ISSN 1064-5624, Doklady Mathematics, 2024, Vol. 110, Suppl. 1, pp. S62–S74. © The Author(s), 2024. This article is an open access publication.
Russian Text © The Author(s), 2024, published in Doklady Rossiiskoi Akademii Nauk. Matematika, Informatika, Protsessy Upravleniya, 2024, Vol. 520, No. 2, pp. 70–84.

Empirical Approach to Sample Size Estimation for Testing of AI Algorithms

M. R. Kodenko[a,b,*], T. M. Bobrovskaya[a,], R. V. Reshetnikov[a,***], K. M. Arzamasov[a,****], A. V. Vladzymyrskyy[a,*****], O. V. Omelyanskaya[a,******], and Yu. A. Vasilev[a,*******]**

Received September 30, 2024; revised October 2, 2024; accepted October 2, 2024

Abstract—Calculation of sample size is one of the basic tasks in the field of correct and objective testing of artificial intelligence (AI) algorithms. Existing approaches, despite their exhaustive theoretical justification, can give results that differ by an order of magnitude under the same initial conditions. Most of the input parameters for such methods are determined by the researcher intuitively or on the basis of relevant literature data in the subject area. Such uncertainty at the research planning stage is associated with a high risk of obtaining biased results, which is especially important to take into account when using AI algorithms for medical diagnosis. Within the framework of this work, an empirical study of the value of the minimum required sample size of radiology diagnostic studies to obtain an objective value of the AUROC metric was conducted. An algorithm for calculating the threshold value of sample size according to the criterion of no statistically significant changes in the metric value in case of increasing this size was developed and implemented in software format. Using datasets containing the results of testing of AI algorithms on mammographic and radiographic studies with the total volume of more than 300 thousand, the empirical threshold for the sample size from 30 to 25 thousand studies with different relative content of pathology—from 10 to 90%—was calculated. The proposed algorithm allows obtaining results invariant to the balance of classes in the sample, the target value of AUROC, the modality of studies, and the AI algorithm. The empirical value of the minimum sufficient sample size for testing the AI algorithm for binary classification, obtained by analyzing over 2 million estimated values, is 400 studies. The results can be used to solve the problems of development and testing of diagnostic tools, including AI algorithms.

Keywords: radiology, sample size, artificial intelligence, testing, ROC, AUC

DOI: 10.1134/S1064562424602063

1. INTRODUCTION

The use of artificial intelligence (AI) algorithms to solve medical problems is associated with high requirements for the objectivity of its results. One of the key aspects of obtaining reliable test results is determining the optimal test sample size [1].

There are many approaches to calculating the minimum sufficient sample size, differing depending on the purpose (descriptive or evidential), study design (for instance, parallel or cross-over, equality, noninferiority, superiority, or equivalence), the number of groups compared, etc. The calculation results also depend largely on the input values, the determination of which is a nontrivial task requiring population studies, which are available for a very limited range of tasks. However, as shown by Norman and Salama [2], even the presence of reasonable inputs can lead to significant differences in results, depending on the chosen calculation method, determined by the target metric. In this paper we examine the descriptive problem of estimating the value of the diagnostic accuracy metric of the AI algorithm for the binary (norm or pathology) classification of diagnostic imaging studies. The metric under study is AUROC (area under receiving operating characteristic curve), which integrally reflects the diagnostic characteristics (sensitivity (Se) and specificity (Sp)) of the algorithm [3]. When estimating the sample size within the framework of the task, the key factors are: errors of the first and second kind, the AUROC value for the null hypothesis, as well as the relative proportion of the target feature in the sample (the so-called prevalence), determined

[a] *Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow Health Care Department, Moscow, Russia*

[b] *Bauman Moscow State Technical University, Moscow, Russia*

**e-mail: mrkodenko@yandex.ru*

***e-mail: BobrovskayaTM@zdrav.mos.ru*

****e-mail: ReshetnikovRV1@zdrav.mos.ru*

*****e-mail: ArzamasovKM@zdrav.mos.ru*

******e-mail: VladzimirskijAV@zdrav.mos.ru*

*******e-mail: OmelyanskayaOV@zdrav.mos.ru*

********e-mail: npcmr@zdrav.mos.ru*

from population data. The result is determined in the course of solving the problem of obtaining AUROC with the desired accuracy or the problem of achieving the desired effect size [4, 5]. The methods are adapted for practical use in the format of online calculators. The methods are united by their theoretical validity, while the issue of the practical applicability of the results traditionally remains unaddressed.

In this paper, an empirical study was conducted to determine the threshold value of the minimum sample size required to obtain an objective AUROC value.

2. MATERIALS AND METHODS

The empirical determination of the threshold sample size that is minimal and sufficient to obtain an objective AUROC value consists of analyzing the results of testing AI algorithms on samples of different sizes. The following hypothesis is postulated as a criterion for the sufficiency of the sample size: we regard the sample size as minimal and sufficient if its further increase does not lead to statistically significant changes in the value of the metric. Additionally, we investigate the behavior of the Se and Sp of an AI algorithm for binary classification of radiation studies. Figure 1 presents the algorithm for checking the fulfillment of the criterion.

At the data preprocessing stage, the initial sample is divided into two basic subsamples based on the classification of the studies into the norm and pathology classes according to the radiological conclusion. Next, the balance of classes k is assigned in the range from 10 to 90%, and the sample size n for testing the AI algorithm is prescribed in the range from 30 to 25 000 in increments of 10. The first stage calculations contain two consecutive operations:

1. For the selected combination of k and n, from the basic subsamples we randomly select $k \cdot n$ studies of the pathology class and $(1-k)\cdot n$ studies of the norm class. We repeat the operation 100 times, which results in 100 subsamples of size n, each of which contains k % of studies of the pathology class, which corresponds to 100-fold repetition of the experiment for a given combination of n and k.

2. We test the AI algorithm on each of the 100 subsamples and record the Se, Sp, and AUROC metrics. As a result of the calculations, we obtain a matrix of size 100×3, that is, 100 values of each metric obtained from a sample of size n with a share of pathological cases k.

We repeat actions under paragraphs 1 and 2 for each possible combination of n and k. The number of iterations of each of the loops in n and k is, respectively,

$$\frac{(N-n_1)}{\text{step}}+1=\frac{25000-30}{10}+1=2498, \qquad (1)$$

$$\frac{(K-k_1)}{\text{step}}+1=\frac{90-10}{10}+1=9. \qquad (2)$$

The result of the calculations is a matrix of size

$$(100\cdot 2498\cdot 9)\times 3 = 2\,248\,200\times 3. \qquad (3)$$

We implemented the presented algorithm for data selection and metric calculation as a program code in Python (version 3.9) [6]. The calculation of Se and Sp is carried out on the basis of contingency tables. False negatives are cases that are positive according to the opinion of the radiologist, but negative according to the results of processing the study by the AI algorithm. The AUROC calculation was performed using the roc_auc_score() function of the sklearn.metrics package [7].

At the calculation stage, we use an iterative one-way implementation of the k-NN analysis method [8] for each sample of the obtained metric values for each balance of classes k. Iteratively, moving from the beginning to the end of the sample, we determine the main reference subsample, corresponding to some n, and select the 15 following samples as the subsamples to compare with. The number of neighbors equal to 15 corresponds to an increase in size n by 150 studies. Next, we carry out a sequential comparison of the means and variance of the main subsample with its neighbors. The choice of comparison methods, parametric or nonparametric, is determined by the results of the statistical test of the compared samples for normality. Simultaneous comparison of means and variance is important, because comparison of means may be biased without taking into account the variance in the data: there are known cases of no differences in means despite differences in variance [9], which, in the context of this work, may lead to an incorrect interpretation of the results. We interpret the results of statistical comparison of samples from the point of view of accepting or rejecting the hypothesis of the absence of statistically significant differences: the total number of neighbors for which the hypothesis is accepted simultaneously for both the variance and the means is recorded. The result of the calculations at the second stage is a sample containing the total number of statistically not significantly different neighboring samples x (*redundant* neighbors), corresponding to each combination of n and k. A 100-fold repetition of calculation for each combination of n and k allows forming a sample of metric values that is sufficient, according to the Altman nomogram, to conduct statistical comparisons at a significance level of 0.05 and 80% of power of statistical tests.

At the final stage, we determine the critical value n_{cr} for which we approximate the obtained dependences $x(n)$ for any k using smooth curves by means of the loss function minimization method [10]. Next, we cut off the minimum number of redundant neighbors by $x = 10$, which corresponds to an increase in sample size n per 100 studies. The minimum abscissa of a point

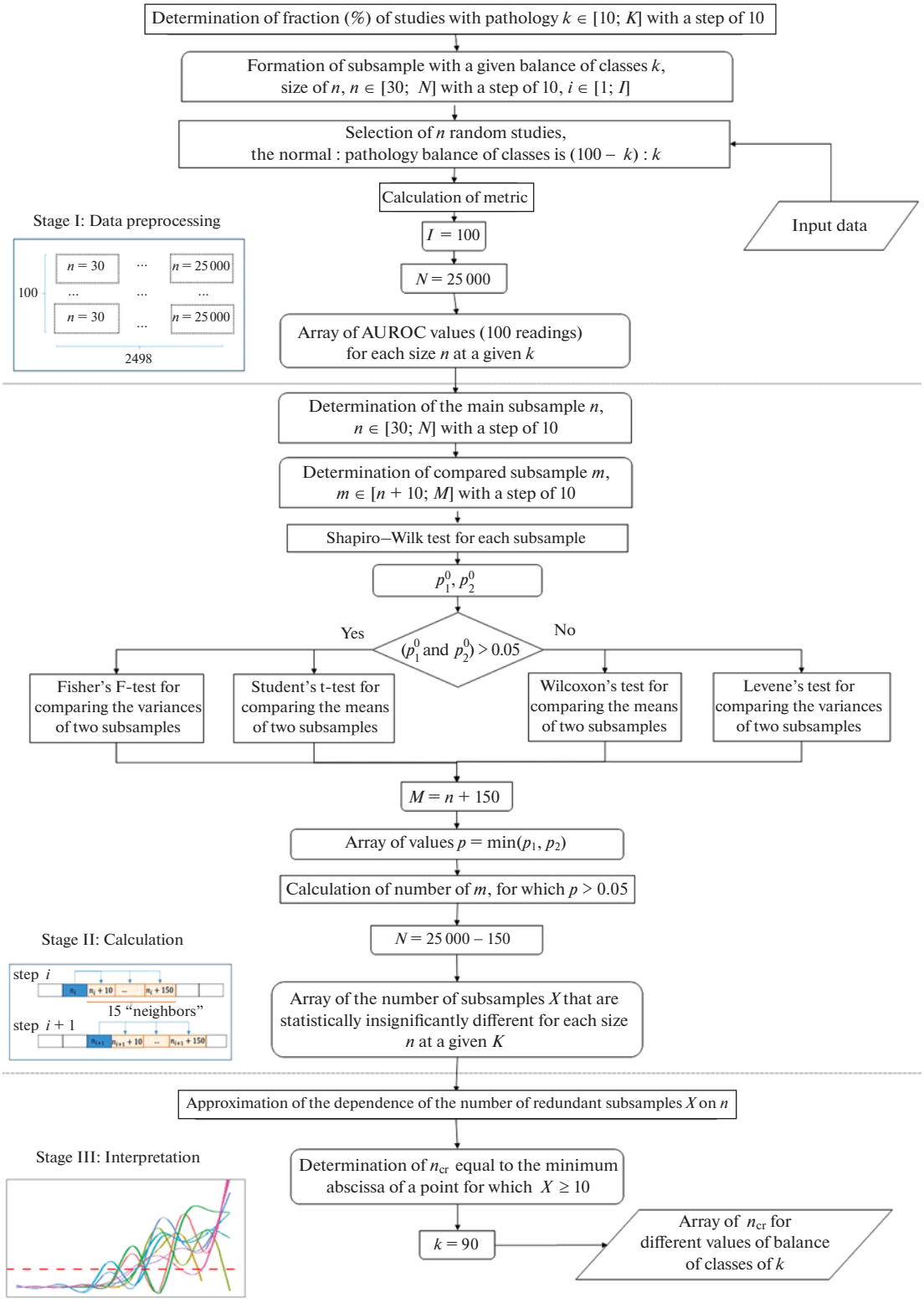

Fig. 1. Scheme of operation of the proposed method of empirical sample size estimation.

whose ordinate satisfies the criterion $x \geq 10$ is taken as the desired value n_{cr}: increasing sample size n_{cr} by a value from 1 to 100 studies does not lead to the emergence of statistically significant changes in the metric for a given k.

To determine the type of dependence n_{cr} from the class balance, we carried out an additional analysis of the results in the area including all values of 95% confidence intervals (CI) n_{cr} for all class balances: all metric values obtained with $n \in [\min n_{cr}; \max n_{cr}]$. We used the parametric Pearson criterion or the nonparametric Spearman criterion and a statistical comparison of groups corresponding to different k for each of the three algorithms and applied the ANOVA or Kruskal–Wallace with Tukey or Dunn's post hoc test, respectively, to identify different groups. To objectify the results, we performed subgroup analysis with Bonferroni correction for multiple comparisons [11].

Statistical calculations were performed using the R language (version 4.2.1) [12]. The choice of statistical methods is based on the distribution of data in the compared samples. We determined the type of data distribution (normal or nonnormal) using the Shapiro–Wilk test (shapiro.test() [13]). Depending on the distribution, we selected Student's t-test (t.test() [14]) or its nonparametric analogue, the Wilcoxon test for independent samples (wilcox.test() [15]) to compare mean values. We similarly carried out comparison of variance using Fisher's parametric test (var.test() [16]) or Levene's nonparametric test (leveneTest() [17]). To plot the curve, we used the geom_smooth() method of the ggplot2 package [10] with minimization of the loss function and assigned the number of points equal to the size of the approximated sample. We performed correlation analysis using the cor.test() function [18] of the basic R package. For variance analysis, we used the pairs of functions aov() [19] and TukeyHSD() [20] of the stats package, kruskal.test() of the stats package [21] and dunn.test() [22] of the like-named package. The significance level for accepting statistical hypotheses was chosen to be 0.05 [23].

For practical testing of the proposed method, we used the results of testing three different AI algorithms participating in the "Experiment on the use of innovative technologies in the field of computer vision for the analysis of medical images and further application in the healthcare system of the city of Moscow" [24], approved by the Independent Ethics Committee and registered on the ClinicalTrials resource (NCT04489992). We compared the calculation results n_{cr} between three AI algorithms.

We used the data sets (DSs) listed below:

1. Mammographic examinations (MMGs) were classified by the presence (pathology) and absence (normal) of signs of malignant neoplasms of the mammary gland according to the Bi-RADS scale [25]: classes 1 and 2 were classified as normal, and classes 3–5 were interpreted as pathology:

1.1 DS containing 143 710 studies with the results of the AI algorithm A1, obtained for the period from February 1, 2022 to October 31, 2022.

1.2 DS containing 123 301 studies with the results of the AI algorithm A2, obtained for the period from September 1, 2021 to December 27, 2021.

2. Radiographic (XR) studies were classified according to the presence (pathological) or absence (normal) of at least one of the following features: pleural effusion, pneumothorax, occult focus, infiltration, consolidation, dissemination, cavity, atelectasis, calcification, mediastinal widening, cardiomegaly, or disruption of the integrity of the cortical layer. DS containing 62 142 studies with the results of the AI algorithm A3, obtained for the period from October 25, 2023 to November 21, 2023.

Each DS is presented in xlsx format as a table containing the results of the processing the study by the AI algorithm and the conclusion of a radiologist in binary form (0 for normal and 1 for pathology).

Sample size reference values were calculated in two ways, with the proportion of pathological cases in the sample ranging from 0.1 to 0.9 and the target AUROC value being set to 0.57 for MMG and 0.7 for XR (Fig. 2):

1. Sample size calculator [26], which allows calculating the sample size by achieving the effect size for a given test power, significance level, and class balance:

— type I error 0.05 [23];

— test power 0.8 [28];

— the null hypothesis corresponds to a random response of AUROC of 0.5 [23].

2. Sample size–confidence interval for AUROC [27], which allows estimating the sample size required to obtain AUROC with the accuracy specified by the confidence interval:

— confidence level 0.95 [23];

— the width of the confidence interval is 0.1 [29].

Additionally, we compared the values of the metric for empirical n_{cr} and conditionally maximally achievable $n = N = 25000$. We carried out statistical comparison of the samples similarly to the method described above.

3. RESULTS

Initially, we analyzed the dependence of the sample mean and median values of each of the three metrics AUROC, Se, and Sp on the proportion of pathological studies in the sample (k). For all three algorithms, these values coincide (Fig. 2a). Figure 2b presents the general view of the dependence of the mentioned metrics, as well as their variances, on the sample size using the example of data from algorithm A1. The dependence for all three metrics has a similar

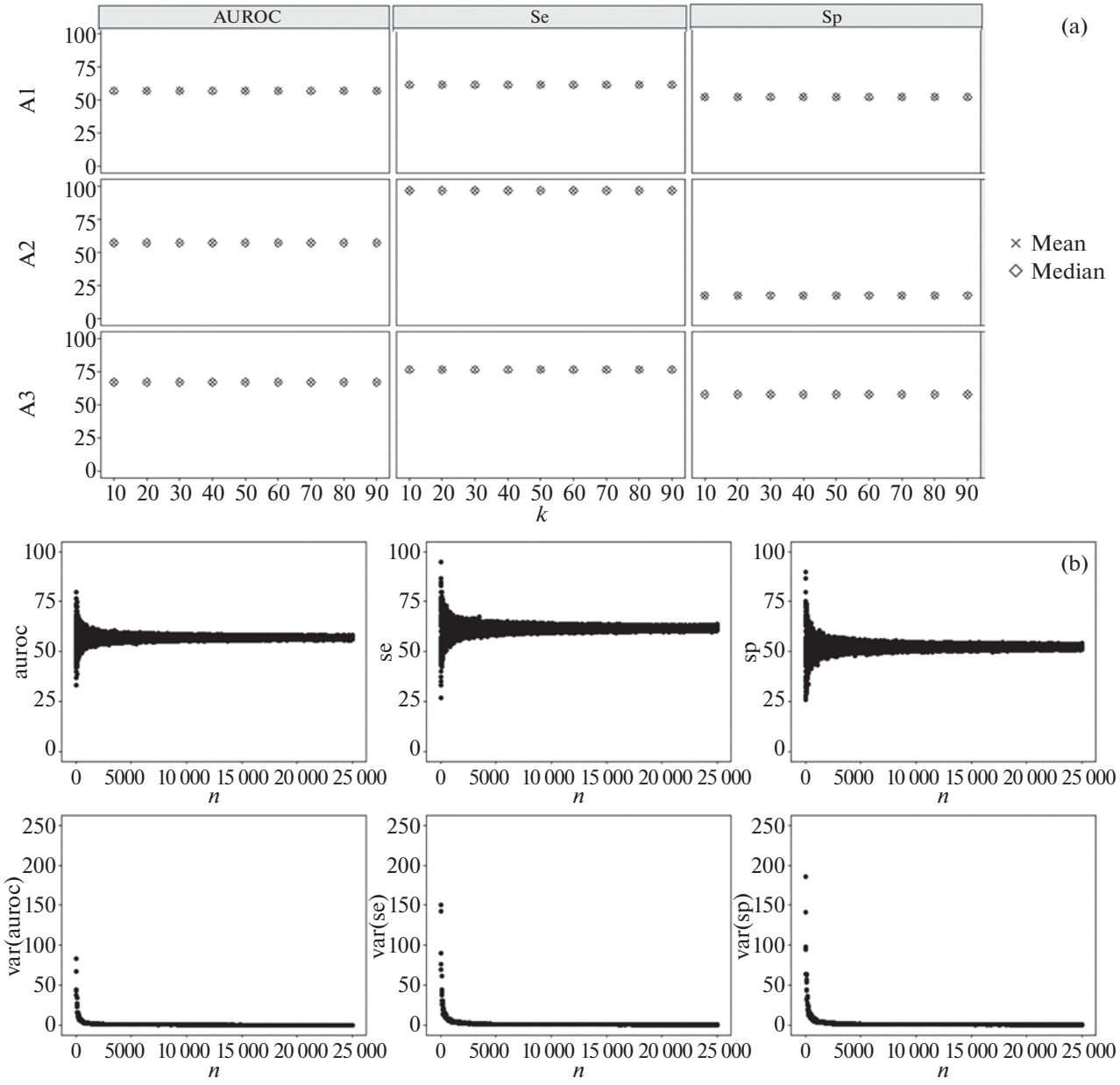

Fig. 2. (a) Sample means and medians of AI algorithm metrics over k. (b) General view of the dependence of metrics and their variances on n on example of operation of the AI algorithm A1 ($k = 50\%$).

form of symmetrical damped oscillation: as n increases, the amplitude of the spread of values decreases, reaching a certain conditionally stable range. The type of dependence of the variance on n (Fig. 2b, bottom row) confirms the advisability of taking into account the variance of subsamples when comparing means. According to the obtained data, the expected average AUROC value for AI algorithms A1 and A2 is 57%, and it is 70% for A3.

Figure 3 shows the results of smoothing approximation for all k of all three metrics for each AI algorithm. The dependence of the number of redundant neighbors x on the sample size n is shown. Also shown is the cutoff line by which the value of n_{cr} is determined for each class. According to the criterion of simultaneous comparison of means and variance of the metric, the number x increases with the growth of the sample size on which the given metric was obtained. For ease of visualization, the maximum sample size is limited to $n = 12000$.

The curves have a similar appearance regardless of the class balance: a linearly increasing section smoothly transitions into a region of small fluctuations near the plateau. The cutoff line corresponding

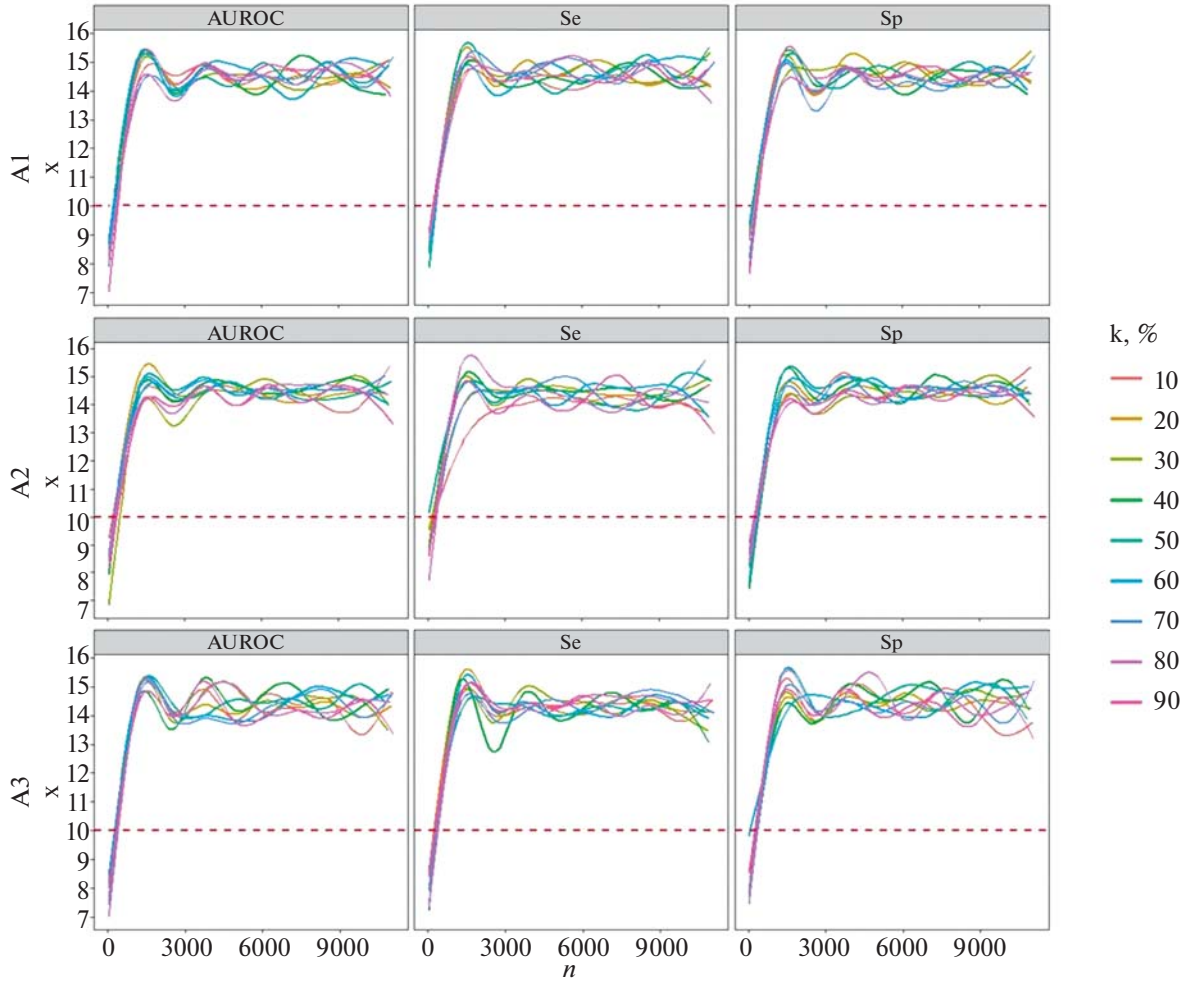

Fig. 3. Approximation of dependence $x(n)$ for different k. The red dotted line denotes the cutoff level $x = 10$.

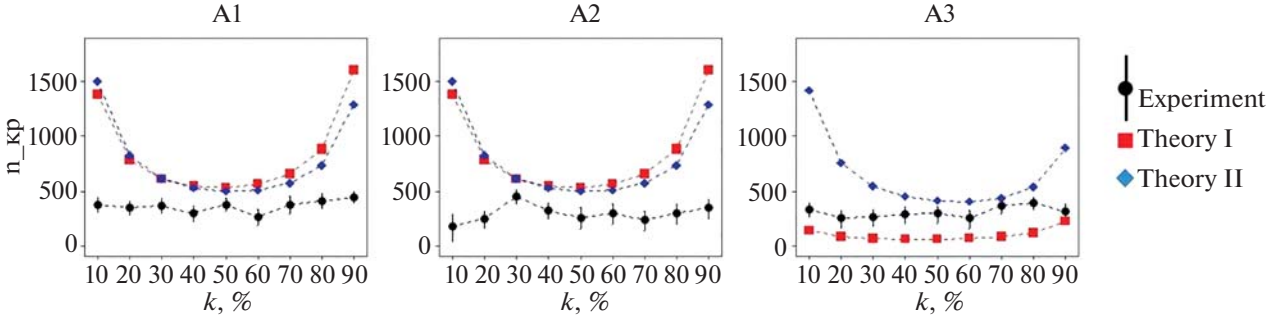

Fig. 4. Calculation of n_{er} using theoretical estimates of precision (red) and effect size (blue) of the AUROC value and empirical estimate (black) with 95% CI.

to 10 redundant neighbors x intersects the graphs on a conditionally linear section.

The theoretical estimate of the sample size (Fig. 4) for algorithms A1 and A2 (expected AUROC value is 57%) has a similar form for both calculation methods: a U-shaped curve with the minimum value for class balance 1:1. For algorithm A3 (expected AUROC value is 70%), the type of dependence differs significantly for the first and second calculation methods. The empirical dependence of the sample size on the class balance demonstrates the absence of a clear dependence on k.

Table 1. Result of subgroup analysis of significance n_{cr} for different class balances[1]

Algorithm	Metric	The result of the n_{cr} vs. k correlation analysis (Spearman's rho)	Result of statistical analysis of groups (Kruskal–Wallace test)	n_{cr} (95% CI)
A1	**AUROC[2]**	**–0.08 ($p = 0.19$)**	**$p = 0.04$***	**365 (324; 407)**
	Se	0.02 ($p = 0.73$)	$p = 0.60$	328 (298; 357)
	Sp	–0.12 ($p = 0.03$*)	$p = 0.31$	328 (280; 375)
A2	**AUROC**	**–0.03 ($p = 0.45$)**	**$p = 0.78$**	**297 (235; 358)**
	Se	–0.02 ($p = 0.69$)	$p = 0.85$	214 (147; 282)
	Sp	0.07 ($p = 0.16$)	$p = 0.60$	314 (265; 364)
A3	**AUROC[3]**	**–0.17 ($p = 0.004$*)**	**$p = 0.03$***	**312 (274; 350)**
	Se	–0.08 ($p = 0.15$)	$p = 0.63$	312 (274; 350)
	Sp	0.04 ($p = 0.43$)	$p = 0.2$	296 (225; 366)

[1]Bold indicates the value of the metric available for calculating reference values,

[2]Dunn's post hoc test with Bonferroni correction for multiple comparisons showed statistically significant differences only for classes $k = 10$ and 30%.

[3]Dunn's post hoc test with Bonferroni correction for multiple comparisons showed statistically significant differences only for classes $k = 20$ and 80%.

*Statistically significant values are marked,

As can be seen from the graphs in Fig. 4, the theoretically calculated values n_{cr} exceed empirical ones for all k for the first two AI algorithms. For the third algorithm, this condition is not met when using the calculation based on the effect size estimation criterion: because the expected AUROC value (70%) is significantly larger than the value for the null hypothesis (50%), the minimum sufficient sample sizes turned out to be small.

Table 1 presents the results of subgroup analysis to assess the relationship between the sample size n_{cr} and the share of pathology k for each algorithm. Tables 2 and 3 detail the calculated values. Figures 5 and 6 show the details of the construction of the smoothing curves presented in Fig. 3.

Because the empirically obtained values of n_{cr} demonstrate invariance to k, for each metric we compared three samples (100 values each) corresponding to the selected $n \geq n_{cr}$. The intergroup analysis of the neighbor samples in the AUROC analysis between the three algorithms for all class balances gives the following result: $p = 0.8$ (Kruskall–Wallis test), which corresponds to the absence of statistically significant differences and allows unifying the 95% CI for n_{cr} up to (245; 398) with an average value of 322. It makes sense to estimate the boundary size at the upper limit of the interval, rounding to the nearest value convenient for compiling samples with different class balances: 400 studies.

Comparison of AUROC values obtained at $n = 400$ and $n = N = 25000$ for all three AI algorithms demonstrates no statistically significant differences at any k (the minimum value of $p = 0.08$). This means that the AUROC obtained with an empirical sample size of 400 studies is not statistically significantly different from the value obtained with a sample size of 25000 studies, regardless of the pathology content proportion (from 10 to 90%), AI algorithm (A1, A2 or A3), and the type of radiographic examination (MMG or XR). In terms of absolute values, we are talking about the change in the average AUROC between the compared groups ($n = 400$ and $n = 25000$) in the third decimal place, which, in addition to the demonstrated lack of statistical significance, also has no practical significance.

4. DISCUSSION

Sample size calculation is traditionally performed at the research planning stage and precedes the experiment. Moreover, the result of estimating this size largely depends on the formulation of the null hypothesis and the chosen method [30]. Classical approaches to calculating sample size also use parameters that are selected by the researcher based on the population data (proportion of pathology in the sample), declared during testing (expected AUROC value) and on the literature data (significance level, test power, and CI width). Differences in approaches often lead to incomparability of the results. For example, based on the calculation results for the same balance of classes (norm-to-pathology is 1:1), the sample size according to the method of achieving 10% CI [27] for the expected AUROC of 99% is 18 studies, and for the expected AUROC of 51% it is as large as 513 studies; while, according to the method of estimating the effect size [26], these values are 40 and 26166 studies, respectively. It is obvious that if the use of classical, statistically based approaches leads to such incomparable results, there is a need to search for alternative approaches that allow objectifying the results.

Table 2. Results of sample size calculation for MMG[4]

k (%)	10	20	30	40	50	60	70	80	90
Theoretical sample size for effect size AUROC	1376	790	614	548	536	569	662	884	1599
Theoretical sample size for accuracy AUROC	1495	827	619	532	501	513	574	739	1286
A1									
Empirical sample size for AUROC (95% CI)	**380 (300 ; 450)**	**350 (280 ; 410)**	**370 (300 ; 440)**	**300 (230 ; 370)**	**380 (320 ; 440)**	**270 (190 ; 340)**	**380 (290 ; 460)**	**410 (340 ; 480)**	**450 (390 ; 500)**
Empirical sample size for Se (95% CI)	300 (190; 390)	310 (230; 370)	370 (290; 440)	320 (230; 400)	370 (310; 420)	360 (280; 420)	290 (200; 370)	360 (260; 440)	270 (160; 350)
Empirical sample size for Sp (95% CI)	270 (170; 350)	270 (190; 340)	300 (170; 390)	380 (310; 440)	370 (300; 440)	230 (120; 310)	390 (310; 450)	340 (240; 420)	400 (340; 450)
A2									
Empirical volume for AUROC (95% CI)	**180 (40; 290)**	**250 (160; 320)**	**460 (390; 520)**	**330 (250; 400)**	**260 (160; 350)**	**300 (200; 390)**	**240 (140; 320)**	**300 (190; 390)**	**350 (240; 430)**
Empirical sample size for Se (95% CI)	200 (30; 380)	150 (30; 280)	200 (80; 280)	220 (110; 310)	30 (30; 160)	250 (140; 330)	290 (160; 380)	330 (260; 390)	260 (150; 340)
Empirical sample size for Sp (95% CI)	350 (250; 420)	260 (140; 350)	410 (330; 480)	360 (290; 420)	280 (200; 350)	380 (300; 450)	240 (140; 320)	230 (70; 350)	320 (210; 410)

[4]Bold indicates the value of the metric available for calculating reference values.

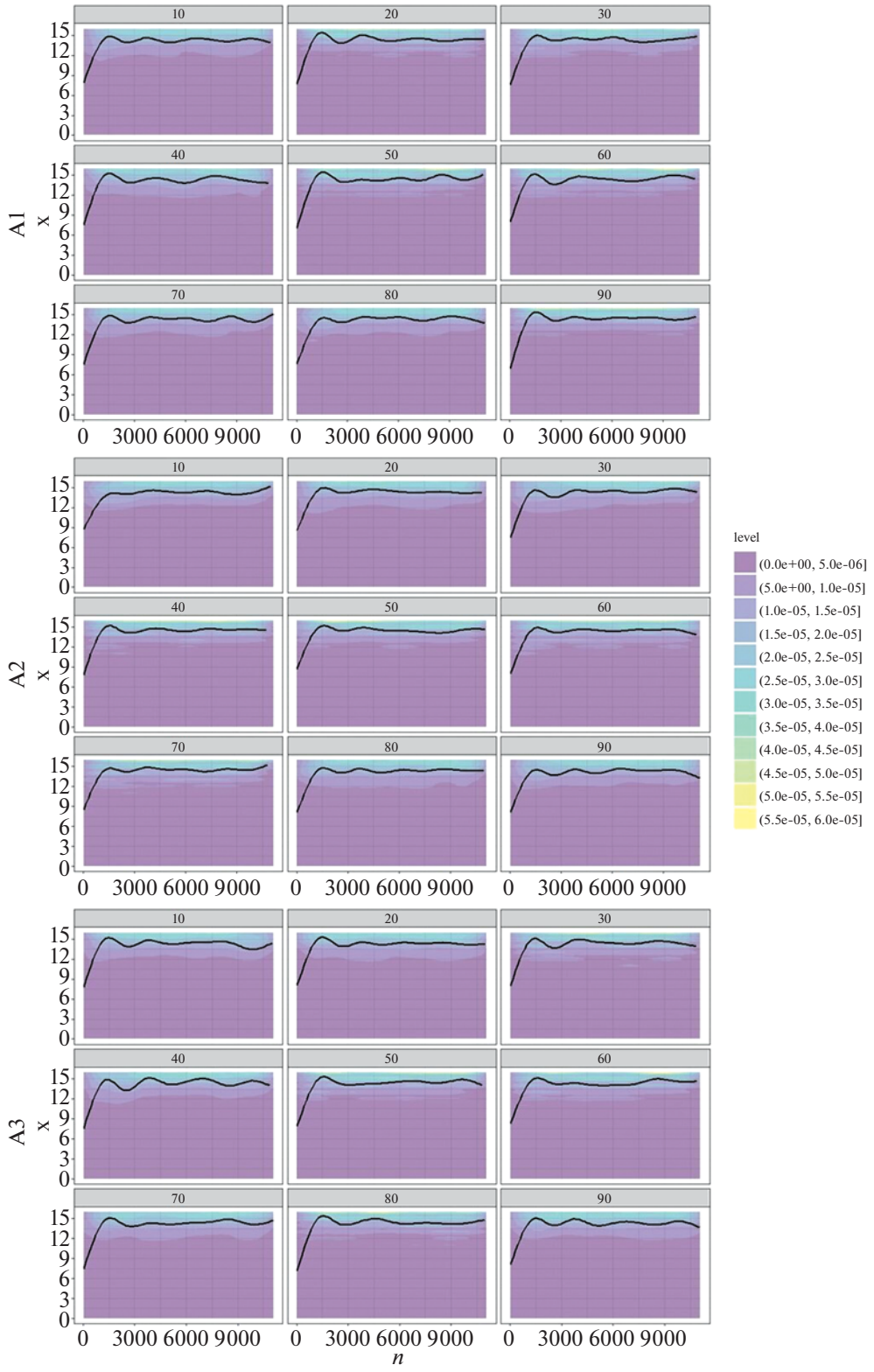

Fig. 5. Smoothing approximation of dependence $x(n)$ for various k: the density of distribution of experimental points (level) is mapped in color.

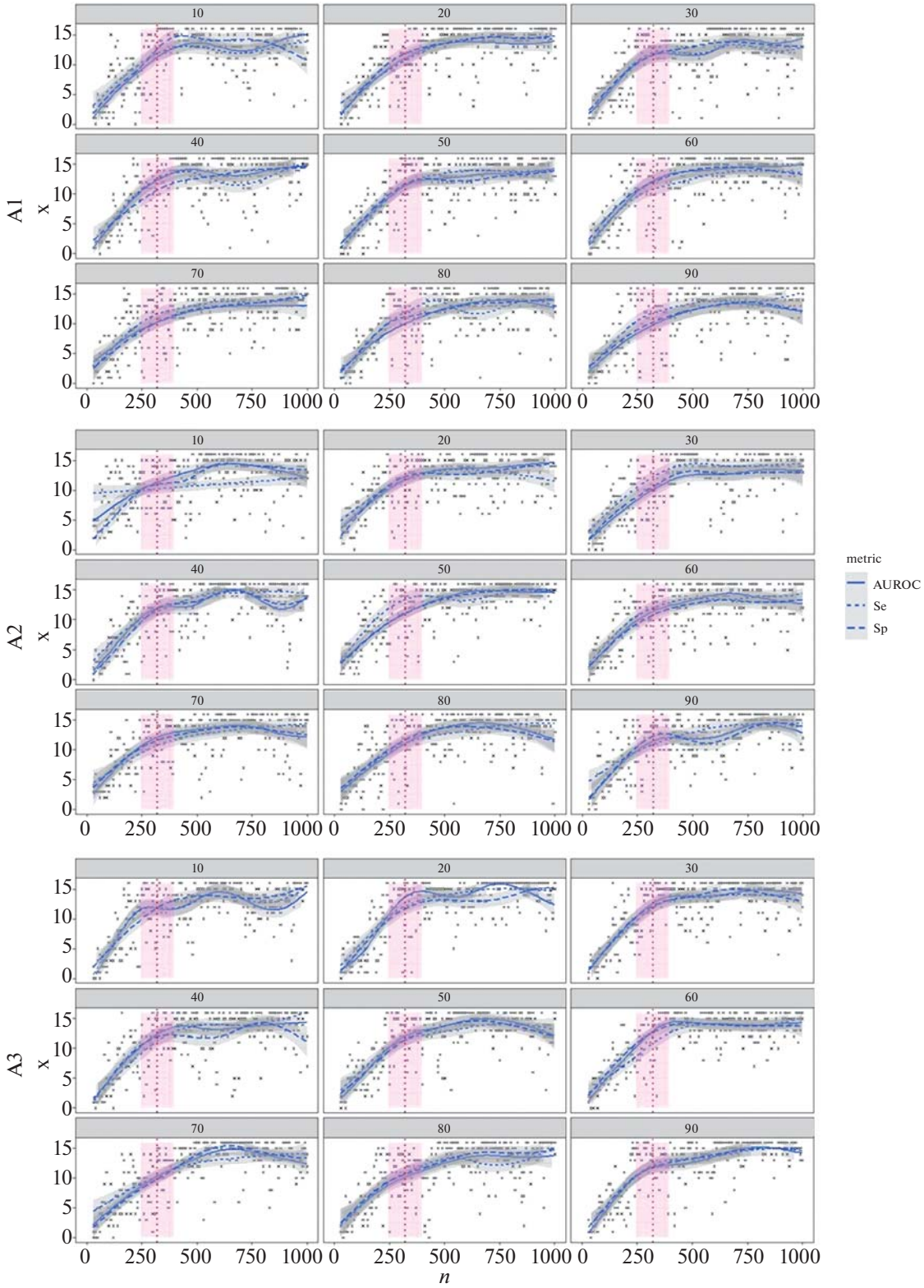

Fig. 6. Detail of dependence $x(n)$ for $n \in [30; 1000]$ at different k. The pink color shows the 95% CI for n_{cr}.

Table 3. Results of sample size calculation for RG[5]

k (%)	10	20	30	40	50	60	70	80	90
Theoretical sample size for effect size AUROC	145	8837	72	67	68	75	90	123	128
Theoretical sample size for accuracy AUROC	1413	759	551	458	416	409	440	540	894
A3									
Empirical sample size for AUROC (95% CI)	**340 (260; 400)**	**260 (160; 330)**	**270 (180; 340)**	**290 (210; 360)**	**300 (210; 360)**	**260 (160; 330)**	**370 (290; 430)**	**400 (330; 450)**	**320 (230; 390)**
Empirical sample size for Se (95% CI)	280 (180; 360)	230 (140; 310)	320 (230; 380)	340 (280; 400)	290 (170; 380)	350 (280; 410)	340 (250; 410)	390 (320; 450)	270 (170; 340)
Empirical sample size for Sp (95% CI)	300 (220; 370)	350 (260; 420)	260 (160; 340)	370 (290; 440)	320 (250; 380)	80 (30; 230)	380 (300; 440)	340 (270; 400)	260 (160; 350)

[5]Bold indicates the value of the metric available for calculating reference values.

In this paper, we attempted to estimate the sample sizes "by reverse": we calculated the behavior of the diagnostic accuracy metric of several different AI algorithms on samples of different sizes and class balances from 30 to 25000 radiographic studies. We examined the results (more than 2 million values for each of the metrics) from the point of view of the presence of a certain sample size, upon reaching which further increase does not make a statistically significant contribution to the change in the metric itself. To implement the criterion, we developed an algorithm that represents a one-way iterative version of the nearest neighbor analysis (k-NN analysis): the metric values obtained for a certain sample size n are compared with values obtained by increasing n by a value from 10 to 150 in increments of 10 studies. The absence of statistically significant differences in the mean values and variances of metrics, reproducible for 10 neighboring samples, is considered necessary and sufficient. We substantiated the reliability of the results by multiple (100) repetitions of measurements at each experimental point, as well as by the actual invariance to the formulation of the null hypothesis. We implemented the algorithm in the format of program code, detailed presentation (Fig. 1) of which makes it available for reproduction and practical application in solving research problems of related profile.. Practical testing of the algorithm was carried out using the results of objective testing of AI algorithms under conditions of maximum approximation to real data. Empirical estimation of the minimum sample size yields results that differ significantly from theoretical ones, not only in their absolute value but also in the dependence on the input data.

First, the empirical minimum sample size is invariant to the effect size, that is, to the difference between the expected AUROC value and the conditional random response boundary (AUROC = 0.5). This allows determining the sample size without first setting the value of the metric that the testing procedure is aimed at obtaining. According to theoretical calculations, the higher the desired AUROC value, the smaller the sample size required; however, it is precisely for small sample sizes that a high risk of introducing systematic error is noted [31]. This is confirmed by the results of the analysis of real data: in Fig. 2a larger spread is characteristic of small samples; therefore, for a sample of formally sufficient size, we can obtain unreasonably over- or underestimated results.

Second, the empirical minimum sample size is invariant to the content of pathology in the sample. At first glance, this thesis contains a logical contradiction, because it is known that the imbalance of classes in a sample can lead to an unjustified overestimation of Se or Sp [31]. At the same time, the share of pathology (the so-called prevalence) is usually assigned based on population research data to achieve sample representativeness, that is, its maximum likelihood in relation to the general population. This is confirmed by the result of comparing the average values of AUROC samples obtained for different class balances (see Fig. 2).

Finally, an empirical minimum sample size can be determined regardless of the study modality and AI algorithm. Three values were obtained for the minimum required sample size: 365 95% CI (324; 407), 297 95% CI (235; 358), and 312 95% CI (274; 350) for A1, A2, and A3. The conducted intergroup analysis showed the absence of statistically significant differences, which allows combining the results and designating a single threshold for testing the AI algorithm as the upper limit of 95% CI, rounded to 400 studies.

Based on the results, we can conclude that, invariant to the formulated hypothesis, the balance of sample classes, and the expected value of the diagnostic metric, the minimum required sample size for testing the AI algorithm with a binary classifier is 400 studies. Further increase in sample size does not provide a statistically significant contribution to the value of the diagnostic metric of the algorithm's accuracy.

5. LIMITATIONS OF THE STUDY

The limitations of the study include

1. Testing the algorithm on three AI algorithms. It should be noted that the proposed method is based on the analysis of the data itself and does not have task-specific constants or ratios: the reproducibility of the results for various testing conditions and different modalities is demonstrated.

2. Setting the number of neighbors to be analyzed to 15. The limitation is related to the optimization of calculations. However, as shown in Fig. 3, the dependence graph reaches a plateau, the value of which is less than 15, which shows the possibility of limiting from above by the specified value. In addition, the result of the metrics comparison on n_{cr} and maximum n showed no statistically significant differences.

3. Setting the threshold for the sufficient number of redundant neighbors to 10. The value is chosen empirically, but it corresponds to an increase in sample size by 100 radiology examinations, which can be considered sufficient in the context of the work.

4. The size of the subsamples from which the calculated metric values were obtained ranges from 30 to 25000 in increments of 10 studies. The exit to a stable plateau is achieved (according to Fig. 3) already at $n = 10000$, the upper limit of the sample is more than twice this value. The lower limit of the sample was chosen to be less than the minimum recommended value of 50 studies per class [32]. The discrete increase in values is compensated by using a smoothing algorithm to estimate the boundary value, as well as to estimate the intervals from above.

6. CONCLUSIONS

We presented an approach to estimating the minimum sample size for testing binary outcome AI algorithms designed for the analysis of radiological studies. The proposed calculation algorithm demonstrates stable results that are invariant to class balance, expected metric values, and research modality. The empirical threshold calculated in the study is 400. The results may have practical implications for conducting research in the field of developing and testing AI algorithms and can be extrapolated to other metrics and areas of medical diagnostics.

FUNDING

This article was prepared by a team of authors within the framework of the R and D project "Development of software for automatic generation of data sets of CT studies of the cardiovascular system with contrast suppression for training and testing algorithms based on artificial intelligence" (EGISU no. 123031500002-1).

CONFLICT OF INTEREST

The authors of this work declare that they have no conflicts of interest.

REFERENCES

1. A. Homeyer et al., "Recommendations on compiling test datasets for evaluating artificial intelligence solutions in pathology," Mod. Pathol. **35** (12), 1759–1769 (2022). https://doi.org/10.1038/s41379-022-01147-y
2. G. Norman, S. Monteiro, and S. Salama, "Sample size calculations: Should the emperor's clothes be off the peg or made to measure?," BMJ. **345**, e5278 (2012). https://doi.org/10.1136/bmj.e5278
3. R. D. Riley et al., "Evaluation of clinical prediction models (part 3): Calculating the sample size required for an external validation study," BMJ **384**, e074821 (2024). https://doi.org/10.1136/bmj-2023-074821
4. J. C. Biesanz and S. M. Schrager, "Sample size planning with effect size estimates," unpublished manuscript (2017).
5. C. C. Serdar et al., "Sample size, power and effect size revisited: Simplified and practical approaches in preclinical, clinical and laboratory studies," Biochem. Med. **31** (1), 27–53 (2021). https://doi.org/10.11613/BM.2021.010502
6. Python Release Python 3.9.0, Python.org. https://www.python.org/downloads/release/python-390. Accessed August 17, 2024.

7. Roc_auc_score, scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.metrics.roc_auc_score.html. Accessed August 17, 2024.
8. P. K. Syriopoulos et al., "kNN classification: A review," Ann. Math. Artif. Intell. (2023). https://doi.org/10.1007/s10472-023-09882-x
9. L. S. Kao and C. E. Green, "Analysis of variance: Is there a difference in means and what does it mean?," J Surg. Res. **144** (1), 158–170 (2008). https://doi.org/10.1016/j.jss.2007.02.053
10. Smoothed conditional means—geom_smooth. https://ggplot2.tidyverse.org/reference/geom_smooth.html. Accessed August 17, 2024.
11. R. A. Armstrong, "When to use the B onferroni correction," Ophthalmic Physiol. Opt. **34** (5), 502–508 (2014). https://doi.org/10.1111/opo.12131
12. Posit. https://www.posit.co. Accessed August 17, 2024.
13. S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," Biometrika **52** (3–4), 591–611 (1965).
14. T.test function—Rdocumentation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test. Accessed May 7, 2022.
15. Wilcox.Test Function—Rdocumentation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/wilcox.test. Accessed May 7, 2022.
16. Var.test function—Rdocumentation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/var.test. Accessed August 17, 2024.
17. LeveneTest function—Rdocumentation https://www.rdocumentation.org/packages/car/versions/3.1-2/topics/leveneTest. Accessed August 17, 2024.
18. Cor.test function—Rdocumentation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test. Accessed August 17, 2024.
19. Aov function—Rdocumentation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aov. Accessed August 17, 2024.
20. TukeyHSD function—Rdocumentation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/TukeyHSD. Accessed August 17, 2024.
21. Kruskal.test function—Rdocumentation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kruskal.test. Accessed August 17, 2024.
22. Dunn.test function—Rdocumentation. https://www.rdocumentation.org/packages/dunn.test/versions/1.3.6/topics/dunn.test. Accessed August 17, 2024.
23. J. Miller and R. Ulrich, "The quest for an optimal alpha," PLoS One **14** (1), e0208631 (2019). https://doi.org/10.1371/journal.pone.0208631
24. *Computer Vision in Radiodiagnosis: The First Stage of the Moscow Experiment*, Ed. by Yu. A. Vasil'ev and A. V. Vladzimirskii (Izdatel'skie Resheniya, Moscow, 2022) [in Russian].
25. Y. Weerakkody, T. Manning, P. Lemos, et al., "Breast imaging-reporting and data system (BI-RADS)," Radiopaedia.org. https://doi.org/10.53347/rID-10003. Accessed August 17, 2024.
26. Sample size calculator (riskcalc.org). https://riskcalc.org/samplesize. Accessed August 17, 2024.
27. J. S. Kohn, "Sample size—confidence interval for AUROC," Sample Size Calculators. https://samplesize.net/sample-size-ci-for-auroc. Accessed August 17, 2024.
28. F. J. Dorey, "In brief: Statistics in brief: Statistical power: What is it and when should it be used?," Clin. Orthop. Relat. Res. **469** (2), 619–620 (2011). https://doi.org/10.1007/s11999-010-1435-0
29. A. Hazra, "Using the confidence interval confidently," J. Thorac. Dis. **9** (10), 4125 (2017). https://doi.org/10.21037/jtd.2017.09.14
30. C. B. Monti, F. Ambrogi, and F. Sardanelli, "Sample size calculation for data reliability and diagnostic performance: A go-to review," Eur. Radiol. Exp. **8** (1), 79 (2024). https://doi.org/10.1186/s41747-024-00474-w
31. R. Aggarwal et al., "Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis," NPJ Digital Med. **4** (1), 65 (2021). https://doi.org/10.1038/s41746-021-00438-z
32. A. Alwosheel, S. van Cranenburgh, and C. G. Chorus, "Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis," J. Choice Model. **28**, 167–182 (2018).