Journal of Biomedical Informatics 48 (2014) 193–204

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Methodological Review

Sample size estimation in diagnostic test studies of biomedical informatics

Karimollah Hajian-Tilaki *

Dept of Social Sciences and Health, Babol University of Medical Sciences, Babol, Iran

ARTICLE INFO

Article history:
Received 11 August 2013
Accepted 17 February 2014
Available online 26 February 2014

Keywords:
Diagnostic studies
Sample size
Sensitivity
Specificity
ROC analysis
Area under the curve

ABSTRACT

Objectives: This review provided a conceptual framework of sample size calculations in the studies of diagnostic test accuracy in various conditions and test outcomes.
Methods: The formulae of sample size calculations for estimation of adequate sensitivity/specificity, likelihood ratio and AUC as an overall index of accuracy and also for testing in single modality and comparing two diagnostic tasks have been presented for desired confidence interval.
Results: The required sample sizes were calculated and tabulated with different levels of accuracies and marginal errors with 95% confidence level for estimating and for various effect sizes with 80% power for purpose of testing as well. The results show how sample size is varied with accuracy index and effect size of interest.
Conclusion: This would help the clinicians when designing diagnostic test studies that an adequate sample size is chosen based on statistical principles in order to guarantee the reliability of study.

1. Introduction

Biomedical informatics deals with health information, its structure, acquisition and use in health care and medical practices [1]. It includes health research, education, and health services, clinical disciplines and health care and information systems that ranging from theoretical model to the building and evaluation of applied diagnostic systems [1]. A specific of its domain is the evaluation of biomarkers and diagnostic systems for classification of diseased from healthy subjects to make a decision in clinical practices [1–4]. The bioinformatics markers/systems in medicine have been developed progressively during the past decades. Primarily, in order to apply the new developed biomarkers/systems for decision in clinical practices and to extract extra useful information from them, they should be evaluated in experimental setting versus a gold standard [4–6]. The collecting experimental data of gold standard is often expensive and time consuming. Thus, the evaluation of classification performance of bioinformatics systems needs annotated training sample since the predictive power in detection difference between two alternative classifiers strongly depends on sample size [7,8]. In estimating the diagnostic accuracy and to obtain a desired level of statistical power to detect an effect size for testing a single modality or a comparative study of two diagnostic tasks in order to know which has a greater diagnostic ability of certain target classification performance, the investigators need to know the minimal sample size required for their experiments. On the other hand, if the experiments are done on with available samples only, the investigators need to know the power of statistical test for detection a desirable effect size in their experiments.

However, the receiver operating characteristic (ROC) analysis used in diagnostic medicine for continuous biomarkers in classification is rather complex with no single approach for analysis [3–6]. Also, there is no single measure of accuracy index in evaluating diagnostic tools for decision support and how to estimate the sample size needed for proposed experiment. All depend on specific application and design used in biomarker experiments. This paper provides the sample size calculation for estimating and testing of accuracy indexes. We included different accuracy indexes with various types of experimental design for single diagnostic test and comparative study of two diagnostic tasks both independent design and matched paired design. First we briefly described some examples of relevant biomedical informatics research and then we addressed the relevant diagnostic accuracy and the main concept of ROC analysis in diagnostic studies and it was followed with review of sample size calculation.

2. Examples

There are several examples of the use of ROC curve analysis in bioinformatics medicine. A large number of computer diagnostic systems have been developed to advise physician on patient

* Fax: +98 111 2229936.
E-mail address: drhajian@yahoo.com

http://dx.doi.org/10.1016/j.jbi.2014.02.013
1532-0464/

diagnosis and management. For example, recently a text classifier model for high quality article retrieval in internal medicine [9] and an automated text classifier to detect radiology report have been evaluated in ROC analysis [10]. In another study, several least square vector machines for prediction of preoperative malignancy of ovarian tumors have been developed and assessed [11]. In particular, a study was designed to evaluate a compute based algorithm for diagnosis of heart disease (HD) in order to provides useful information that can improve the cardiac diagnosis in a typical clinical setting [12]. In this study, 127 patients were entered in two cohorts of 60 and 67 subjects. The follow up information was available on 114 subjects with mean age of 60 years for their future cardiac problem for final diagnosis as gold standard. The heart disease program (HDP) algorithm was designed to assist physician in diagnosis of HD, in particular, condition leading to homodynamic dysfunction and heart failure. This diagnostic algorithm is based on casual probability in term of severity of necessary causes and the possible mechanism and the risk profile. The program uses the input data to specialize the patient profile including a set of prior probability of disease based on demographic characteristics and risk profile and to put a set of assumption and to compute the risk score. The authors wished to compare the diagnostic performance of physician alone and heard disease program (HDP) alone with combination of physician and HDP in prediction of cardiac problem. They used sensitivity and specificity and also ROC curve analysis but in their ROC analysis, comparison of different diagnostic tasks was done with descriptive method regardless of performing statistical test. However, it is most helpful to justify the required sample size. The question would be raised whether the achieved sample size has power to detect a desirable effect size of accuracy index and how to calculate the optimal sample size for their study. In addition, the power calculation would be helpful in interpretation of lack of difference between diagnostic tasks with achieved sample size. These questions are also relevant for any other ROC diagnostic studies in bioinformatics research.

3. Diagnostic accuracy and classification performance

In diagnostic studies of biomedical informatics which the test yields dichotomized outcome (positive or negative results), the accuracy is evaluated by sensitivity and specificity. These two measures determine the inherent ability of diagnostic test versus a dichotomized gold standard and they are not influenced by prior probability of disease (or prevalence) in population [2,4]. The gold standard may be another test without errors but a more expensive diagnostic method or invasive method. It can be the combination of tests that may be available in clinical follow up, surgical verification, autopsy, and biopsy or by panel of experts [5,6]. The sensitivity indicates the proportion of diseased subject with positive test result and specificity determines the proportion of nondiseased subject with negative test results. The sensitivity and specificity can be combined as one-dimensional index that is called likelihood ratio (LR). The positive LR is the ratio of probability of positive test in diseased to nondiseased and negative LR is the ratio of probability of negative test in diseased to nondiseased. In fact, the positive LR is the ratio of sensitivity to 1-specificity and the negative LR is the ratio of 1-sensitivity to specificity [13]. The higher value of positive LR corresponds with greater information of positive test result while the lower value of negative LR associates with more information of negative test results. In particular, the positive and negative LR is of greater interest in comparative studies of two diagnostic tests.

For a quantitative diagnostic test or the test results are recorded on ordinal scale, the sensitivity and specificity varies across the different thresholds and the sensitivity is inversely related with specificity [2,4,14]. Then, the plot of sensitivity versus 1-specificity is called receiver operating characteristic (ROC) curve and the area under the curve (AUC), as an effective measure of accuracy has been considered and it has a meaningful interpretations [15]. This curve plays a central role in evaluating diagnostic ability of tests to discriminate the true state of subjects and comparing two alternative diagnostic tasks when each task is performed on the same subject [5,14–16].

4. A review of sample size consideration

In evaluating the accuracy of diagnostic test in medicine, the sample size plays an important role either for estimation or testing of accuracy. A small sample size produces an imprecise estimate of accuracy with wide confidence interval [17] which is non-informative for decision makers in medical context. On the other hand, unduly large sample size is wastage of resources especially when the new diagnostic test is expensive [18]. Unfortunately, sample sizes calculations are rarely reported by clinical investigators for diagnostic studies [19,20] and few clinicians are aware of them. Researchers often decide about the sample size arbitrary either for their conveniences or from the previous literature. For example, among 40 (out of 1698 articles) published studies on non-screening diagnostic accuracy in five higher impact factors of ophthalmology journal in 2005, only one study (2.5%) reported a prior sample size calculation for a planned sensitivity and specificity of 80% and 95% confidence level [19]. Another report of eight journals published in 2002, 43 articles (out of 8999 articles) were non-screening on diagnostic accuracy and two of 43 studies (5%) reported a prior calculation of sample size but no study reported that the sample size had been calculated on the base of pre-planned subgroup analysis while twenty articles (47%) reported results for subgroup of patients [20].

In sample size calculation for estimating both sensitivity and specificity, Buderer [21] incorporated prevalence of disease in sample size formula for sensitivity/specificity and provided the table of sample size for sensitivity and specificity but only for precision of 10%. Malhotra and Indrayan [18] argued that the sample size without considering prevalence would be adequate for sensitivity or specificity alone but not for both while Obuchowski [22] addressed that this is because of unknown true disease status at time of sampling from target population. Charley et al. [23] have provided monogram for estimation of sensitivity and specificity with too many lines and curves make complexity in reading. Malhotra and Indrayan [18] presented a table of sample size calculation based on Borderer formula only for estimating sensitivity and specificity but not for testing. Simel et al. [24] deal with sample size based on desired likelihood ratios (LR) confidence interval but not calculate sample size for a wide range of marginal errors around LR. Obuchowski [22] also provided a review of sample size formula for a various diagnostic accuracy but did not provide practical tables for calculating sample sizes. Several other authors also developed methods for sample size calculation in diagnostic medicine [25–28] because of complexity of their methods for clinician and the lack of availability of software their methods were not used frequently in clinical practices. In this article, a review of the critical elements of sample size calculations for diagnostic accuracy were addressed conceptually and the formula was driven based on statistical principles with respect to study purpose (estimation or testing) and accuracy of interest (sensitivity/specificity, LR and AUC). We also calculated the required sample size for various situations and the calculated sample size were tabulated for practical convenience of clinician in diagnostic test evaluation.

5. Factors affecting on sample size for diagnostic accuracy

Based on statistical principle, as a general rule of sample size calculation for proportions, since sensitivity (or specificity) is a

proportion, it is intuitively appealing that the four essential elements are required for calculation of sample size in estimating sensitivity (or specificity): (1) a pre-determined value of sensitivity (or specificity) that is available from previous published studies or clinical judgment; this is because the standard error of sensitivity (or specificity) depends on its value; (2) the confidence level $(1-\alpha)$ for statistical judgment where α is the probability of type I error; (3) the precision of estimates of sensitivity (or specificity) i.e. the maximum difference between estimated sensitivity (or specificity) and the true value. Additionally, the prevalence of disease in population is needed to be ascertained and also to be taken into account in sample size calculation [21]. In effect, we are planning for the number of affected subjects and the number of unaffected subjects separately, so we need a total number of subjects that will make both these groups large enough. In practice usually it is the sensitivity, not the specificity that determines the total number of subjects to be used. When the true status or condition is known before undergoing subjects into new diagnostic test, no longer the prevalence is incorporated into sample size calculation for sensitivity/specificity [22]. For the purpose of testing, instead of third element, the difference of sensitivity (or specificity) under the null and alternative hypothesis is required (i.e. the maximum difference to be detected in statistical test with power of $1-\beta$ where β is the probability of type II error). Thus, the power of statistical test (the compliment of type II error) should be considered the prior sample size calculation. As a general rule, with higher precision (i.e. the lower marginal error: the half wide of confidence interval) in estimating accuracy and detecting a small difference of effect in testing of accuracy with higher power, a greater sample size is required.

6. Sample size for studies with binary test outcome

6.1. Sample size for adequate sensitivity/specificity

First, assume we wish to determine the number of cases to estimate sensitivity (Se) of new diagnostic test. Similarly, one may estimate specificity (Sp). Since sensitivity (or specificity) is a proportion, for estimation of sensitivity (or specificity) alone when the diseased status is known, the formula for sampler size with $(1-\alpha)$% confidence level and with maximum marginal error of estimate of d for constructing confidence interval of true value of sensitivity (or specificity) using normal approximation is driven as follows:

$$n=\frac{Z^2_{\frac{\alpha}{2}}\widehat{P}(1-\widehat{P})}{d^2} \tag{6.1}$$

where $\widehat{P}$ is pre-determined value of sensitivity (or specificity) that is ascertained by previous published data or clinician experience/judgment and for $\alpha=0.05$, $Z_{\frac{\alpha}{2}}$ is inserted by 1.96. This is an estimate of sample size for sensitivity or specificity alone when the true condition of disease status is known. Buderer [21] incorporated the prevalence of disease in formula for sample sizes calculation when the true disease status is not known at the time of sampling. This might occur in prospective study when consecutive subjects undergoing the test are used as samples [22]. In practice, the clinicians would like to estimate the number required both in sensitivity and specificity within a study population containing cases and controls. In this situation, to ensure the study sample which the test will be applied is a representative of study population, the proportion of cases and controls should be taken into account by the prevalence of the disease in population.

Lets n_{cases}, n_{total} (cases and control) and Prev denote the number of cases, the total sample sizes (cases and control) and the prevalence of disease respectively, then based on sensitivity, the overall sample size (both cases and controls) is

$$n_{\text{total}}=\frac{n_{\text{cases}}}{\text{Prev}} \tag{6.2}$$

In fact, if one divides the right hand side of Eq. (6.1) by the prevalence of disease in target population, it gives the total number of subjects (cases and controls) need for sensitivity. The required sample size of specificity is estimated by dividing the right hand side of Eq. (6.1) by (1-prevalence) that gives the total subject for specificity. Then, if one is interested for both sensitivity and specificity, the largest value of two calculated total sample sizes will be considered as total study samples.

If one knows the n_{total} and n_{cases}, one can simply calculate the number of control that is needed to estimate specificity of new diagnostic test as follows:

$$n_{\text{controls}}=n_{\text{total}}-n_{\text{cases}}=n_{\text{total}}(1-\text{Prev}) \tag{6.3}$$

Thus, based on specificity

$$n_{\text{total}}=\frac{n_{\text{controls}}}{1-\text{Prev}} \tag{6.4}$$

Simply one can drive the proportion of cases to controls as follows

$$\frac{n_{\text{cases}}}{n_{\text{controls}}}=\frac{\text{Prev}}{1-\text{Prev}} \tag{6.5}$$

Thus, the total sample sizes based on sensitivity and specificity respectively are

$$n_{\text{Se}}=\frac{Z^2_{\frac{\alpha}{2}}\widehat{\text{Se}}(1-\widehat{\text{Se}})}{d^2\times\text{Prev}} \tag{6.6}$$

$$n_{\text{Sp}}=\frac{Z^2_{\frac{\alpha}{2}}\widehat{\text{Sp}}(1-\widehat{\text{Sp}})}{d^2\times(1-\text{Prev})} \tag{6.7}$$

For $\alpha=0.05$, $Z_{\frac{\alpha}{2}}$ is inserted by 1.96; $\widehat{\text{Se}}, \widehat{\text{Sp}}$, and Prev are the pre-determined values of sensitivity, specificity and prevalence of disease respectively and d as the precision of estimate (i.e. the maximum marginal error) is pre-determined by clinical judgment of investigators.

For example, if the Se is primary interested in diagnostic screening purpose and lets the pre-determined values of Se and prevalence of disease as 80% and 10% respectively. In order the maximum marginal error of estimate does not exceed from 7% with 95% confidence level, the total required sample size can be driven by plugging the above values in Eq. (6.6) as follows:

$$n_{\text{Se}}=\frac{1.96^2\times 0.8\times 0.20}{0.07^2\times 0.10}=1254$$

In Section 8 (Tables 1 and 2), we calculated and tabulated the required total sample sizes with various values of Se, Sp, Prev and marginal errors.

Obviously, the formula based on sensitivity and specificity yield a similar sample size at prevalence of 0.5. With low prevalence, the required sample size based on sensitivity is much higher than that of specificity while the prevalence of diseased becomes more than 0.50 which is less intuitive appealing, the sample size based on sensitivity lower than that of specificity. In practice, clinicians may be guided first to calculate the number of cases based on sensitivity and then uses Eq. (6.5) to estimate the number of controls but our thought intuitively hints that first the maximum total number of subjects based on sensitivity and specificity should be taken into account and then the number of cases (or controls) be calculated based on Eq. (6.5).

Table 1
The required total samples sizes for estimating sensitivity with respect to marginal error, sensitivity and the prevalence of disease in target population.

Sensitivity	Marginal error	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.40	0.50	0.01	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
		n	*n*	*n*	*n*	*n*	*n*	*n*	*n*	*n*									
Prevalence																			
0.70	0.03	89,637	17,927	8964	5976	4482	3585	2988	2241	1793	905	944	996	1055	1120	1195	1281	1494	1793
0.75	0.03	80,033	16,007	8003	5336	4002	3201	2668	2001	1601	808	842	889	942	1000	1067	1143	1334	1601
0.80	0.03	68,295	13,659	6830	4553	3415	2732	2277	1707	1366	690	719	759	803	854	911	976	1138	1366
0.85	0.03	54,423	10,885	5442	3628	2721	2177	1814	1361	1088	550	573	605	640	680	726	777	907	1088
0.90	0.03	38,416	7683	3842	2561	1921	1537	1281	960	768	388	404	427	452	480	512	549	640	768
0.95	0.03	20,275	4055	2028	1352	1014	811	676	507	406	205	213	225	239	253	270	290	338	406
0.70	0.05	32,269	6454	3227	2151	1613	1291	1076	807	645	326	340	359	380	403	430	461	538	645
0.75	0.05	28,812	5762	2881	1921	1441	1152	960	720	576	291	303	320	339	360	384	412	480	576
0.80	0.05	24,586	4917	2459	1639	1229	983	820	615	492	248	259	273	289	307	328	351	410	492
0.85	0.05	19,592	3918	1959	1306	980	784	653	490	392	198	206	218	230	245	261	280	327	392
0.90	0.05	13,830	2766	1383	922	691	553	461	346	277	140	146	154	163	173	184	198	230	277
0.95	0.05	7299	1460	730	487	365	292	243	182	146	74	77	81	86	91	97	104	122	146
0.70	0.07	16,464	3293	1646	1098	823	659	549	412	329	166	173	183	194	206	220	235	274	329
0.75	0.07	14,700	2940	1470	980	735	588	490	368	294	148	155	163	173	184	196	210	245	294
0.80	0.07	12,544	2509	1254	836	627	502	418	314	251	127	132	139	148	157	167	179	209	251
0.85	0.07	9996	1999	1000	666	500	400	333	250	200	101	105	111	118	125	133	143	167	200
0.90	0.07	7056	1411	706	470	353	282	235	176	141	71	74	78	83	88	94	101	118	141
0.95	0.07	3724	745	372	248	186	149	124	93	74	38	39	41	44	47	50	53	62	74
0.70	0.10	8067	1613	807	538	403	323	269	202	161	81	85	90	95	101	108	115	134	161
0.75	0.10	7203	1441	720	480	360	288	240	180	144	73	76	80	85	90	96	103	120	144
0.80	0.10	6147	1229	615	410	307	246	205	154	123	62	65	68	72	77	82	88	102	123
0.85	0.10	4898	980	490	327	245	196	163	122	98	49	52	54	58	61	65	70	82	98
0.90	0.10	3457	691	346	230	173	138	115	86	69	35	36	38	41	43	46	49	58	69
0.95	0.10	1825	365	182	122	91	73	61	46	36	18	19	20	21	23	24	26	30	36

Table 2
The required total sample sizes for estimating specificity with respect to marginal error, specificity and the prevalence of disease in target population.

Specificity	Marginal error	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.40	0.5
		n	*n*	*n*	*n*	*n*	*n*	*n*	*n*	
Prevalence										
0.70	0.03	905	944	996	1055	1120	1195	1281	1494	1793
0.75	0.03	808	842	889	942	1000	1067	1143	1334	1601
0.80	0.03	690	719	759	803	854	911	976	1138	1366
0.85	0.03	550	573	605	640	680	726	777	907	1088
0.90	0.03	388	404	427	452	480	512	549	640	768
0.95	0.03	205	213	225	239	253	270	290	338	406
0.70	0.05	326	340	359	380	403	430	461	538	645
0.75	0.05	291	303	320	339	360	384	412	480	576
0.80	0.05	248	259	273	289	307	328	351	410	492
0.85	0.05	198	206	218	230	245	261	280	327	392
0.90	0.05	140	146	154	163	173	184	198	230	277
0.95	0.05	74	77	81	86	91	97	104	122	146
0.70	0.07	166	173	183	194	206	220	235	274	329
0.75	0.07	148	155	163	173	184	196	210	245	294
0.80	0.07	127	132	139	148	157	167	179	209	251
0.85	0.07	101	105	111	118	125	133	143	167	200
0.90	0.07	71	74	78	83	88	94	101	118	141
0.95	0.07	38	39	41	44	47	50	53	62	74
0.70	0.10	81	85	90	95	101	108	115	134	161
0.75	0.10	73	76	80	85	90	96	103	120	144
0.80	0.10	62	65	68	72	77	82	88	102	123
0.85	0.10	49	52	54	58	61	65	70	82	98
0.90	0.10	35	36	38	41	43	46	49	58	69
0.95	0.10	18	19	20	21	23	24	26	30	36

Alternatively, Li and Fine developed sample size for sensitivity and specificity in prospective studies when disease status may not be known at the time of enrolment since the gold standard applied at pre-determined time after initial screening. They developed a formal method of sample size calculation based on unconditional power property of statistical test [26].

6.2. Sample size for testing sensitivity (or specificity) of single diagnostic test

Suppose P_0 denote the pre-determined value of sensitivity or specificity of new diagnostic test. In comparing the test's accuracy to fixed value of P_0, the null and alternative hypothesis is

$$H_0: \text{Se} = P_0 \text{ versus } H_1: \text{Se} \neq P_0 \ (\text{or Se} = P_1)$$

where P_1 is the value of sensitivity (or specificity) under alternative hypothesis. A general sample size formula for comparison of proportion with fixed value can be applied for evaluation of single diagnostic test. With $(1-\alpha)\%$ confidence level and $(1-\beta)\%$ power for detection an effect of $P_1 - P_0$ using normal approximation as a general rule, Z-score under the null and alternative hypothesis can defined and thus the required sample size for cases is driven as follows:

$$n = \frac{\left[Z_{\frac{\alpha}{2}}\sqrt{P_0(1-P_0)} + Z_{\beta}\sqrt{P_1(1-P_1)}\right]^2}{(P_1 - P_0)^2} \tag{6.8}$$

where $Z_{\frac{\alpha}{2}}$ and Z_β denote the upper $\frac{\alpha}{2}$ and β percentiles of standard normal distribution and α, β are the probability of type I and type II errors respectively. For $\alpha = 0.05$ and $\beta = 0.20$, they are inserted by $Z_{\frac{\alpha}{2}} = 1.96$ and $Z_\beta = 0.84$ respectively. In this paper, we used consistently two side tests instead of one side test in our sample size calculation; for one side test Z_α and Z_β should be used.

For example, an investigator compares H_0: Se = 0.70 versus H_1: Se ≠ 0.70. The sample size one would need to have 95% confidence and 80% power to detect a difference of 10% from presumption value of Se = 70%, can be calculated by plugging in the above information in Eq. (6.8) as follows:

$$n = \frac{(1.96 \times \sqrt{0.70 \times 0.30} + 0.84 \times \sqrt{0.80 \times 0.20})^2}{(0.10)^2} = 153$$

In Eq. (6.8), one may argue that why the prevalence was not appeared in the formula. As we already mentioned since the true status of disease is know in comparative situations. Therefore, the prevalence is not relevant for sample size calculation in this condition.

6.3. Sample size for comparing the sensitivity (or specificity) of two diagnostic tests

A general formula of sample size calculation for comparing two independent proportions can be used to estimate sample size for studies comparing sensitivity and/or specificity of two tests of unpaired design. In comparing the diagnostic accuracy of two alternative tasks for two independent samples, suppose P_1 and P_2 denote the expected proportion (Se or Sp) of two alternative diagnostic tests respectively. For testing hypothesis: H_0: $P_1 = P_2$ versus H_1: $P_1 \neq P_2$, the required sample size with equal size based on normal approximation of binomial data with $1-\alpha$ confidence level and $1-\beta$ power is

$$n = \frac{\left[Z_{\frac{\alpha}{2}}\sqrt{2 \times \overline{P}(1-\overline{P})} + Z_\beta \sqrt{P_1(1-P_1) + P_2(1-P_2)}\right]^2}{(P_1 - P_2)^2} \tag{6.9}$$

where $\overline{P}$ the average of P_1 and P_2 and Z_α, Z_β is are the standard normal Z values corresponding to α and β (the probability of type I and type II errors respectively).

Suppose, one wishes to compare the Se of two alternative diagnostic tasks H_0: $P_1 = P_2$ versus H_1: $P_1 \neq P_2$. The sample size would one need to have 95% confidence and 80% power to detect a difference of 10% from a Se of 70% (i.e. $P_1 = 0.70$, $P_2 = 80\%$ and $\overline{P} = 0.75$) can be calculated by inserting this information in Eq. (6.9) as follows:

$$n = \frac{(1.96 \times \sqrt{2 \times 0.75 \times 0.25} + 0.84 \times \sqrt{0.70 \times 0.30 + 0.80 \times 0.20})^2}{(0.10)^2}$$

$$= 293$$

Epi info software can be used to perform these calculations and to estimate the required sample size for proportion. Also the approach can be extended for paired designs when multiple tests are performed on the same subjects, then the proportion (Se or Sp) should be considered as dependent. Beam [29] presented formulae for calculation of sample size for paired designs; he has also written a program in FORTRAN for calculation of sample sizes. To avoid the increased complexity of this text for clinician, we referred the interesting readers to previously published papers [29–32].

6.4. Sample size for likelihood ratio estimation

As we described when test yields positive or negative results, sensitivity and specificity are the two inherent indexes of accuracy. One may estimate sample size based on one of these two indexes regarding preference of researcher either uses sensitivity or specificity as primary of interest but LR that combines the sensitivity and specificity of test as uni-dimensional index, is a greater of interest. A test with higher LR^+ has a greater value of rule in the disease while a test with lower value of LR^- has a higher value of rule out disease. These two indexes are particularly interesting in comparative studies of two or multiple tests. The test with greater value of LR^+ and lower values of LR^- has more diagnostic abilities in the classification of true status of diseased and nondiseased. Thus, positive LR and negative LR play an important rule for clinical decision and they can be used in estimating sample size in diagnostic test. Simel et al. [24] proposed the confidence interval of positive LR (or negative LR) to be used for sample size estimation. For example, a clinical investigator wishes to calculate sample size where the LR^+ is greater from a pre-determined value of LR with $(1-\alpha)$% confidence interval (i.e. a pre-determined value lies within the confidence bound with $(1-\alpha)$% confidence level.

Suppose $\widehat{P}_1$ and $\widehat{P}_2$ denote the sensitivity and 1-specificity of a test respectively and n_1 and n_2 denote the sample size for diseased and nondiseased. The ratio estimator of $LR^+ = \frac{\widehat{P}_1}{\widehat{P}_2}$ is skewed and the logarithm transformation can be used to convert its distribution to normal approximately. Thus, $\log \frac{\widehat{P}_1}{\widehat{P}_2}$ can be assumed that asymptotically normally distributed with standard error of $\sqrt{\frac{1-\widehat{P}_1}{n_1\widehat{P}_1} + \frac{1-\widehat{P}_2}{n_2\widehat{P}_2}}$ and therefore $(1-\alpha)$% confidence interval for $\log(LR^+)$ is as follows:

$$\log(LR^+) = \log\frac{\widehat{P}_1}{\widehat{P}_2} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{1-\widehat{P}_1}{n_1\widehat{P}_1} + \frac{1-\widehat{P}_2}{n_2\widehat{P}_2}} \tag{6.10}$$

With the presumption of equal sample size for diseased and nondiseased (i.e. $n_1 = n_2 = n$), then the required sample size for each group of cases and controls can be calculated by solving the Eq. (6.10) as follows:

$$n = \frac{\left(Z_{\frac{\alpha}{2}}\sqrt{\frac{1-\widehat{P}_1}{\widehat{P}_1} + \frac{1-\widehat{P}_2}{\widehat{P}_2}}\right)^2}{\left(\log(LR+) - \log\frac{\widehat{P}_1}{\widehat{P}_2}\right)^2} \tag{6.11}$$

For example, an investigator wishes to estimate the sample size of a study where LR^+ has more valuable when $LR^+ \geqslant 2$. Given the result of pilot study, sensitivity = 0.8 and specificity = 0.7 and thus $LR^+ = \frac{Se}{1-Sp} = 2.96$. For sample size calculation, we substituted LR = 2 for lower bond of confidence interval. With the presumption of equal sample size for diseased and nondiseased (i.e. $n_1 = n_2 = n$), then the required sample size can be calculated by solving the following equation:

$$2 = \exp\log\frac{0.8}{0.7} - 1.96\sqrt{\left(\frac{1}{n}\right)\left[\left(\frac{0.2}{0.8} + \frac{0.7}{0.3}\right)\right]} \tag{6.12}$$

$n = 74$ for each group and thus, the total sample size would be 148. With these sample sizes the investigator 95% of time to be confident that positive LR is greater than 2 (i.e. the LR of 2 lies below the lower bound of 95% confidence interval). One also can assume the ratio of r for controls (n_1) to cases (n_2), then $n_2 = r \times n_1$. By replacing $r \times n_1$ instead of n_2, the solution of Eq. (6.10) yields sample size for cases (n_1).

In another condition, a diagnostic test may be useful, if negative LR is lower than a pre-determined value. Then, the sample size

would be calculated based on the confidence bond of negative LR. Similarly, one could drive a confidence interval for LR^- as follows:

$$\log(LR^-) = \left(\log\frac{\widehat{P}_1}{\widehat{P}_2} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{1-\widehat{P}_1}{n_1\widehat{P}_1}+\frac{1-\widehat{P}_2}{n_2\widehat{P}_2}}\right) \quad (6.13)$$

where $\widehat{P}_1 = Sp$ and $\widehat{P}_2 = 1 - Se$.

For example, from a clinical point of view, a test is useful when the maximum value of negative LR about 0.4 and from literature review Se = 0.9 and Sp = 0.5. With presumption of $n_1 = n_2 = n$, the required sample size is calculated by solving the following equation with respect to n,

$$\log 0.4 = \log\frac{0.9}{0.5} + 1.96\sqrt{\left(\frac{1}{n}\right)\left(\frac{0.9}{0.1}+\frac{05}{0.5}\right)} \quad (6.14)$$

Thus, $n = 80$ for each group and total sample size is 160.

7. Sample size for studies of ROC index of accuracy

7.1. Sample size for estimating accuracy index

So far, we discussed sample size calculations when the test yields a binary outcome. As we addressed already, for a quantitative test or the test results are recorded on ordinal scale, ROC curves show the trade off between sensitivity and specificity and the area under the curve (AUC) is considered as an index of accuracy. The AUC can be estimated parametric (binormal model) and nonparametric (Wilcoxon statistic) approaches. Both approaches allow estimating the sampling variability of AUC. In diagnostic studies, involving ROC analysis, for the purpose of estimating or testing AUC, the clinicians should decide the number of patients and controls needed in study protocol. Suppose a clinical researcher wishes to estimate the diagnostic accuracy as defined by AUC in which the marginal error of estimate (i.e. the difference between true AUC and its estimate) does not exceed from a pre-determined value of d with $(1-\alpha)$% confidence level (e.g. 95%). Using normal approximation in constructing confidence interval for AUC, we have

$$Z_{\frac{\alpha}{2}}SE(\widehat{AUC}) \leqslant d \quad (7.1)$$

By squaring two sides of equation, then

$$Z^2_{\frac{\alpha}{2}}Var(\widehat{AUC}) = d^2 \quad (7.2)$$

Lets $V(\widehat{AUC}) = nVar((\widehat{AUC})$. Thus, the required sample size for each group of nondiseased and diseased is

$$n = \frac{Z^2_{\frac{\alpha}{2}}V(\widehat{AUC})}{d^2} \quad (7.3)$$

The variance of $\widehat{AUC}$ denoted as $Var(\widehat{AUC})$ can be estimated can be estimated parametrically based on binormal assumption [22,33,34] (see Appendix B) or exponential approximation using Hanley and McNeil formula [15] (see Appendix A).

In numerical results that we produced the required sample size in Section 9, we used binormal based variance of AUC as described in Appendix B and we assumed $n_1 = n_2 = n$. Thus, the $V(AUC)$ can easily be driven from Eq. (A3) (see Appendix B) as follows:

$$nVar(\widehat{AUC}) = V(AUC) = (0.0099 \times e^{-a^2/2}) \times (6a^2 + 16) \quad (7.4)$$

where $a = \varphi^{-1}(AUC) \times 1.414$ and φ^{-1} is the inverse of standard cumulative normal distribution.

For example, Suppose, the pre-determined value of AUC = 0.70, in order to estimate AUC with 95% confidence the degree of precision of estimate about 0.07, the required sample size can be calculated as follows:

For estimating $V(AUC)$, first one should calculate $a = \varphi^{-1}(0.70) \times 1.414 = 0.741502$ and then, $V(AUC)$ can be calculated using Eq. (7.4) as:

$$V(AUC) = (0.0099 \times e^{-0.741502^2/2}) \times (6 \times 0.741502^2 + 16) = 0.145136$$

Therefore, the required sample size is obtained by inserting the $V(AUC)$ and $d = 0.07$ in Eq. (7.3) as follows:

$$n = \frac{1.96^2 \times 0.145136}{0.07^2} = 114$$

Alternatively, the pure nonparametric method was proposed by Delong et al. [35–37] (see Appendix C). This method is based on structure components of individual based data which we called as pseudo accuracies for both diseased and nondiseased subjects. Hanley and McNeil formula [15] is a convenient method for estimation of SE of AUC that only depends on estimate of accuracy index (AUC) but it does not consider the ratio of standard deviation of nondiseased to diseased populations. While the normal based standard error is more flexible to consider the ratio of two standard deviations [33]. Both formulae take into account the ratio of controls to cases. Obuchowski [33] reported that Hanley and McNeil method underestimate the SE of AUC for rating data but not for continuously distributed data while the normal based standard error is more flexible to consider the ratio of two standard errors produces a conservative estimate of SE In a Monte Carlo simulation studies with continuously distributed data, Hajian-Tilaki and Hanley [38] showed that overall the three methods of SE of AUC (binormal model, exponential approximation and Delong's method) worked well in reflecting actual variation for various configurations of nonbinormal data while for bimodal data, the binormal estimator of SE produces as more conservative estimate of SE than others.

7.2. Sample size for testing accuracy of quantitative diagnostic test of single modality

Assume a clinical researcher may wish to test the accuracy (AUC as unknown parameter of interest) of a new diagnostic method with a pre-specified value of AUC_0. The null and alternative hypothesis is

$H_0: AUC = AUC_0$ versus $H_1: AUC \neq AUC_0$ (i.e. $AUC = AUC_1$)

Using the normal approximation, the required sample size for each group of cases and controls (assuming the ratio of cases to controls is one) in detecting an effect of $\delta = AUC_1 - AUC_0$ with $(1-\alpha)$% confidence level and $(1-\beta)$% power is as follows

$$n = \frac{\left[Z_{\frac{\alpha}{2}}\sqrt{V_{H0}(\widehat{AUC})} + Z_\beta\sqrt{V_{H1}(\widehat{AUC})}\right]^2}{[AUC_1 - AUC_0]^2} \quad (7.5)$$

where $V(\widehat{AUC} = nVar(\widehat{AUC})$. $Var_{H0}(\widehat{AUC})$ and $Var_{H1}((\widehat{AUC})$ denote the variance of $\widehat{AUC}$ under the null and alternative hypothesis respectively.

7.3. Sample size for comparing accuracy of two diagnostic tests

In comparative study of two diagnostic tasks in the context of ROC analysis, for example, a clinical investigator has a plan to compare the accuracy of MRI and CT in detecting abnormal condition. The accuracy of these two diagnostic tasks is determined by AUC and the same subjects are undergoing two alternative tasks for

the purpose of efficiency of design for the assessment of conditions. Let AUC_1 and AUC_2 are the true two diagnostic accuracies for the two diagnostic modalities respectively. The null and alternative hypothesis are

$H_0: AUC_1 = AUC_2$ versus $H_1: AUC_1 \neq AUC_2$

The investigator wants to decide how many cases and controls are needed to detect an effect between two diagnostic tasks as defined by $\delta = AUC_1 - AUC_2$ under alternative hypothesis with $(1-\alpha)\%$ confidence level and $(1-\beta)\%$ power. By constructing confidence interval for parameter of interest $AUC_1 - AUC_2$ using normal approximation under the null and alternative hypothesis, then the required sample sizes for each group are driven as

$$n = \frac{\left[Z_{\frac{\alpha}{2}}\sqrt{V_{H0}(\widehat{AUC}_1 - \widehat{AUC}_2)} + Z_{\beta}\sqrt{V_{H1}(\widehat{AUC}_1 - \widehat{AUC}_2)}\right]^2}{[AUC_1 - AUC_2]^2} \quad (7.6)$$

where

$$V(\widehat{AUC}_1 - \widehat{AUC}_2) = n\mathrm{Var}(\widehat{AUC}_1) + n\mathrm{Var}(\widehat{AUC}_2) - 2n\mathrm{Cov}(\widehat{AUC}_1, \widehat{AUC}_2) \quad (7.7)$$

In Eq. (7.6), the $\mathrm{Var}(\widehat{AUC}_1 - \widehat{AUC}_2)$ needs to be estimated under the null (H_0) and alternative hypothesis (H_1). In a case, two diagnostic tasks do not apply on the same subjects, then the two AUC's are independent and the Eq. (7.6) can be written as:

$$n = \frac{\left[Z_{\frac{\alpha}{2}}\sqrt{2V_{H0}(\widehat{AUC})} + Z_{\beta}\sqrt{V(\widehat{AUC}_1) + V(\widehat{AUC}_2)}\right]^2}{[AUC_1 - AUC_2]^2} \quad (7.8)$$

where the AUC under the H_0 is the average of AUC_1 and AUC_2.

For example for detection of 10% difference in estimating two independent diagnostic systems with 95% confidence and 80% power, assuming AUC1 = 0.70 and AUC2-AUC1 = 0.10 using binormal based variance of AUC as described in Section 7.1, on could easily calculate $V(AUC)_{H0} = 0.13480$; $V(AUC_1) = 0.14513$; $V(AUC_2)$ - = 0.11946. Then the required sample size for each task with 80% power and 95% confidence for detection 10% difference in accuracy index (AUC) using Eq. (7.8) as

$$n = \frac{[1.96 \times \sqrt{2 \times 0.1348} + 0.84 \times \sqrt{0.14513 + 0.11946}]^2}{(0.10)^2} = 211$$

In the following Section 9, we also calculated the required sample sizes in different situation of AUC and effect size and it is tabulated in Table 7 for two independent AUC.

The $\mathrm{Var}(\widehat{AUC}_1)$ and $\mathrm{Var}(\widehat{AUC}_2)$ can also be estimated using Hanley and McNeil formula [15] but it does not gives the covariance between two correlated AUC's. While the advantage of Delong's method [35,36] is that the covariance between two correlated AUC's can be estimated from its components of variance and covariance matrix as well. In addition, CORROC software (Metz Software) also estimates the covariance between the two correlated AUC's in parametric approach of comparative study of two diagnostic tasks [39].

$$\mathrm{Cov}(\widehat{AUC}_1, \widehat{AUC}_2) = r\mathrm{SE}(\widehat{AUC}_1) \cdot \mathrm{SE}(\widehat{AUC}_2) \quad (7.9)$$

where and SE denote the correlation between the two estimated AUC's and the standard error (i.e. the square root of variance) of estimate of AUC's respectively. If the two diagnostic tasks are not examined on the same subjects, the two estimated AUC to be independent and thus the covariance term will be zero. The investigator may assume the ratio of sample size for the controls to the cases to be as *R*. Then, the sample size can be estimated for each group with this ratio.

Although AUC as an overall accuracy index is admired as robust index with meaningful interpretation and primarily interesting in diagnostic accuracy, in comparing of AUC from two diagnostic tasks when two ROC curves crossing each other, this overall index may be not useful. The partial area at clinical relevant of false positive and true positive fraction at specific false positive rate (TPF_{FPF}) are the two other indexes of accuracy that had been considered in ROC analysis. The methods have been developed for sample size calculations of partial area and TPF_{FPF} [22]. In addition, ROC studies in particular for rating data may be involved with multiple readers. Obuchowski provided tables of calculated sample size for multiple reader studies. The interesting readers are referred to some published articles in this area [22,40].

8. Sample size calculation

We calculated sample size using Excel software (Windows office 2007) for purpose of estimating of different diagnostic accuracies (sensitivity, specificity, LR and AUC) and of testing as well, assuming the ratio of sample sizes for cases and controls to be one. We provided tables of required samples for combinations of various conditions. In particular, for estimating and testing of AUC in single modality and comparative study of two correlated and independent accuracy index (AUC), we used binomial based variance with standard ratio of one. For the two correlated AUC when the two alternative tasks are applied on the same subject, we assumed the correlation of 0.5. A correlation between two accuracies close to this value was reported by Rockette et al. [41]. Obviously, if the different samples of patients underwent the two different diagnostic tasks, then the correlation would be zero. We examined the required sample sizes with a wide range of accuracy indexes from 0.60 to 0.98 and a rational range of marginal errors and different effects (varied from 0.03 to 0.15) with 95% confidence level and 80% power.

9. Results of sample size calculation

Table 1 shows that the required total sample size for sensitivity is substantially varied in relation to marginal errors (i.e. the difference between the estimates and true value that is expected to be detected or one half the desired width of the confidence interval), the degree of sensitivity and the prevalence of disease in population. The greatest sample size of 89,637 was calculated for low marginal errors (0.03) and low sensitivity (0.70) and low prevalence (0.01). For a given sensitivity and a marginal error, the sample size substantially decreased as prevalence reached to 0.50. The smallest sample size for sensitivity was calculated when the prevalence was about 0.50. For example, for the sensitivity of 0.70 and the marginal error of 0.03, the required sample size basically varied from 89,637 to 179 as prevalence elevated from 0.01 to 0.50. In addition, for a given prevalence and marginal error, sample size decreased as sensitivity increased. Table 2 shows that the required total sample size for specificity increases by higher prevalence of disease. The largest sample size was calculated for low specificity and low marginal errors and high prevalence (the first row of Table 2). The smallest one was computed for high accuracy, high marginal error and low prevalence. In Table 3, the third column shows the point estimate of LR^+, as confidence bound of LR^+ becomes wider which included with higher marginal error, the sample size decreased since LR as ratio estimator is more sensitive with respect to variation of sensitivity and specificity. A wider confidence bound of LR^+ (the difference between the third column and the boundary value) (i.e. lower precision) which is unacceptable statistically and clinically as well, produces a small sample size. Similarly, Table 4 shows the narrower of confidence bound for LR^- (i.e. the lower

Table 3
Calculation of sample sizes for estimation of LR^+ within a 95% confidence level.

Sensitivity	Specificity	Estimated LR^+	$LR^+ = 2$	$LR^+ = 2.5$	$LR^+ = 3$	$LR^+ = 3.5$	$LR^+ = 4$	$LR^+ = 4.5$	$LR^+ = 5$	$LR^+ = 6$				
0.70	0.60	1.75	416	58	26	15	11	8	7	5	2	3	2	2
0.75	0.60	1.88	1691	85	32	18	12	9	7	5	2	3	2	3
0.80	0.60	2.00	–	135	41	21	14	10	8	6	2	3	2	3
0.85	0.60	2.13	1752	244	54	26	16	11	9	6	2	4	2	3
0.90	0.60	2.25	446	558	75	32	19	13	10	6	2	4	2	3
0.95	0.60	2.38	202	2267	109	40	22	15	11	7	2	4	2	3
0.70	0.70	2.33	447	2229	168	65	37	25	18	12	3	7	2	5
0.75	0.70	2.50	206	–	308	90	46	30	21	13	3	8	2	5
0.80	0.70	2.67	120	2383	715	134	60	36	25	15	3	8	2	6
0.85	0.70	2.83	79	615	2951	216	81	45	30	17	3	9	2	6
0.90	0.70	3.00	57	282	–	395	113	57	36	20	3	10	3	6
0.95	0.70	3.17	43	164	3136	915	168	74	44	22	3	11	3	7
0.70	0.80	3.50	54	150	716	–	954	269	134	59	5	25	4	15
0.75	0.80	3.75	42	101	334	3497	3997	501	201	75	5	29	4	17
0.80	0.80	4.00	34	74	197	916	–	1177	328	99	6	34	4	19
0.85	0.80	4.25	28	57	132	426	4365	4911	607	135	6	40	5	22
0.90	0.80	4.50	24	46	96	250	1138	–	1423	191	7	48	5	25
0.95	0.80	4.75	21	38	74	167	527	5326	5917	285	8	57	5	28
0.70	0.90	7.00	23	34	50	75	116	186	320	1524	45	2031	17	285
0.75	0.90	7.50	21	30	43	62	91	137	218	720	93	8608	21	433
0.80	0.90	8.00	18	26	37	52	74	107	161	429			27	714
0.85	0.90	8.50	17	24	33	45	62	87	125	291	−98	9592	37	1335
0.90	0.90	9.00	15	21	29	39	53	73	101	213	−50	2523	56	3153
0.95	0.90	9.50	14	20	26	35	46	62	84	165	−34	1178	115	13,218
0.70	0.95	14.00	20	25	31	39	48	58	70	104	−15	238	−26	659
0.75	0.95	15.00	18	23	29	35	43	51	62	88	−14	188	−21	452
0.80	0.95	16.00	17	21	26	32	38	46	55	77	−12	154	−18	335
0.85	0.95	17.00	16	20	24	29	35	42	49	68	−11	130	−16	262
0.90	0.95	18.00	15	19	23	27	32	38	45	61	−11	112	−15	212
0.95	0.95	19.00	14	18	21	26	30	35	41	55	−10	98	−13	178

Table 4
Calculation of sample sizes for estimation of LR^- within 95% confidence level for pre-determined value of LR^-.

Sensitivity	Specificity	Estimated LR^-	$LR^- = 0.05$	$LR^- = 0.10$	$LR^- = 0.15$	$LR^- = 0.20$	$LR^- = 0.30$	$LR^- = 0.40$	$LR^- = 0.50$	$LR^- = 0.60$
0.70	0.60	0.50	2	4	8	14	44	231	–	347
0.75	0.60	0.42	3	7	13	26	131	8453	424	106
0.80	0.60	0.33	5	12	28	69	1615	539	109	52
0.85	0.60	0.25	9	29	93	489	732	110	51	32
0.90	0.60	0.17	26	142	3345	1117	107	48	31	23
0.95	0.60	0.08	290	2273	219	99	46	31	24	19
0.70	0.70	0.43	2	5	10	18	83	2229	447	94
0.75	0.70	0.36	3	8	18	39	433	1026	116	49
0.80	0.70	0.29	6	15	41	134	7147	150	54	31
0.85	0.70	0.21	11	40	184	4919	207	60	33	22
0.90	0.70	0.14	33	285	15,216	320	66	34	23	18
0.95	0.70	0.07	587	659	136	70	36	25	20	16
0.70	0.80	0.38	2	6	12	25	199	2383	120	45
0.75	0.80	0.31	4	10	23	63	7492	205	57	29
0.80	0.80	0.25	6	19	63	328	491	74	34	21
0.85	0.80	0.19	13	58	456	5457	103	40	24	17
0.90	0.80	0.13	42	714	1069	161	46	26	18	14
0.95	0.80	0.06	1485	335	96	55	30	21	17	14
0.70	0.90	0.33	3	6	15	36	846	282	57	27
0.75	0.90	0.28	4	11	31	111	2018	90	35	20
0.80	0.90	0.22	7	25	102	1423	175	46	24	16
0.85	0.90	0.17	15	85	1999	668	64	29	18	14
0.90	0.90	0.11	55	3153	389	101	35	21	15	12
0.95	0.90	0.06	6614	212	74	45	26	19	15	13
0.70	0.95	0.32	3	7	17	44	3484	164	43	22
0.75	0.95	0.26	4	13	37	156	683	67	28	17
0.80	0.95	0.21	8	28	135	5917	124	38	21	14
0.85	0.95	0.16	17	105	8351	393	53	25	17	12
0.90	0.95	0.11	63	13,218	277	84	32	20	14	11
0.95	0.95	0.05	27,819	178	67	41	24	18	14	12

difference between the third column and above) is corresponded with higher sample sizes.

Table 5 shows for estimation of AUC with a given marginal error of 0.03 and 95% confidence level, sample sizes varied from 665 to 42 for low (AUC = 0.60) to high (AUC = 0.98) accuracy index. For moderate marginal error (0.05), the required sample size was 240 and 50 from low to high accuracy. Table 6 shows for testing AUC with pre-determined fixed value and for detecting an effect of 0.03 with 95% confidence level and 80% power, the sample sizes varied from 1317 to 716 for low to high accuracy while for

Table 5
The calculated sample sizes for each group of diseased and nondiseased in estimating diagnostic accuracy (AUC) in various configurations of AUC and marginal errors with 95% confidence level.

AUC	$d = 0.03$ n	Marginal $d = 0.05$ n	Errors $d = 0.07$ n	$d = 0.10$ n
0.60	665	240	122	60
0.65	648	234	119	59
0.70	620	232	114	58
0.73	596	215	110	54
0.75	576	208	106	52
0.78	540	194	99	49
0.80	510	184	94	46
0.83	458	165	85	42
0.85	418	151	77	38
0.88	347	125	64	32
0.90	293	106	57	27
0.93	202	73	38	–
0.95	137	50	–	–
0.98	42	–	–	–

Table 6
The required sample sizes for each group of diseased and nondiseased for testing the accuracy of single diagnostic method (H_0: AUC = AUC_0 versus H_1: AUC = $AUC_0 + \delta$) with different values of AUC and effect (δ) with 95% confidence level and 80% power.

AUC	n $\delta = 0.03$	n $\delta = 0.05$	n $\delta = 0.07$	n $\delta = 0.10$	n $\delta = 0.12$	n $\delta = 0.15$
0.60	1317	469	240	117	80	51
0.65	1293	464	234	113	78	49
0.70	1259	458	226	108	74	46
0.73	1233	455	220	105	71	43
0.75	1212	452	215	102	69	41
0.78	1175	448	206	96	64	38
0.80	1145	445	199	92	61	35
0.83	983	441	187	85	55	30
0.85	1053	438	177	78	49	24
0.88	979	433	158	66		–
0.90	920	429	143	54	–	–
0.93	811	425	–	–	–	–
0.95	716	–	–	–	–	–

Table 7
The required sample sizes for each group of diseased and nondiseased for comparison of two diagnostic methods for two independent samples for detection an effect of $\delta = AUC_1 - AUC_2$ and for different AUCs and effects (δ) with 95% confidence level and 80% power.

AUC_1	N $\delta = 0.03$	n $\delta = 0.05$	N $\delta = 0.07$	n $\delta = 0.10$	n $\delta = 0.12$	n $\delta = 0.15$
0.60	2696	966	490	238	164	103
0.65	2615	933	472	227	156	97
0.70	2482	881	442	211	143	88
0.73	2369	837	418	197	133	81
0.75	2278	801	398	186	125	80
0.78	2111	736	362	167	110	65
0.80	1979	685	335	152	100	57
0.83	1744	595	286	126	80	44
0.85	1562	526	248	106	66	36
0.88	1248	407	185	74	44	–
0.90	1012	319	139	52	–	–
0.93	626	202	70	–	–	–
0.95	361	90	–	–	–	–

Table 8
The required sample sizes for each group of diseased and nondiseased for comparison of two diagnostic methods on the same subjects[a] for detection an effect of $\delta = AUC_1 - AUC_2$ and for different AUCs and effects (δ) with 95% confidence level and 80% power.

AUC_1	n $\delta = 0.03$	n $\delta = 0.05$	n $\delta = 0.07$	n $\delta = 0.10$	n $\delta = 0.12$	n $\delta = 0.15$
0.60	2243	804	408	198	136	86
0.65	2176	777	393	189	130	81
0.70	2065	733	369	176	120	74
0.73	1972	697	348	165	111	68
0.75	1896	667	332	156	105	67
0.78	1758	614	303	140	93	56
0.80	1648	571	280	128	84	50
0.83	1453	497	240	107	69	41
0.85	1301	439	209	91	59	35
0.88	1041	342	158	66	44	–
0.90	846	270	121	52	–	–
0.93	527	172	69	–	–	–
0.95	310	–	–	–	–	–

[a] The correlation between two AUCs was assumed as 0.50.

detecting a moderate effect (0.07), the required sample size was 240 for low and 143 for high AUC.

Tables 7 and 8 show that for the comparison of two independent diagnostic tasks, as one expected the required sample size was greater than that of the two correlated indexes in similar conditions. For example the required sample size for each group for detecting an effect of 0.07 with 95% confidence and 80% power in comparison of two independent AUC is equal to 490 for low accuracy and 70 for high accuracy while for two correlated AUC for detecting of the same effect and the same confidence and power, the required sample size decreased to 408 for low and to 69 for high accuracy (Table 8).

10. Illustration of analysis and sample size with an example

As an example of biomedical informatics study illustrated in Section 2, the performance of three diagnostic tasks (physician alone, HDP alone and their combination) was assessed for prediction of heart failure [12]. First the author used sensitivity and specificity in their analysis. Then, they acknowledged the limitation of sensitivity and specificity that depending on threshold used. Thus, they performed ROC curve analysis. The ROC curves were depicted for physician alone and HDP alone and also some points for their combination. They only tested for comparison of sensitivity and specificity but not for AUC. They just descriptively reported AUC were 68.8%, 70% for physician alone and HDP alone respectively. They descriptively presented and concluded that the combination of HDP and physician gave better discrimination than HDP program and physician alone without reporting the p-values. It is more helpful to report the AUC for the combination of two tasks and the SE of each task and their confidence interval of difference of accuracy of two alternative tasks. Since this study used a matched paired design, we recommend the SE of difference of AUCs to be calculated using Delong's method. However, it is not clear with the small difference observed in ROC curves this difference would be significant with the achieved sample size. In addition, they also reported the sensitivity of HDP alone was significantly higher than physician alone (0.53% versus 34.8%, $p < 0.001$) while physician alone had higher specificity than HDP alone (93.9% versus 75.6%) but their depicted ROC curves did not shows such a differences in ROC space. Although, they reported p-values for comparison of sensitivity and specificity of different tasks, reporting the CI would have been more informative. Also reporting SE of AUC or its CI and the p-value also would be informative instead of a visualized comparison. Sample size consideration was necessary and the power calculation would help in particular the difference between two diagnostic classifiers was not detected to be significant with achieved sample size. The calculated sample size in our Table 8 shows with paired design for AUC about 70% and for detection of an effect of 10%, the required sample size is 108 subjects for each group of cases and

controls with 80% power and 95% CI but for a desirable effect of 12%, this sample size is reduced to 71 for each group of cases and control.

11. Summary guideline

Listed below is recommendation for sample size calculation in diagnostic studies with respect type of study and classifier.

1. In a single diagnostic test with dichotomized outcomes, if an investigator interested in sensitivity and specificity of performance of diagnostic test, depending on conditions for estimating accuracy and desirable marginal error of estimates as outlined in formula in Sections 6.1 and 6.2, the required sample size can be chosen with respect to sensitivity (or specificity) and prevalence of disease in Tables 1 and 2.
2. For comparison of sensitivity (or specificity) of two diagnostic tests with dichotomized outcome for detection of a desirable effect between two tasks with 95% CI, choose the formula in Sections 6.2 and 6.3 for sample size calculation.
3. For a dichotomized classifier when the choice of accuracy index is the LR^+ or LR^-, depending on condition which is more interested from clinical prospective, one can choose the sample size in Tables 3 and 4 or using formula in Section 6.4.
4. When the diagnostic test results are recorded in ordinal or continuous scale, ROC curve analysis is the choice of interest and AUC is primarily of interested accuracy index. In estimating purpose, for a given AUC and desired marginal errors of the estimates as outlined in Table 5, select the optimal sample size.
5. For testing AUC with a pre-specified value in a single diagnostic task, depending on the AUC and the effect to be detected with 80% power and 95% CI, choose Table 6 for the required sample size.
6. For comparison of two independent AUCs, given a base value for AUC and desirable effect size with 80% power and 95% CI, select Table 7 for minimum sample size of each group.
7. For comparison of two diagnostic tasks in paired design, choose Delong method for estimation of SE of difference of AUCs of two correlated tasks. Select the possible value of AUC and the desirable effect size with 80% power and 95% CI in Table 8 for the optimal sample size of each group.
8. Additionally, in any of the above conditions, if the required sample size is not achieved in practice, the calculation of power of achieved sample size is needed if it differs from required sample size.

12. Discussion

In this review, we have shown the various sample size estimator methods for diagnostic studies. We provided the tables for the required samples size with different configurations of accuracy indexes and effect sizes/marginal errors. Our findings show that the required sample size not only depend on the effect size/marginal errors but also depend on the accuracy index for a given confidence level and power. Higher accuracy produces smaller sample size since higher accuracy has less room for sampling variations (i.e. less SE) in ROC space. A similar discussion is relevant regarding the range of ROC curve. The ROC curve is progressively located in the right corner of ROC space (AUC > 0.90), corresponding to lower sampling variability, as our results shows the required sample size for a given effect size and power is lower than ROC curve located toward the diagonal line (AUC = 0.60). This happens because of high sampling variability of AUC around diagonal line. Thus, it is important that the investigators compute sample size using an accuracy index that is relevant in clinical practice or from previous published data. Thus, the values of sensitivity/specificity, LR and AUC that are expected in clinical settings, should be considered. The various degrees of these parameters were presented in the 1st column of our tables of required sample sizes.

On the other hand, the sampling variability of accuracy index may depends on the underlying distribution of data of diagnostic test results and the model used to estimate ROC index and its SE. If the underlying distribution of test results is close to binormal, then the binormal SE of AUC is more reliable [37]. In a case of non-binormal data, particularly bimodal data, the Delong method of nonparametric estimate of SE of AUC represents the actual variation but binormal model produces a conservative estimate of SE [38].

In relation of sample size with LR^+ (or LR^-), the combination of various boundary values of LR^+ and point estimate that is derived from observed sensitivity and specificity, as boundary value largely differs from point estimate that corresponds with higher marginal errors, the calculated sample size becomes very small ($n < 30$). These sample sizes produce unreliable estimate of LR. Thus, we do not recommend such a sample size in clinical practice. For example, as presented in our results (Table 3) with observed value of $LR^+ = 1.75$, assuming the minimum value of $LR^+ = 6$ produced a small sample size of 5 for each group. The corresponding marginal errors are relatively high (the wide of confidence interval is 6–1.75 = 4.25 and thus the percentage size of marginal error is about $\frac{4.25}{1.75} = 242.8\%$ while for minimum boundary of 2.5 for LR^+, the percentage of error in wide of CI is about $\frac{0.75}{1.75} = 42.3\%$ that produced a sample size of 416 for each group.

The major component of sample size estimator is involved with sampling variability of accuracy under the null and alternative hypothesis. Regarding to AUC, at least three methods (Hanley and McNeil method [15], binormal estimator [22,33,24] and Delong's method [35,36] have been proposed for SE of AUC. Hanley and McNeil formula which uses exponential approximation for SE of nonparametric AUC (Wilcoxon statistic), has important role in inferences on ROC analysis for simplicity of its calculation but it only involves one parameter distribution (AUC). This method allows to estimate SE in a single modality and also for difference of two independent AUC. However, it does not provide the covariance between two correlated AUC. Hanley and McNeil method of sample size estimator underestimates the actual variation for rating data, particularly, when the standard deviation (SD) ratio differs greatly from one [25] while the binormal estimator of SE is more flexible and it allows including a wide range of ratio of two standard deviations. It produces a conservative estimate of SE that is preferable in sample size calculation [38]. With continuously distributed data, Hajian-Tilaki and Hanley [38] showed that the binormal estimator of SE of AUC reflects actual variation or it tends slightly to be greater than empirical SE with binormal data while it produces more conservative estimates of SE with nonbinormal data, particularly for bimodal data. In panning ROC study, generally, the method produces a conservative estimates of SE is a great of interest to determine sample size that will allow achieving a higher statistical power. Thus, we used binormal estimator in our sample size calculations and for simplicity the ratio of two SD was considered to be close to one.

The required sample size presented in Tables 1–8 do not guarantee for subgroup analysis because of lack of power in statistical test due to paucity of number of patients in each subgroup. The investigator may decide such analysis for exploration of his findings but we emphasize any subgroup analysis must be decided at the stage of design and sample size calculation in study protocol. However, it was rarely taken into account at this stage in clinical practices [19,20].

The methods of sample size calculation presented in this article, are based on asymptotic theory of normal approximation. This

assumption might be violated for a high accuracy indexes that is close to one but this occurs rarely in diagnostic studies. The new diagnostic test may usually have the accuracy at intermediate level (0.70–0.90). However, the exact method of confidence bound for accuracy based on binomial distribution particularly for sensitivity and specificity are recommended [42,43].

In comparative studies of two diagnostic tasks, obviously sample sizes are influenced by study design. In clinical practice, paired design is more efficient in term of precision of estimates (i.e. less SE for difference of two estimates of AUC) for a given sample size and power. In our sample size calculation for paired design when the two tests are applied consecutively on the same subjects, we assumed the correlation between two accuracy indexes to be close to 0.50. Such a correlation has been reported in clinical practices of diagnostic accuracy by Rocket et al. [41].

The half-width of a confidence interval as one of the quantities required in order to plan a sample size. While this is true for quantities that have symmetrical confidence intervals, this may be rather inappropriate for proportions especially ones such as sensitivity and specificity, for which we aim for high values. Thus, the simple symmetrical Wald interval has some well recognized limitations for use when actually calculating a CI for proportion, to such a degree that caution is needed even for its use in planning sample sizes. The better intervals are asymmetrical about the empirical estimates that have been discussed by several authors [44–46]. A similar issue applies to the intervals for the area under the ROC curve.

The other aspects of study design affect on sample size, for example, in some clinical settings, more than one unit of observation per patient are taken (2 or more lesion per person in evaluation of breast cancer screening by mammography) [40]. These units of observations from the same subjects are not independent statistically and cannot be analyzed as independent required sample size. Hajian-Tilaki et al. [47] developed methods for multiple signal data per patient that is based on structure components of Delong's method that are called as pseudo accuracy for each signal and each person. However, our tables of sample size assume one unit per patient. This might be a new area of research to develop statistical tools for sample size calculations when multiple lesions are as a unit of observation per patient.

It is necessary to determine that the required sample size for corresponding new statistical methods uses in diagnostic study and more simulation study needs to assess the performance of the different approaches of SE estimator of diagnostic accuracies in sample size and power calculations in diagnostic studies.

Appendix A

The variance of nonparametric AUC (Wilcoxon statistic) is estimated using the methods that proposed by Bamber [48] as

$$\mathrm{Var}(\widehat{\mathrm{AUC}}) = \frac{\mathrm{AUC}(1-\mathrm{AUC}) + (n_1-1)(Q_1-\mathrm{AUC}^2) + (n_2-1)(Q_2-\mathrm{AUC}^2)}{n_1 n_2} \tag{A1}$$

Hanley and McNeil [7] used exponential approximation to estimate Q_1 and Q_2 as

$$Q_1 = \frac{\mathrm{AUC}}{2-\mathrm{AUC}} \quad \text{and} \quad Q_2 = \frac{2\mathrm{AUC}^2}{1+\mathrm{AUC}}$$

that allows one to estimate the variance of AUC and its SE. Thus, the variance of AUC under the null and also alternative hypothesis can be estimated easily.

For studies with continuous test results $\mathrm{Var}(\widehat{\mathrm{AUC}})$ can be written approximately

$$\mathrm{Var}(\widehat{\mathrm{AUC}}) = \frac{Q_1}{r} + Q_2 - \mathrm{AUC}^2\left(\frac{1}{r}+1\right) \tag{A2}$$

where r the ratio of sample size of controls to cases ($r = \frac{n_2}{n_1}$).

Appendix B

The two parameters of ROC curves based on binormal assumption are defined as $a = \frac{\mu_2-\mu_1}{\sigma_1}$ and $b = \frac{\sigma_1}{\sigma_2}$ where μ_1 and σ_1 the mean and standard deviation of distribution for nondiseased and μ_2 and σ_2 are for diseased distribution respectively. The area under curve with binormal model is $\mathrm{AUC} = \varphi[\frac{a}{1+b^2}]$ where φ is the cumulative distribution function. Delta method can be used to estimate variance and SE of AUC. With an approximation when the ratio of SD is close to one (i.e. $b = 1$) the binormal estimator of variance of $(\widehat{\mathrm{AUC}})$ is

$$\mathrm{Var}(\widehat{\mathrm{AUC}}) = (0.0099 \times e^{-a^2/2}) \times \left(\frac{5a^2+8}{n_2} + \frac{a^2+8}{n_1}\right) \tag{A3}$$

where $a = \varphi^{-1}(\mathrm{AUC}) \times 1.414$ and n_1 and n_2 are the sample size for nondiseased and diseased [22].

Appendix C

Let X_i, $i = 1, 2, \ldots, n$ denote test results for a sample of n nondiseased subjects, and Y_j, $j = 1, 2, \ldots, m$ denote for m diseased subjects. For each (X_i, Y_j) pair, an indicator function $I(X_i, Y_j)$ is defined as follows:

$$\begin{array}{ll} I(X_i, Y_j) = 1 & \text{if } Y_j > X_i \\ 1/2 & \text{if } Y_j = X_i \\ 0 & \text{if } Y_j < X_i \end{array} \tag{A4}$$

The average of these I's over all $n \times m$ comparisons is the Wilcoxon or Mann Whitney U-statistic. The U is equivalent to the area under the empirical ROC curve, and the expected value of U is the area under the theoretical (population) ROC curve [15]. An alternative representation, used by DeLong et al. [35] is to define the components of the U-statistic for each of the n nondiseased subjects and for each of m diseased subjects as

$$V_N(X_i) = \frac{1}{m}\sum_{j=1}^{m} I(X_i, Y_j) \tag{A5}$$

$$V_D(Y_j) = \frac{1}{n}\sum_{i=1}^{n} I(X_i, Y_j) \tag{A6}$$

We will call the individual $V_N(X_i)$'s and $V_D(Y_j)$'s "pseudo-values" or "pseudo-accuracies". The pseudo-value, $V_N(X_i)$, for the ith subject in the nondiseased group, is defined as the proportion of Y's in the sample of diseased subjects where Y is greater than X_i. Likewise $V_D(Y_j)$, for the jth subject in the diseased group is defined as the proportion of X's in the sample of nondiseased subjects whose X is less than Y_j. The average of the $n\{V_N\}$'s and the average of the $m\{V_D\}$'s are both equivalent to the U-statistic.

Hanley and Hajian-Tilaki [36] restated the DeLong's method of calculating the variance of the accuracy index in a single diagnostic test is as follows:

$$\mathrm{Var}[U] = \frac{\mathrm{Var}[V_N]}{n} + \frac{\mathrm{Var}[V_D]}{m} \tag{A7}$$

The advantage of DeLong's method is to compare the areas under two or more ROC curves (1 curve per diagnostic system) obtained from the same sample of subjects. In a setting with two diagnostic systems, if we denote the estimates of accuracy indices of two compared systems by A_1 and A_2, then the $\mathrm{SE}(A_1 - A_2)$ is

based on the variation of pseudo-values $V_1 = \{V_{1N}, V_{1D}\}$, and $V_2 = \{V_{2N}, V_{2D}\}$ and the sample sizes investigated. The variance–covariance matrix is

$$\mathrm{Var}[A_1, A_2] = \frac{\boldsymbol{S_N}}{n} + \frac{\boldsymbol{S_D}}{m} \quad \text{(A8)}$$

where $\boldsymbol{S_N}$ and $\boldsymbol{S_D}$ represent the estimated variance–covariance matrices of the paired pseudo-values in the nondiseased and diseased groups, respectively. Therefore,

$$\mathrm{Var}(A_1, A_2) = \mathrm{Var}(A_1) + \mathrm{Var}(A_2) - 2\mathrm{Cov}(A_1, A_2) \quad \text{(A9)}$$

where each variance and each covariance is a sum of the corresponding components from S_N/n and S_D/m.

References

[1] Perry GP, Roderer NK, Asnar SA. Current perspective of medical informatics and health sciences librarianship. J Med Libr Assoc 2005;33:199–206.
[2] Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. Invest Radiol 1979;14:109–21.
[3] Lasko TA, Bhagwat JG, Zou KH, Ohno-Machao L. The use of receiver operating characteristic curves. JBM 2005;38:404–15.
[4] Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging 1989;29:307–35.
[5] Kummar R, Indrawn A. Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatr 2011;48:277–89.
[6] Zou KH, O'Malley AJ, Mauri L. Receiver operating characteristic analysis for evaluation diagnostic tests and predictive models. Circulation 2007;115:654–7.
[7] Figoria RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Med Inform Dec Mak 2012;12:8.
[8] Yao X, Wilczynski NL, Walter SD, Haynes RB. Sample size determination for bibliographic retrieval studies. BMC Med Inform Dec Mak 2008;8:43.
[9] Aphinaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc 2005;12:207–16.
[10] Chapman WW, Cooper GF, Hanbury P, Chapmann BE, Harrison LH, Wagner MM. Creating a text classifier to detect radiology reports describing mediastinal findings. Associated with inhalational anthrax and other disorders. J Am Med Inform Assoc 2003;10:494–503.
[11] Lu C, Van Gestel T, Suykens JAK, Van Huffel SV, Vergote I, Timmerman D. Preoperative prediction of malignancy of ovarian tumors using least squares support vector machines. Artif Intell Med 2003;28:281–306.
[12] Faraser HSF, Lonc WJ, Naimi S. Evaluation of a cardiac diagnostic program in a typical clinical setting. J Am Med Inform Assoc 2003;10:373380.
[13] Kramer M. Clinical epidemiology and biostatistics: a primer for clinical investigation and decision making. Berlin: Springer-Verlag; 1988. p. 201–19.
[14] Linnet K. Comparison of quantitative diagnostic tests: type I error, power, and sample size. Stat Med 1987;6:147–58.
[15] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
[16] Hanley JA, McNeil BJ. A method of comparing the area under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43.
[17] Jones SR, Charely S, Marison M. An introduction to power and sample size calculation. Emerg Med J 2003;20:453–8.
[18] Malhorta RK, Indrayan A. A simple nomogram for sample size for estimating sensitivity and specificity of medical tests. Indian J Ophthalmol 2010;58:519–22.
[19] Bochmann F, Johmson Z, Azuara-Balanco A. Sample size in studies on diagnostic accuracy in ophthalmology: a literature survey. Br J Opthalmol 2007;91:898–9000.
[20] Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample size of studies on diagnostic accuracy: literature survey. BMJ 2006;332:1127–9.
[21] Buderer NM. Statistical methodology: I. Incorporating the prevalence of disease into sample size calculation for sensitivity and specificity. Acad Emerg Med 1996;3:895–900.
[22] Obuchowski NA. Sample size calculation in studies of test accuracy. Stat Methods Med Res 1998;7:371–92.
[23] Charley S, Dosman S, Jones SR, Harison M. Simple nomograms to calculate sample size in diagnostic studies. Emerg Med J 2005;22:180–1.
[24] Simel DL, Samsa GP, Matchar DB. Likelihood ratio with confidence: sample size estimation for diagnostic test studies. J Clin Epidemiol 1991;44:763–70.
[25] Fosgate GT. Practical sample size calculations for surveillance and diagnostic investigations. J Vet Diagn Invest 2009;21:3–14.
[26] Li J, Fine J. On sample size for sensitivity and specificity in prospective diagnostic accuracy studies. Stat Med 2004;23:2537–50.
[27] Steinberg DM, Fine J, Chappell R. Sample size for positive and negative predictive value in diagnostic research using case control designs. Biostatistics 2009;10:94–105.
[28] Kumar R, Indrayan A. A nomogram for single-stage cluster-sample survey in community for estimation of a prevalence rate. Int J Epidemiol 2002;31:463–7.
[29] Beam CA. Strategies for improving power in diagnostic radiology research. Am J Roentgenol 1992;159:631–7.
[30] Fleiss JL, Levin B. Sample size determination in studies with matched pairs. J Clin Epidemiol 1988;58:859–62.
[31] Lachin JM. Power and sample size evaluation for the McNemar test with application to matched case-control studies. 1992; 11:1239–51.
[32] Connor RJ. Sample size for testing differences in proportions for the paired sample design. Biometrics 1987;43:207–11.
[33] Obuchowski NA. Computing sample size for receiver operating characteristic studies. Invest Radiol 1994;29:238–43.
[34] Metz CE, Herman BA, Shen J-H. Maximum-likelihood estimation of ROC curves from continuously-distributed data. Stat Med 1998;17:1033–53.
[35] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.
[36] Hanley JA, Hajian-Tilaki KO. Sampling variability of nonparametric estimates of the area under receiver operating characteristic curves: an update. Acad Radiol 1977;4(1):49–58.
[37] Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP. A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. Med Decis Making 1997;17:94–102.
[38] Hajian-Tilaki KO, Hanley JA. Comparison of three methods for estimation the standard error of the area under the curve in ROC analysis of quantitative data. Acad Radiol 2002;9:1278–85.
[39] Metz CE, Wang PL, Kronman HB. A new approach for testing the significance of differences between ROC curves from correlated data. In: Deconick F, editor. Information processing in medical imaging. The Hague: Nijhoff; 1984. p. 432–45.
[40] Obuchowski NA. Sample size tables for receiver operating characteristic studies. AJR 2000;175:603–8.
[41] Rockette HE, Campbell WL, Britton CA, Holbert JM, King JL, Gur D. Empirical assessment of parameters that affect the design of multiobserver receiver operating characteristic studies. Acad Radiol 1999;6:723–9.
[42] Flahault A, Cadilhac M, Thomas G. Sample size should be performed for design accuracy in diagnostic test studies. J Clin Epidemiol 2005;58:859–62.
[43] Cho H, Cole SR. Sample size calculation using exact methods in diagnostic test studies. J Clin Epidemiol 2007;60:1201–2.
[44] Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat 1998;52:119–26.
[45] Newcombe RG. Two sided confidence intervals for the single proportion: comparison of seven methods. Stat Med 1998;17:857–72.
[46] Newcombe RG. Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 2: asymptotic methods and evaluation. Stat Med 2006;25:559–73.
[47] Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP. An extension of ROC analysis to data concerning multiple signals. Acad Radiol 1977;4:225–9.
[48] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. J Math Psychol 1975;12:387–415.