1-Score	43.53	43.21	37.3
MCC	0.194	0.195	0.079

NLR = neutrophil-to-lymphocyte ratio; dNLR = derived neutrophil-to-leucocyte ratio; PLR = platelet-to-lymphocyte ratio; TP = true positive; TN = true negative; FP = false positive; FN = false negative; Se = sensitivity; Sp = specificity; Acc = accuracy; LR = likelihood ratio; MCR = miss-classification rate; DOR = diagnostic odds ratio; NND = number needed to diagnose.

3.2. Second Scenario: Unrestricted Random Samples

At least one unrestricted random sample retrieves the same Youden's cut-off value for each investigated ratios as the full predicted model (*URun01* for NLR and dNLR; *URun02* for dNLR; *URun05* for PLR; see Tables 1 and 3).

Table 3. Number of total subjects in the run, distribution among those with and without metastasis, cut-off values and models characteristics for unrestricted random samples.

	AUC [95%CI]	StdErr	p-Value	Gini Index	Cut-Off Value *
<i>URun01—1222 patients (72.4%); 307 with (73.4%) and 915 without (72.0%) metastasis</i>					
NLR	0.658 [0.621 to 0.695]	0.019	<0.0001	0.316	4.255
dNLR	0.648 [0.611 to 0.685]	0.019	<0.0001	0.297	2.745
PLR	0.565 [0.527 to 0.604]	0.020	0.0009	0.131	246.650
<i>URun02—1050 patients (62.2%); 267 with (63.9%) and 783 without (61.7%) metastasis</i>					
NLR	0.643 [0.602 to 0.683]	0.0208	<0.0001	0.285	3.635
dNLR	0.625 [0.584 to 0.666]	0.0208	<0.0001	0.250	2.745
PLR	0.586 [0.544 to 0.627]	0.0211	<0.0001	0.171	257.565
<i>URun03—1003 patients (59.4%); 258 with (61.7%) and 745 without (58.7%) metastasis</i>					
NLR	0.649 [0.609 to 0.689]	0.0203	<0.0001	0.298	3.625
dNLR	0.635 [0.595 to 0.676]	0.0205	<0.0001	0.271	2.755
PLR	0.584 [0.543 to 0.626]	0.0212	<0.0001	0.169	246.410
<i>URun04—905 patients (53.6%); 218 with (52.2%) and 687 without (54.1%) metastasis</i>					
NLR	0.678 [0.637 to 0.719]	0.0210	<0.0001	0.356	3.625
dNLR	0.656 [0.615 to 0.698]	0.0214	<0.0001	0.313	2.655
PLR	0.576 [0.531 to 0.621]	0.0229	0.0009	0.152	185.145
<i>URun05—848 patients (50.2%); 207 with (49.5%) and 641 without (50.5%) metastasis</i>					
NLR	0.623 [0.577 to 0.670]	0.0235	<0.0001	0.247	4.465
dNLR	0.597 [0.551 to 0.644]	0.0237	<0.0001	0.195	2.985
PLR	0.558 [0.512 to 0.605]	0.0236	0.0135	0.117	255.680
<i>URun06—772 patients (45.7%); 201 with (48.1%) and 571 without (45.0%) metastasis</i>					
NLR	0.639 [0.595 to 0.683]	0.0226	<0.0001	0.278	3.635
dNLR	0.621 [0.576 to 0.666]	0.0229	<0.0001	0.242	2.670
PLR	0.576 [0.53 to 0.622]	0.0236	0.0013	0.152	162.995
<i>URun07—664 patients (39.3%); 154 with (36.8%) and 664 without (52.3%) metastasis</i>					
NLR	0.654 [0.602 to 0.705]	0.0264	<0.0001	0.307	5.905
dNLR	0.637 [0.585 to 0.689]	0.0266	<0.0001	0.274	3.135
PLR	0.574 [0.520 to 0.627]	0.0272	0.0069	0.147	255.845
<i>URun08—597 patients (35.4%); 158 with (37.8%) and 439 without (34.6%) metastasis</i>					
NLR	0.616 [0.564 to 0.668]	0.0266	<0.0001	0.233	5.855
dNLR	0.607 [0.554 to 0.66]	0.0270	<0.0001	0.214	2.740
PLR	0.53 [0.476 to 0.584]	0.0276	0.2726	0.061	335.900
<i>URun09—507 patients (30.0%); 120 with (28.7%) and 387 without (30.5%) metastasis</i>					
NLR	0.639 [0.58 to 0.698]	0.0299	<0.0001	0.278	3.645
dNLR	0.622 [0.562 to 0.682]	0.0304	<0.0001	0.244	3.575
PLR	0.56 [0.502 to 0.619]	0.0299	0.0436	0.121	150.380
<i>URun10—403 patients (23.9%); 107 with (25.6%) and 296 without (23.3%) metastasis</i>					
NLR	0.684 [0.623 to 0.746]	0.0316	<0.0001	0.369	4.545
dNLR	0.673 [0.612 to 0.734]	0.0311	<0.0001	0.346	2.645
PLR	0.608 [0.541 to 0.675]	0.0343	0.0016	0.216	265.875
<i>URun11—320 patients (19.0%); 78 with (18.7%) and 242 without (19.1%) metastasis</i>					
NLR	0.665 [0.591 to 0.74]	0.0379	<0.0001	0.331	3.635
dNLR	0.664 [0.589 to 0.738]	0.0381	<0.0001	0.328	2.625
PLR	0.574 [0.501 to 0.647]	0.0372	0.0480	0.147	258.540
<i>URun12—260 patients (15.4%); 67 with (16.0%) and 193 without (15.2%) metastasis</i>					
NLR	0.627 [0.547 to 0.708]	0.0410	0.0019	0.255	4.250
dNLR	0.619 [0.538 to 0.7]	0.0413	0.0039	0.238	3.145
PLR	0.575 [0.489 to 0.661]	0.0441	0.0885	0.150	265.230
<i>URun13—136 patients (8.1%); 34 with (8.1%) and 102 without (8.0%) metastasis</i>					
NLR	0.674 [0.568 to 0.780]	0.0541	0.0013	0.348	3.215
dNLR	0.666 [0.556 to 0.775]	0.0561	0.0032	0.331	3.080
PLR	0.521 [0.408 to 0.634]	0.0578	0.7193	0.042	134.670

$n = 1688$; AUC = area under the curve; [95%CI] = lower bound to upper bound of the AUC 95% confidence interval; StdErr = standard error; * Youden index ($Se + Sp - 1$), where Se = sensitivity; Sp = specificity; NLR = neutrophil-to-lymphocyte ratio; dNLR = derived neutrophil-to-leucocyte ratio; PLR = platelet-to-lymphocyte ratio.

The obtained cut-off values were lower than the values reported in Table 1 for NRL (8/13, 61.5%; Table 3). Changes in the significance of the AUC were observed only for PLR for unrestricted random samples with 16% of the subjects in the development set (Table 3).

3.3. Third Scenario: Restricted Random Samples

The classification models in cross-validation, with 70% of the available data in the development set and 30% in the validation set, showed similar performances of AUCs for NLR and dNLR (*p*-values in the same range), but with AUCs close to 0.5 for PLR (in all *RRun*, Table 4).

Table 4. Models characteristics in restricted random samples.

	AUC [95%CI]	StdErr	<i>p</i> -Value	Gini Index	Cut-Off Value *
<i>RRun01</i>					
NLR	0.647 [0.610 to 0.684]	0.0189	<0.0001	0.295	4.225
dNLR	0.629 [0.591 to 0.666]	0.0190	<0.0001	0.257	2.755
PLR	0.559 [0.520 to 0.598]	0.0199	0.0032	0.117	268.675
<i>RRun02</i>					
NLR	0.638 [0.600 to 0.676]	0.0195	<0.0001	0.275	4.065
dNLR	0.616 [0.577 to 0.654]	0.0195	<0.0001	0.231	2.835
PLR	0.564 [0.525 to 0.603]	0.0200	0.0013	0.128	225.480
<i>RRun03</i>					
NLR	0.650 [0.612 to 0.689]	0.0195	<0.0001	0.300	4.475
dNLR	0.635 [0.596 to 0.673]	0.0196	<0.0001	0.269	3.135
PLR	0.575 [0.536 to 0.614]	0.0200	0.0002	0.151	255.560
<i>RRun04</i>					
NLR	0.643 [0.606 to 0.681]	0.0191	<0.0001	0.287	3.635
dNLR	0.626 [0.589 to 0.664]	0.0193	<0.0001	0.253	3.075
PLR	0.570 [0.531 to 0.609]	0.0200	0.0005	0.140	255.845
<i>RRun05</i>					
NLR	0.634 [0.596 to 0.672]	0.0192	<0.0001	0.268	4.225
dNLR	0.619 [0.581 to 0.657]	0.0193	<0.0001	0.238	2.755
PLR	0.570 [0.532 to 0.609]	0.0196	0.0003	0.141	261.695

NLR = neutrophil-to-lymphocyte ratio; dNLR = derived neutrophil-to-leucocyte ratio; PLR = platelet-to-lymphocyte ratio; AUC = area under the curve; [95%CI] = lower bound to upper bound of the AUC 95% confidence interval; StdErr = standard error; * Youden index ($Se + Sp - 1$), where *Se* = sensitivity; *Sp* = specificity.

The true positive, true negative, false positive and false negative in restricted random samples varied in the same way in the development and validation sets (Table 5). The above-mentioned values varied between runs in different amounts, ranging, for example, in the case of true positives from a difference between maximum and minimum form of 3.4% observed for PLR and 46.2% observed for AGR (Table 5).

The cross-validation ROC analysis applied on restricted random samples identified the NLR and dNLR as the most potent markers for metastasis according to the highest F1-Score (Table 6). However, better performances are obtained when the whole sample was investigated for NLR and dNLR, while on the opposite, better performances are obtained in cross-validation for PLR, all of them showing the instability of the classification model dictated by the input data (Tables 2 and 6).

Table 5. Number and frequencies of true positive, true negative, false positive and false negative in the restricted random samples according to run.

Ratio	Run	Development				Validation			
		TP	TN	FP	FN	TP	TN	FP	FN
NLR	RRun01	153 (52.4)	629 (70.8)	259 (29.2)	139 (47.6)	69 (54.8)	264 (69.1)	118 (30.9)	57 (45.2)
	RRun02	158 (54.1)	616 (68.6)	282 (31.4)	134 (45.9)	72 (57.1)	246 (66.1)	126 (33.9)	54 (42.9)
	RRun03	145 (49.7)	669 (75.3)	219 (24.7)	147 (50.3)	61 (48.4)	275 (72.0)	107 (28.0)	65 (51.6)
	RRun04	182 (62.3)	534 (60.1)	354 (39.9)	110 (37.7)	81 (64.3)	226 (59.2)	156 (40.8)	45 (35.7)
	RRun05	150 (51.4)	618 (69.6)	270 (30.4)	142 (48.6)	72 (57.1)	275 (72.0)	107 (28.0)	54 (42.9)
dNLR	RRun01	153 (52.4)	606 (68.2)	282 (31.8)	139 (47.6)	74 (58.7)	256 (67)	126 (33)	52 (41.3)
	RRun02	147 (50.3)	630 (70.2)	268 (29.8)	145 (49.7)	70 (55.6)	261 (70.2)	111 (29.8)	56 (44.4)
	RRun03	127 (43.5)	703 (79.2)	185 (20.8)	165 (56.5)	53 (42.1)	289 (75.7)	93 (24.3)	73 (57.9)
	RRun04	124 (42.5)	692 (77.9)	196 (22.1)	168 (57.5)	81 (64.3)	226 (59.2)	156 (40.8)	45 (35.7)
	RRun05	152 (52.1)	597 (67.2)	291 (32.8)	140 (47.9)	75 (59.5)	265 (69.4)	117 (30.6)	51 (40.5)
PLR	RRun01	86 (29.5)	729 (82.1)	159 (17.9)	206 (70.5)	38 (30.2)	324 (84.8)	58 (15.2)	88 (69.8)
	RRun02	108 (37.0)	668 (74.4)	230 (25.6)	184 (63.0)	48 (38.1)	282 (75.8)	90 (24.2)	78 (61.9)
	RRun03	95 (32.5)	725 (81.6)	163 (18.4)	197 (67.5)	39 (31.0)	301 (78.8)	81 (21.2)	87 (69)
	RRun04	94 (32.2)	722 (81.3)	166 (18.7)	198 (67.8)	39 (31.0)	305 (79.8)	77 (20.2)	87 (69)
	RRun05	85 (29.1)	734 (82.7)	154 (17.3)	207 (70.9)	43 (34.1)	306 (80.1)	76 (19.9)	83 (65.9)

NLR = neutrophil-to-lymphocyte ratio; dNLR = derived neutrophil-to-leucocyte ratio; PLR = platelet-to-lymphocyte ratio; TP = true positive; TN = true negative; FP = false positive; FN = false negative.

Table 6. Models performances in restricted random samples according to run.

	NLR	dNLR	PLR
	<i>RRun01</i>		
Se	52.4 [47.4 to 57.4]	52.4 [47.3 to 57.4]	29.5 [25.1 to 34]
Sp	70.8 [69.2 to 72.5]	68.2 [66.6 to 69.9]	82.1 [80.7 to 83.6]
Acc	66.3 [63.8 to 68.7]	64.3 [61.8 to 66.8]	69.1 [66.9 to 71.3]
+LR	1.8 [1.54 to 2]	1.65 [1.42 to 2]	1.65 [1.3 to 2]
-LR	0.67 [0.59 to 0.76]	0.7 [0.61 to 0.79]	0.86 [0.79 to 0.93]
MCR	33.7 [31.3 to 36.2]	35.7 [33.2 to 38.2]	30.9 [28.7 to 33.1]
DOR	2.7 [2 to 3.5]	2.4 [1.8 to 3.1]	1.9 [1.4 to 2.6]
NND	3 [2.8 to 3.2]	2.8 [2.6 to 3]	3.2 [3 to 3.5]
MCC _{dev}	0.025	0.022	0.016
PPV	36.9 [31.6 to 42]	37 [32 to 41.7]	39.6 [30.9 to 48.6]
NPV	82.2 [79.2 to 85.2]	83.1 [79.9 to 86.2]	78.6 [76.6 to 80.7]
MCC _{val}	0.038	0.041	0.032
F1-Score	43.3	43.4	33.8
	<i>RRun02</i>		
Se	54.1 [49 to 59.1]	50.3 [45.3 to 55.3]	37.0 [32.3 to 41.8]
Sp	68.6 [66.9 to 70.2]	70.1 [68.5 to 71.7]	76.7 [75.1 to 78.3]
Acc	65 [62.6 to 67.5]	65.3 [62.8 to 67.7]	66.7 [64.4 to 69.1]
+LR	1.72 [1.48 to 2]	1.69 [1.44 to 2]	1.59 [1.3 to 2]
-LR	0.67 [0.58 to 0.76]	0.71 [0.62 to 0.8]	0.82 [0.74 to 0.9]
MCR	35 [32.5 to 37.4]	34.7 [32.3 to 37.2]	33.3 [30.9 to 35.6]
DOR	2.6 [1.9 to 3.4]	2.4 [1.8 to 3.1]	1.9 [1.4 to 2.6]
NND	2.9 [2.7 to 3.1]	2.9 [2.7 to 3.1]	3 [2.8 to 3.2]
MCC _{dev}	0.024	0.022	0.013
PPV	36.4 [31.4 to 41.2]	38.7 [33.2 to 43.9]	34.8 [28.1 to 41.6]
NPV	82 [78.7 to 85.2]	82.3 [79.2 to 85.3]	78.3 [75.8 to 81]
MCC _{val}	0.037	0.042	0.025
F1-Score	43.5	43.7	35.9

Table 6. Cont.

	NLR	dNLR	PLR
<i>RRun03</i>			
Se	49.7 [44.7 to 54.6]	43.5 [38.7 to 48.3]	32.5 [28 to 37.2]
Sp	75.3 [73.7 to 76.9]	79.2 [77.6 to 80.7]	81.6 [80.2 to 83.2]
Acc	69 [66.5 to 71.4]	70.3 [68 to 72.7]	69.5 [67.3 to 71.8]
+LR	2.01 [1.7 to 2]	2.09 [1.73 to 3]	1.77 [1.41 to 2]
-LR	0.67 [0.59 to 0.75]	0.71 [0.64 to 0.79]	0.83 [0.76 to 0.9]
MCR	31 [28.6 to 33.5]	29.7 [27.3 to 32]	30.5 [28.2 to 32.7]
DOR	3 [2.3 to 4]	2.9 [2.2 to 3.9]	2.1 [1.6 to 2.9]
NND	3.2 [3 to 3.5]	3.4 [3.1 to 3.7]	3.3 [3.1 to 3.5]
MCC _{dev}	0.027	0.026	0.018
PPV	36.3 [30.5 to 42]	36.3 [29.9 to 42.8]	32.5 [25.3 to 40.2]
NPV	80.9 [78 to 83.7]	79.8 [77.2 to 82.5]	77.6 [75.4 to 79.9]
MCC _{val}	0.033	0.031	0.019
F1-Score	42	39.6	32.5
<i>RRun04</i>			
Se	62.3 [57.2 to 67.2]	42.5 [37.6 to 47.3]	32.2 [27.7 to 36.8]
Sp	60.1 [58.5 to 61.7]	77.9 [76.3 to 79.5]	81.3 [79.8 to 82.8]
Acc	60.7 [58.2 to 63.1]	69.2 [66.8 to 71.5]	69.2 [66.9 to 71.4]
+LR	1.56 [1.38 to 2]	1.92 [1.59 to 2]	1.72 [1.37 to 2]
-LR	0.63 [0.53 to 0.73]	0.74 [0.66 to 0.82]	0.83 [0.76 to 0.91]
MCR	39.3 [36.9 to 41.8]	30.8 [28.5 to 33.2]	30.8 [28.6 to 33.1]
DOR	2.5 [1.9 to 3.3]	2.6 [1.9 to 3.5]	2.1 [1.5 to 2.8]
NND	2.5 [2.4 to 2.7]	3.2 [3 to 3.5]	3.2 [3 to 3.5]
MCC _{dev}	0.024	0.023	0.018
PPV	34.2 [30 to 38.1]	38 [31.9 to 44]	33.6 [26.2 to 41.5]
NPV	83.4 [79.7 to 86.8]	81.1 [78.4 to 83.9]	77.8 [75.6 to 80.1]
MCC _{val}	0.038	0.037	0.021
F1-Score	44.2	40.1	32.9
<i>RRun05</i>			
Se	51.4 [46.3 to 56.4]	52.1 [47 to 57.1]	29.1 [24.8 to 33.6]
Sp	69.6 [67.9 to 71.2]	67.2 [65.6 to 68.9]	82.7 [81.2 to 84.1]
Acc	65.1 [62.6 to 67.6]	63.5 [61 to 66]	69.4 [67.3 to 71.6]
+LR	1.69 [1.45 to 2]	1.59 [1.36 to 2]	1.68 [1.32 to 2]
-LR	0.7 [0.61 to 0.79]	0.71 [0.62 to 0.81]	0.86 [0.79 to 0.93]
MCR	34.9 [32.4 to 37.4]	36.5 [34 to 39]	30.6 [28.4 to 32.7]
DOR	2.4 [1.8 to 3.2]	2.2 [1.7 to 2.9]	2 [1.4 to 2.7]
NND	2.9 [2.7 to 3.1]	2.7 [2.6 to 2.9]	3.3 [3.1 to 3.5]
MCC _{dev}	0.022	0.020	0.016
PPV	40.2 [34.7 to 45.5]	39.1 [33.9 to 43.9]	36.1 [28.7 to 43.9]
NPV	83.6 [80.6 to 86.4]	83.9 [80.7 to 86.8]	78.7 [76.4 to 81]
MCC _{val}	0.047	0.047	0.027
F1-Score	45.1	44.7	32.2

NLR = neutrophil-to-lymphocyte ratio; dNLR = derived neutrophil-to-leucocyte ratio; PLR = platelet-to-lymphocyte ratio; Se = sensitivity; Sp = specificity; Acc = accuracy; +LR = positive likelihood ratio; -LR = negative likelihood ratio; PPV = positive predictive value; NPV = negative predictive value; MCR = miss-classification rate; DOR = diagnostic odds ratio; NND = number needed to diagnose/predict; MCC = Matthews correlation coefficient; _{dev} = development set; _{val} = validation set.

The visual representation of the AUCs for each scenario related to the full model for NLR, dNLR, and PLR showed higher variability in the case of unrestricted random samples (Figure 2). The AUC representation of each run as compared to the full model is available in the Supplementary Material (Figures S1–S3).

Figure 2. The pattern of the AUCs: (a) neutrophil-to-lymphocyte ratio (NLR), (b) derived neutrophil-to-leucocyte ratio (dNLR), and (c) platelet-to-lymphocyte ratio (PLR). *URun* refers to unrestricted cross-validation, *RRun* refers to restricted cross validation.

4. Discussion

4.1. Summary of Main Findings

Cross-validation with random split of the data set under restrictions, namely respecting the proportion of those with and without metastasis, showed that the ROC classification models of the investigated ratios had low variability reported to the full model. The variability of ROC classification models in the unrestricted samples is very similar to the full model, when almost 70% of cases are in the development set with identical cut-off for 4/5 ratios (excepting PLR). For all other unrestricted samples, higher variability of the AUC was observed as compared to the restricted samples (Figure 2, Figures S1–S3).

4.2. Findings Discussion

The cut-off values are different, when different methods are used (Table 1), as expected and this affects the number of absolute frequencies in the confusion matrix. The Youden index maximizes both sensitivity and specificity and retrieves higher cut-off values for our dataset, as compared to minimization of the distance to the point with (0,1) coordinates. The number of false positives increases and the number of false negatives decreases, when the cut-off values are identified by the Tangent method as compared with the Youden's index, which reflects on both Se (increases) and Sp (decreases), but without significant effects on the LR or CUIs (Table 2). Even if the Tangent method showed superiority in identification of the true cut-off value [58], this was not observed in our sample. Despite the method used to define the threshold values, there are very poor case-findings. NLR and dNLR are fair for screening, while PLR is poor for screening (Table 2). On the other hand, they are not recommended as biomarkers for metastasis in patients with colorectal cancer.

As expected, different cut-off values were observed with few exceptions in both restricted and unrestricted cross-validation approaches (Tables 1, 3 and 4). Generally, not only for the unrestricted, but also for the restricted cross-validation, the cut-off values were lower than those of the full model. It resulted exactly the opposite for dNLR and PLR. Different cut-off values reported in cross-validation (Tables 3 and 4) indicate a variability, most probably caused by the characteristics of the included observations (the input data) and it has been observed in previously reported researches on patients with colorectal cancer [47–50].

The characteristics of the models were similar for both scenarios, with the AUCs values confined in the confidence interval of the AUC of the full model, without exception and with the AUC of the full model confined in the confidence interval of each individual classification model, regardless of the approach applied (Tables 1, 3 and 4). Opposite significances of the AUCs in unrestricted cross-validation in comparison to the full model was observed for PLR (2/13 classification models, one model that included 35.4% subject from the cohort and the other one that included 15.4%). The consistency of the significance across different samples indicates the stability of the classification models for the investigated ratios and the presence of metastasis as the outcome.

The cut-off values for 4/5 ratios were identical with those of the full model for the first unrestricted run, with almost 70% of subjects randomly selected for the development set. The relative value of the cut-off value from the full model $((\text{Cut-Off}_{\text{model}} - \text{Cut-Off}_{\text{full-model}}) / \text{Cut-Off}_{\text{full-model}} \times 100)$ was generally lower for the restricted models as compared to the unrestricted model. It can be explained by the small differences for those runs with ~70% of observations in the development sets. A pattern is observed, when the cut-off values are investigated. There are lower cut-off values in the majority of the cases for NLR (61.5% in the unrestricted and 80% in the restricted) and higher cut-off values for dNLR (61.5% in the unrestricted runs and 100% in the restricted runs) or PLR (53.8% in the unrestricted runs and 80% in the restricted runs). The cut-off values are reflected in the number of TP, TN, FN and FN and thus in the performances of the classification model (Tables 2 and 6). The evaluation of the full model performances showed that none of the investigated ratios is a good predictor for metastasis on patients with colorectal cancer (Table 2), with low +LR, high -LR, a relatively high

misclassification rate and MCC values less than 0.25. The low performances of the classification model are also observed on restricted cross-validation (Table 6), supporting the results reported by Martens et al. [59], namely that small changes in AUCs bring small changes in performances of the classification model. The ROC analysis is used to overcome this shortcoming, namely poor correlation of one predictor with a certain outcome, in order to find the threshold for classification and, based on the threshold, to consecutively identify a multivariable predictive model [42–50,60,61]. A small number of articles reported the threshold for the investigated inflammatory ratios on patients with colorectal cancer regarding metastasis as the outcome. Anuk and Yildirim reported a cut-off value for PLR equal to 194.7 for liver metastasis (Se = 74.5%, Sp = 72.7%) and of 163.95 (Se = 56.8%, Sp = 56.3%) for lymph node tumour cells invasion [62] on a sample of 152 patients. The thresholds for PLR reported in this study are higher than those reported by Anuk and Yildirim [62], nonetheless the specificity is higher at the expense of sensitivity (Tables 2 and 6). In our study we did not divided into the type of metastasis, but the differences in thresholds and Se(s) are important and could be possible explained by the characteristics of the target population. Guo et al. showed a large variability of PLR (25.4 to 300) thresholds in the evaluation of survival of the patients with colorectal cancer [45]. High PLR was associated with poor overall survival and high LMR was associated with favourable overall survival. The different cut-off values reported by Guo et al. should be explain by different population included in the analysis of the primary researches [45]. This explanation is also supported by the results reported by Xia et al. on a sample of 154 patients with T1-2 rectal cancer with cut-off of 2.8 for NLR and of 140.0 for PLR, when the primary outcome was considered death and the secondary one it was the occurrence of postoperative complications [46]. Peng et al. reported the optimal cut-off values of 4.63 for NLR (AUC = 0.568, 95%CI (0.477 to 0.660), $p = 0.152$) and 150.17 for PLR (AUC = 0.518, 95%CI (0.424 to 0.611), $p = 0.707$) on patients with colorectal cancer liver metastasis only [63]. The optimal cut-off values reported in this study are different by the those previously reported in the scientific literature. However, larger studies are needed to find the explanation for the reported thresholds and models performances.

The use of area under the ROC curve (AUC) in the assessment of classification model performances is known to be biased especially for small samples [64–68] and different cross-validation approaches have been proposed for model validation. The cross-validation approaches are usually used when computer algorithms are applied to identify the best performing classification model [64–69]. Moreover, leave-pair-out (LPO) cross-validation showed low bias in AUC estimation [70]. In our study, we applied cross-validation outside the computer algorithm in identification of the classification model and the unrestricted random sample with 70% of observations in the development set and it proved similar performances with the restricted cross-validation also with 70% of observations in the development set (Tables 1–3 and 5, Figures S1–S3). Our result supports thus the conclusion that, for large samples, an appropriate random sample with inclusion in the study of 70% of eligible subjects will closely reflect the target available population. Similarly, for large samples, the experimental design is the key factor for a valid ROC classification model, so the dictum “Garbage in, garbage out” seems true even for large data sets [71].

4.3. Study Limitations

Despite of a rigorous experimental design we need to list some limitations of our study. First limitation refers to the cross-validation methods used. The unrestricted cross-validation method was applied using the options of the program and unfortunately, the split was not saved, thus the performances of the models could not be reported. The simple random sampling proved its ability of appropriate splitting of a dataset in development (training) and validation (test) set, in the context of a predictive linear multivariate regression model [72]. Furthermore, only two cross-validation methods were applied and the use of other cross-validation methods (e.g., minimum of p -value on the confusion matrix [73], $\max(\text{Se} \times \text{Sp})$ [74], $\min(|\text{Se} - \text{Sp}|)$ [75]) could retrieve different cut-off values and thus different performances of classification models. However, since simple random

splitting in development and validation sets with ~70% of the observations in the development set and the restricted cross-validation (that also included 70% of the observations in the development sets) perform closely to the full model, no significant changes in terms of clinical utility are expected. Second limitation refers to the number of runs for each scenario, a higher number of runs being able to better reflect the reality. As it could be observed, we decided to report the performances in individual runs, not only to present an average of the performances, which is a common practice, but also to closely monitor at what length the performances differ from each other. Nevertheless, since the classification models showed stability in both restricted and unrestricted cross-validations, we are not expecting significant add-on with the increase in the number of runs. Third limitation is related to the considered outcome. We did not separately evaluate different types of colorectal cancer metastasis (e.g., liver, lung, brain, peritoneum, distant lymph nodes) and it can be expected that the thresholds of the investigated ratios vary with the type of metastasis. However, for such evaluation we need to expand the time frame for the retrospective evaluation, in order to assure a sufficient number of observations for each type of metastasis. Fourth limitation is related to the investigated inflammatory ratios. Investigation of the behaviour of other inflammatory ratios, such as lymphocyte-to-monocyte ratio, systemic immune inflammation index or prognostic nutritional index is also of interest. All these markers are taken into consideration by our research team and they will be tested for prediction abilities and proof of clinical utility.

5. Conclusions

The investigated ratios proved low clinical utility in predicting metastasis on patients with colorectal cancer. The full models showed fair clinical utility in screening, but the values of positive and negative likelihood ratios did not support their application in clinical settings. The classification models identified in the development sets, regardless of the use of unrestricted or restricted (percentage of patients with and without metastasis) random split, showed characteristics and performances similar to the full models. Our results showed that a simple random split of observations, weighting or not the patients with and without metastasis, in a ROC analysis assures the performances similar to the full models when at least 70% of the available population is included in the evaluation. Moreover, the sample size of the development set in the case of ROC classification analysis of investigated inflammatory ratios, considered as predictors for metastasis on patients with colorectal cancer, had little interest, in most of the cases, either on the models characteristics or on the models performances. The cut-off values of the investigated scenarios should probably be explained by the characteristics of the input data and are, most likely, linked to the percentage of correct and false classification, but without a significant impact on the models characteristics or performances. This behaviour supports the stability of the classification models in the context of a proper experimental design.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2227-7390/8/10/1741/s1>, Figure S1: “AUCs of neutrophil-to-lymphocyte ratio (NLR) in unrestricted (*URun*) and restricted (*RRun*) cross-validation related to the full model”, Figure S2: “AUCs of derived neutrophil-to-leucocyte ratio (dNLR) on unrestricted (*URun*) and restricted (*RRun*) cross-validation related to the full model.”, Figure S3: “AUCs of platelet-to-lymphocyte ratio (PLR) on unrestricted (*URun*) and restricted (*RRun*) cross-validation related to the full model”.

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

Funding: AC was funded by the Doctoral School of the “Iuliu Hațieganu” University of Medicine and Pharmacy, grant number 2461/8/22/17.01.2020.

Acknowledgments: We gratefully acknowledge the support of the University of Medicine and Pharmacy “Iuliu Hațieganu” Cluj-Napoca, Romania for administrative and technical support.

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

References

1. Egan, J.; Schulman, A.I.; Greenberg, G.Z. Operating Characteristics Determined by Binary Decisions and by Ratings. *J. Acoust. Soc. Am.* **1959**, *31*, 768. [\[CrossRef\]](#)
2. Emmerich, D.S. ROCs obtained with two signal intensities presented in random order, and a comparison between yes-no and rating ROCs. *Percept. Psychophys.* **1968**, *3*, 35–40. [\[CrossRef\]](#)
3. Lusted, L.B. Signal Detectability and Medical Decision-Making. *Science* **1971**, *171*, 1217–1219. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Bolboacă, S.D. Medical Diagnostic Tests: A Review of Test Anatomy, Phases, and Statistical Treatment of Data. *Comput. Math. Methods Med.* **2019**, *2019*, 1891569. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Campbell, G. General methodology I: Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Stat. Med.* **1994**, *13*, 499–508. [\[CrossRef\]](#)
6. Li, W.; Luo, S.; Zhu, Y.; Wen, Y.; Shu, M.; Wan, C. C-reactive protein concentrations can help to determine which febrile infants under three months should receive blood cultures during influenza seasons. *Acta Paediatr.* **2017**, *12*, 106. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Kampfrath, T.; Levinson, S.S. Brief critical review: Statistical assessment of biomarker performance. *Clin. Chim. Acta* **2013**, *419*, 102–107. [\[CrossRef\]](#)
8. Wolk, D.M. Clinical and Evidence-Based Research in the Clinical Laboratory. In *Clinical Laboratory Management*; Garcia, L.S., Ed.; ASM Press: Washington, DC, USA, 2013; pp. 832–848. [\[CrossRef\]](#)
9. Swets, J.A. ROC Analysis Applied to the Evaluation of Medical Imaging Tests. *Investig. Radiol.* **1979**, *14*, 109–121. [\[CrossRef\]](#)
10. Obuchowski, N.A. Receiver operating characteristic curves and their use in radiology. *Radiology* **2003**, *229*, 3–8. [\[CrossRef\]](#)
11. Gatsonis, C.A. Receiver Operating Characteristic Analysis for the Evaluation of Diagnosis and Prediction. *Radiology* **2009**, *253*, 593–596. [\[CrossRef\]](#)
12. Crivellaro, C.; Landoni, C.; Elisei, F.; Buda, A.; Bonacina, M.; Grassi, T.; Monaco, L.; Giuliani, D.; Gotuzzo, I.; Magni, S.; et al. Combining positron emission tomography/computed tomography, radiomics, and sentinel lymph node mapping for nodal staging of endometrial cancer patients. *Int. J. Gynecol. Cancer* **2020**, *30*, 378–382. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Lasko, T.A.; Bhagwat, J.G.; Zou, K.H.; Ohno-Machado, L. The use of receiver operating characteristic curves in biomedical informatics. *J. Biomed. Inf.* **2005**, *38*, 404–415. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Vihinen, M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genom.* **2012**, *13*, S2. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Pahari, S.; Li, G.; Murthy, A.K.; Liang, S.; Fragoza, R.; Yu, H.; Alexov, E. SAAMBE-3D: Predicting Effect of Mutations on Protein–Protein Interactions. *Int. J. Mol. Sci.* **2020**, *21*, 2563. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Fluss, R.; Reiser, B.; Faraggi, D.; Rotnitzky, A. Estimation of the ROC Curve under Verification Bias. *Biom. J.* **2009**, *51*, 475–490. [\[CrossRef\]](#)
17. Alonzo, T.A. Verification Bias—Impact and Methods for Correction when Assessing Accuracy of Diagnostic Tests. *Revstat. Stat. J.* **2014**, *12*, 67–83.
18. Metz, C.E. Basic principles of ROC analysis. *Semin. Nucl. Med.* **1978**, *8*, 283–298. [\[CrossRef\]](#)
19. Shapiro, D.E. The interpretation of diagnostic tests. *Stat. Methods Med. Res.* **1999**, *8*, 113–134. [\[CrossRef\]](#)
20. Zou, K.H.; O'Malley, A.J.; Mauri, L. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation* **2007**, *115*, 654–657. [\[CrossRef\]](#)
21. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [\[CrossRef\]](#)
22. Versi, E. “Gold standard” is an appropriate term. *BMJ* **1992**, *305*, 187. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Valenstein, P.N. Evaluating diagnostic tests with imperfect standard. *Am. J. Clin. Pathol.* **1990**, *93*, 252–258. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Phelps, C.E.; Hutson, A. Estimating diagnostic test accuracy using a “fuzzy gold standard”. *Med. Decis. Mak.* **1995**, *15*, 44–57. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Johnson, W.O.; Gastwirth, J.L.; Pearson, L.M. Screening without a “gold standard”: The Hui-Walter paradigm revisited. *Am. J. Epidemiol.* **2001**, *153*, 921–924. [\[CrossRef\]](#) [\[PubMed\]](#)

26. Van Smeden, M.; Naaktgeboren, C.A.; Reitsma, J.B.; Moons, K.G.M.; de Groot, J.A.H. Latent class models in diagnostic studies when there is no reference standard—A systematic review. *Am. J. Epidemiol.* **2014**, *179*, 423–431. [[CrossRef](#)]
27. Haaksma, M.L.; Calderón-Larrañaga, A.; Olde Rikkert, M.G.M.; Melis, R.J.F.; Leoutsakos, J.M.S. Cognitive and functional progression in Alzheimer disease: A prediction model of latent classes. *Int. J. Geriatr. Psychiatry* **2018**, *33*, 1057–1064. [[CrossRef](#)]
28. Wiegand, R.E.; Cooley, G.; Goodhew, B.; Bannietts, N.; Kohlhoff, S.; Gwyn, S.; Martin, D.L. Latent class modeling to compare testing platforms for detection of antibodies against the Chlamydia trachomatis antigen Pgp3. *Sci. Rep.* **2018**, *8*, 4232. [[CrossRef](#)]
29. Hulsen, T.; Jamuar, S.S.; Moody, A.R.; Karnes, J.H.; Varga, O.; Hedensted, S.; Spreafico, R.; Hafler, D.A.; McKinney, E.F. From Big Data to Precision Medicine. *Front. Med.* **2019**, *6*, 34. [[CrossRef](#)]
30. Cawley, G.C.; Talbot, N.L.C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.
31. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [[CrossRef](#)]
32. Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J. R. Stat. Soc. Ser. B* **1974**, *36*, 111–147. [[CrossRef](#)]
33. Tao, K.; Bian, Z.; Zhang, Q.; Guo, X.; Yin, C.; Wang, Y.; Zhou, K.; Wane, S.; Shi, M.; Bao, D.; et al. Machine learning-based genome-wide interrogation of somatic copy number aberrations in circulating tumor DNA for early detection of hepatocellular carcinoma. *EBioMedicine* **2020**, *56*, 102811. [[CrossRef](#)] [[PubMed](#)]
34. Hong, X.; Mitchell, R.J. Backward elimination model construction for regression and classification using leave-one-out criteria. *Int. J. Syst. Sci.* **2007**, *38*, 101–113. [[CrossRef](#)]
35. Shao, J. Linear model selection by cross-validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486–494. [[CrossRef](#)]
36. Geisser, S. The predictive sample reuse method with applications. *J. Am. Stat. Assoc.* **1975**, *70*, 320–328. [[CrossRef](#)]
37. Xie, J.; Qiu, Z. Bootstrap technique for ROC analysis: A stable evaluation of Fisher classifier performance. *J. Electron.* **2007**, *24*, 523–527. [[CrossRef](#)]
38. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [[CrossRef](#)]
39. Torre, L.A.; Bray, F.; Siegel, R.L.; Ferlay, J.; Lortet-Tieulent, J.; Jemal, A. Global cancer statistics, 2012. *CA Cancer J. Clin.* **2015**, *65*, 87–108. [[CrossRef](#)]
40. Ferlay, J.; Soerjomataram, I.; Dikshit, R.; Eser, S.; Mathers, C.; Rebelo, M.; Parkin, D.M.; Forman, D.; Bray, F. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **2015**, *136*, 359–386. [[CrossRef](#)]
41. Li, L.; Zhang, L.; Tian, Y.; Zhang, T.; Duan, G.; Liu, Y.; Yin, Y.; Hua, D.; Qi, X.; Mao, Y. Serum Chemokine CXCL7 as a Diagnostic Biomarker for Colorectal Cancer. *Front. Oncol.* **2019**, *9*, 921. [[CrossRef](#)]
42. Pine, J.K.; Morris, E.; Hutchins, G.G.; West, N.P.; Jayne, D.G.; Quirke, P.; Prasad, K.R. Systemic neutrophil-to-lymphocyte ratio in colorectal cancer: The relationship to patient survival, tumour biology and local lymphocytic response to tumour. *Br. J. Cancer* **2015**, *113*, 204–211. [[CrossRef](#)] [[PubMed](#)]
43. Azab, B.; Mohammad, F.; Shah, N.; Vonfrolio, S.; Lu, W.; Kedia, S.; Bloom, S.W. The value of the pretreatment neutrophil lymphocyte ratio vs. platelet lymphocyte ratio in predicting the long-term survival in colorectal cancer. *Cancer Biomark.* **2014**, *14*, 303–312. [[CrossRef](#)] [[PubMed](#)]
44. Li, Y.; Wu, H.; Xing, C.; Hu, X.; Zhang, F.; Peng, Y.; Li, Z.; Lu, T. Prognostic evaluation of colorectal cancer using three new comprehensive indexes related to infection, anemia and coagulation derived from peripheral blood. *J. Cancer* **2020**, *11*, 3834–3845. [[CrossRef](#)] [[PubMed](#)]
45. Guo, Y.H.; Sun, H.F.; Zhang, Y.B.; Liao, Z.J.; Zhao, L.; Cui, J.; Wu, T.; Lu, J.R.; Nan, K.J.; Wang, S.H. The clinical use of the platelet/lymphocyte ratio and lymphocyte/monocyte ratio as prognostic predictors in colorectal cancer: A meta-analysis. *Oncotarget* **2017**, *8*, 20011–20024. [[CrossRef](#)] [[PubMed](#)]
46. Xia, L.J.; Li, W.; Zhai, J.C.; Yan, C.W.; Chen, J.B.; Yang, H. Significance of neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio, lymphocyte-to-monocyte ratio and prognostic nutritional index for predicting clinical outcomes in T1–2 rectal cancer. *BMC Cancer* **2020**, *20*, 208. [[CrossRef](#)] [[PubMed](#)]

47. Stojkovic, L.M.; Pavlovic, M.A.; Stankovic, S.; Stojkovic, M.; Dimitrijevic, I.; Radoman, V.I.; Lalic, D.; Milovanovic, T.; Domic, I.; Krivokapic, Z. Combined Diagnostic Efficacy of Neutrophil-to-Lymphocyte Ratio (NLR), Platelet-to-Lymphocyte Ratio (PLR), and Mean Platelet Volume (MPV) as Biomarkers of Systemic Inflammation in the Diagnosis of Colorectal Cancer. *Dis. Markers* **2019**, *2019*, 6036979. [\[CrossRef\]](#)
48. Haram, A.; Boland, M.R.; Kelly, M.E.; Bolger, J.C.; Waldron, R.M.; Kerin, M.J. The prognostic value of neutrophil-to-lymphocyte ratio in colorectal cancer: A systematic review. *J. Surg. Oncol.* **2017**, *115*, 470–479. [\[CrossRef\]](#)
49. Ofllazoglu, U.; Alacacioglu, A.; Somali, I.K.; Yuce, M.; Buyruk, M.A.; Varol, U.; Salman, T.; Taskaynatan, H.; Yildiz, Y.; Kucukzeybek, Y.; et al. Prognostic value of neutrophil/lymphocyte ratio (NLR), platelet/lymphocyte ratio (PLR) and mean platelet volume (MPV) in patients with colorectal carcinoma [Izmir OncologyGroup (IZOG) study]. *Ann. Oncol.* **2016**, *27*, 149–206. [\[CrossRef\]](#)
50. Ying, H.; Deng, Q.; He, B.; Pan, Y.; Wang, F.; Sun, H.; Chen, J.; Liu, X.; Wang, S. The prognostic value of preoperative NLR, d-NLR, PLR and LMR for predicting clinical outcome in surgical colorectal cancer patients. *Med. Oncol.* **2014**, *31*, 305. [\[CrossRef\]](#)
51. Liu, C.; White, M.; Newell, G. Measuring and comparing the accuracy of species distribution models with presence absence data. *Ecography* **2011**, *34*, 232–243. [\[CrossRef\]](#)
52. Powers, D.M.W. Evaluation: From Precision, Recall and F-Score to ROC, Informedness, Markedness & Correlation. *J. Mach. Learn. Tech.* **2011**, *2*, 37–63.
53. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. *Proceedings of Advances in Artificial Intelligence (AI 2006), Lecture Notes in Computer Science, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation*; Springer: Heidelberg, Germany, 2006; p. 4304.
55. Youden, W.J. Index for rating diagnostic tests. *Cancer* **1950**, *3*, 32–35. [\[CrossRef\]](#)
56. Hwang, Y.-T.; Hung, Y.-H.; Wang, C.C.; Terng, H.-J. Finding the optimal threshold of a parametric ROC curve under a continuous diagnostic measurement. *Revstat. Stat. J.* **2018**, *16*, 23–43.
57. Mitchell, A.J. Sensitivity $\times$ PPV is a recognized test called the clinical utility index (CUI+). *Eur. J. Epidemiol.* **2011**, *26*, 251–252. [\[CrossRef\]](#) [\[PubMed\]](#)
58. Rota, M.; Antolini, L. Finding the optimal cut-point for Gaussian and GAMMA distributed biomarkers. *Comput. Stat. Data Anal.* **2014**, *69*, 1–14. [\[CrossRef\]](#)
59. Martens, F.K.; Tonk, E.C.M.; Kers, J.G.; Janssens, A.; Cecile, J.W. Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks. *J. Clin. Epidemiol.* **2016**, *79*, 159–164. [\[CrossRef\]](#)
60. Walsh, S.R.; Cook, E.J.; Goulder, F.; Justin, T.A.; Keeling, N.J. Neutrophil-lymphocyte ratio as a prognostic factor in colorectal cancer. *J. Surg. Onco.* **2005**, *91*, 181–184. [\[CrossRef\]](#)
61. Dell'Aquila, E.; Cremolini, C.; Zeppola, T.; Lonardi, S.; Bergamo, F.; Masi, G.; Stellato, M.; Marmorino, F.; Schirripa, M.; Urbano, F.; et al. Prognostic and predictive role of neutrophil/lymphocytes ratio in metastatic colorectal cancer: A retrospective analysis of the TRIBE study by GONO. *Ann. Oncol.* **2018**, *29*, 924–930. [\[CrossRef\]](#)
62. Anuk, T.; Yildirim, A.C. Clinical Value of Platelet-to-Lymphocyte Ratio in Predicting Liver Metastasis and Lymph Node Positivity of Colorectal Cancer Patients. *Turk. J. Colorectal. Dis.* **2017**, *27*, 50–55. [\[CrossRef\]](#)
63. Peng, J.; Li, H.; Ou, Q.; Lin, J.; Wu, X.; Lu, Z.; Yuan, Y.; Wan, D.; Fang, Y.; Pan, Z. Preoperative lymphocyte-to-monocyte ratio represents a superior predictor compared with neutrophil-to-lymphocyte and platelet-to-lymphocyte ratios for colorectal liver-only metastases survival. *OncoTargets Ther.* **2017**, *27*, 3789–3799. [\[CrossRef\]](#)
64. Airola, A.; Pahikkala, T.; Waegeman, W.; De Baets, B.; Salakoski, T. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Comput. Stat. Data Anal.* **2011**, *55*, 1828–1918. [\[CrossRef\]](#)
65. Parker, B.J.; Gunter, S.; Bedo, J. Stratification bias in low signal microarray studies. *BMC Bioinform.* **2007**, *8*, 326. [\[CrossRef\]](#) [\[PubMed\]](#)
66. Molinaro, A.; Simon, R.; Pfeiffer, R. Prediction error estimation: A comparison of resampling methods. *Bioinformatics* **2005**, *21*, 3301–3307. [\[CrossRef\]](#) [\[PubMed\]](#)

67. Braga-Neto, U.M.; Dougherty, E.R. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* **2004**, *20*, 374–380. [[CrossRef](#)]
68. Varma, S.; Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **2006**, *7*, 91. [[CrossRef](#)]
69. Barlow, H.; Mao, S.; Khushi, M. Predicting High-Risk Prostate Cancer Using Machine Learning Methods. *Data* **2019**, *4*, 129. [[CrossRef](#)]
70. Perez, I.M.; Airola, A.; Bostrom, P.J.; Jambor, I.; Pahikkala, T. Tournament leave-pair-outcross-validation for receiver operating characteristic analysis. *Stat. Methods Med. Res.* **2019**, *28*, 2975–2991. [[CrossRef](#)]
71. Beam, A.L.; Kohane, I.S. Big data and machine learning in health care. *JAMA* **2018**, *319*, 1317–1318. [[CrossRef](#)]
72. Bolboacă, S.D. Assessment of Random Assignment in Training and Test Sets using Generalized Cluster Analysis Technique. *Appl. Med. Inform.* **2011**, *28*, 9–14.
73. Miller, R.; Siegmund, D. Maximally selected chi square statistics. *Biometrics* **1982**, *38*, 1011–1016. [[CrossRef](#)]
74. Liu, X. Classification accuracy and cut point selection. *Stat. Med.* **2012**, *31*, 2676–2686. [[CrossRef](#)] [[PubMed](#)]
75. Unal, I. Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach. *Comput. Math. Methods Med.* **2017**, *2017*, 3762651. [[CrossRef](#)] [[PubMed](#)]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).