

# SAMPLE SIZE DETERMINATION USING ROC ANALYSIS

Viktoriya Stalbovskaya<sup>1</sup>, Brahim Hamadicharef<sup>2</sup> and Emmanuel Ifeachor<sup>1</sup>

<sup>1</sup>University of Plymouth, UK; <sup>2</sup>Institute for Infocomm Research, Singapore  
e.ifeachor@plymouth.ac.uk

**Abstract:** The paper presents a new method of sample size determination (SSD) based on performance evaluation of systems under study. The method builds upon previous work on Bayesian approach to nonparametric receiver operating characteristics (ROC) analysis with estimation of probability density functions and confidence intervals for parameters of ROC curve. Technical details of the method together with an illustration of its application to a real clinical example is given. The advantages of the method include ability to handle and compare models built using multidimensional, multimodal data, with no need for prior estimation of the effect size and variability. The disadvantages of the approach include high computational requirements, especially when the method of SSD deploys AUC as a measure of systems performance.

**Keywords:** sample size determination, ROC analysis, decision support systems

## INTRODUCTION

Determination of sample size is a very important part of the design of biomedical studies. A sufficient sample size allows the researcher to report his/her results with sufficient degree of confidence and acceptable statistical power.

Sample size determination for clinical studies involves identification of the following parameters: minimum expected difference, estimated measurement variability, desired statistical power, and statistical significance criterion [1]. The first parameter, minimum expected difference or effect size, is the smallest measured difference between groups that is likely to be detected. The smaller the difference is, the larger the sample size should be. The measurement variability inversely affects sample size. Identification of these two parameters requires either previous pilot studies or a literature review. Statistical power for a randomised clinical trials is customarily set to  $\geq 0.80$ . Significance criterion is the maximum  $p$ -value for which a difference is to be considered statistically significant.

Although the area of SSD is well established, extensive review of SSD methods in the area of descriptive studies by Adock [2] shows that there are significant gaps in the availability of SSD procedures, particularly in multivariate analysis. Traditional methods of SSD are not suitable for the design of intelligent decision support systems which are based on complex non-linear methods (e.g. artificial neural networks). For biomarker development traditional sample size calculations are not relevant as there are too many factors that have to be taken into

account, including the prevalence of the disease, number of biomarkers, ability of the biomarker to discriminate among different subgroups, and the type of analysis technique that are used (e.g. statistical or machine learning [3]).

One way of finding another paradigm for SSD methods is to exploit the theory and analysis of receiver operating characteristics. Sample size calculations for phase 2 of clinical studies using ROC were described in [4]. The underlying assumption was based on identification of acceptable and anticipated performance levels ( $\{TP, FP\}_{0,1}$ ), and for a chosen significance level ( $\alpha$ ) and power ( $1-\beta$ ) the sizes of the main and control groups were calculated. Another method of SSD for diagnostic accuracy studies involves binormal ROC curve indices for continuous test results [5].

In this paper we propose a new method of sample size determination for intelligent decision support systems. The method uses ROC parameters of system's performance. It is a continuation of our work on Bayesian estimation of confidence intervals for ROC curves [6].

The method is particularly useful in the design of clinical studies and testing medical decision support systems. We use our method to answer a question: *how many cases do we need to employ in order to prove that there is a significant difference in performance between two diagnostic systems under study?*

To confirm that two systems have a significantly different performance one has to build confidence intervals for performance indicators for each system (e.g. AUC, partial AUC, ROC points) and then demonstrate a significant difference between them. Our method works particularly well for small sample size problems and therefore has a high potential applicability for either evaluation and design of clinical studies, such as the development of complex biomarker for diseases involving multimodal and multidimensional data.

The remainder of the paper is organised as follows. In the next section, the new method of sample size estimation is presented followed by example of its application. Finally, we present the discussion, conclusion and future work.

## METHOD

The underlying method constructs a probability density function over the entire ROC graph for each point of the curve. A pdf surface for the hit rate and false alarm rate is generated according to [6] so that contour lines of the surface can be drawn that represents 95% confidence interval for a ROC point. To represent the surface, it is

divided into a fine grid and the probability of each quantised grid cell is calculated by integrating the probability at a point, over the area of each grid cell. Considering  $\mathbf{X}$  and  $\mathbf{Y}$  as discrete distribution of hit rate and false alarm rate, respectively, the surface is defined as

$$\text{Surface} = \mathbf{X} \cdot \mathbf{Y}^T \quad (1)$$

Full details of the rationale for the method and the formulas for calculation of  $\mathbf{X}_i$  and  $\mathbf{Y}_i$  are given in [6].

To illustrate the method lets consider a hypothetical example of building confidence intervals using bayesian approach for ROC curve [7]. The confusion matrix is as follows. The columns give the Gold standard diagnosis, the rows the test results, correct or incorrect.

		Gold standard	
		Diseased	Healthy
Test	Diseased	$b_0 = 4$	$a_0 = 0$
	Unknown	$b_1 = 1$	$a_1 = 3$
	Healthy	$b_2 = 0$	$a_2 = 7$

There are two ROC points out of this table with coordinates:

$$\begin{aligned} x_0 &= \frac{a_0}{a_0 + a_1 + a_2} \\ y_0 &= \frac{b_0}{b_0 + b_1 + b_2} \\ x_1 &= \frac{a_0 + a_1}{a_0 + a_1 + a_2} \\ y_1 &= \frac{b_0 + b_1}{b_0 + b_1 + b_2} \end{aligned}$$

Assuming a binomial distribution of the events one can estimate pdf functions of hit rate and false alarm rate. Probability of an ROC point with coordinates  $x$  and  $y$  is calculated using the *Beta* function [6]:

$$\text{PointProb}_{xy} = \frac{x^{a_0}(1-x)^{a_1}}{\frac{a_0!a_1!}{(a_0+a_1+1)!}} \cdot \frac{y^{b_0}(1-y)^{b_1}}{\frac{b_0!b_1!}{(b_0+b_1+1)!}} \quad (2)$$

To define a surface introduced in (1) one has to identify vectors  $\mathbf{X}$  and  $\mathbf{Y}$ . Please refer to [6] for details the rationale and derivation of formulas. For simplicity, here we provide only final formulas for multiple ROC points in (3) and (4).

To calculate factorial for large numbers of  $a_{0,1}$ ,  $b_{0,1}$  we used gamma function  $n! = \Gamma(n + 1)$  for any natural number  $n$ .

Confidence boundaries for ROC points were obtained by taking a contour from the ROC surface for a given threshold value.

Fig. 1 shows a pdf surface for hit rate and false alarm rate, or in terms of medical diagnostic systems, sensitivity and 1-specificity. Fig. 2 presents the confidence boundaries for two ROC points. The distribution of AUC value for the simple example is given in Fig. 3

The area under the ROC curve (AUC) was calculated using two methods: (i) Bayesian posterior distribution

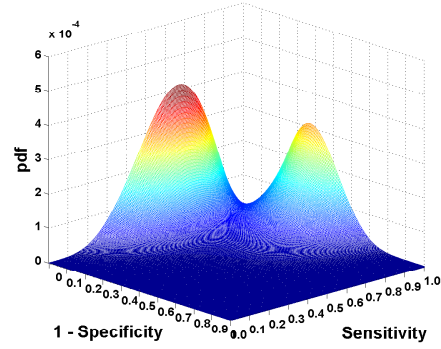


Fig. 1: Example of ROC surface.

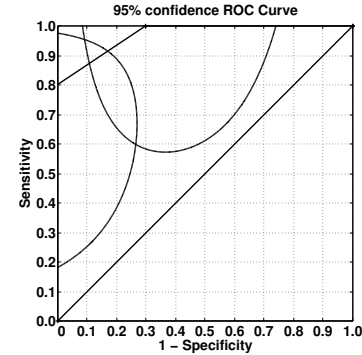


Fig. 2: Example of contour of ROC surface representing 95% confidence boundaries for ROC points.

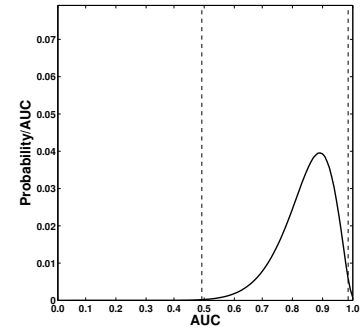


Fig. 3: Example of bayesian posterior distribution of AUC,  $G = 128$ ,  $t = 20.5$  minutes.

and (ii) trapezoid rule for the ROC points. The algorithm for finding Bayesian posterior distribution of AUC is based on an algorithm of finding the shortest path between two nodes on a general graph, where beginning and end points are (0,0) and (1,1), respectively, and the intermediate points presented by ROC points,  $x, y_i$ . The ROC graph is quantized into a grid of  $G \times G$  cells, with a curve approximated by a series of lines joining these cells. However the algorithm is computationally expensive and requires a time proportional to  $G^5$  and memory proportional to  $G^3$ .

The formulas for confidence limits of AUC for frequentist approach

$$AUC_1 = \frac{AUC + \frac{z^2}{2n} - z\sqrt{\frac{AUC(1-AUC)}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

$$AUC_2 = \frac{AUC + \frac{z^2}{2n} + z\sqrt{\frac{AUC(1-AUC)}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

where  $AUC_{1,2}$  are the lower and upper limit of confidence interval,  $z$  has normal distribution and for 95% confidence interval has value of 1.96,  $n$  is the sample size.

*Comparison of confidence intervals for ROC points* involves the comparison of two-dimensional contours of CIs. Rule for significantly different contours is

$$A_1 \cup A_2 = 0, \quad (5)$$

where  $A_m$  is an area bounded by ROC contour for model  $m = 1, 2$ .

*Comparison of confidence intervals for AUC* is straightforward as it is a one-dimension value. The models have significantly different AUCs if

$$LL_1 > UL_2 \vee UL_1 < LL_2, \quad (6)$$

where  $LL_m$  is a lower confidence limit of AUC for model  $m = 1, 2$ ,  $UL_m$  is an upper confidence limit of AUC for model.

The algorithm for sample size determination using confidence estimates of ROC involves the following steps:

1. Select indicator of models' performance: ROC points, AUC.
2. Select method of sample size increment.
3. Calculate confidence intervals for a chosen ROC parameter using either bayesian or frequentist approach.

$$\mathbf{X}_i = (a_0 + a_1 + 1)! \sum_{k=0}^{a_1} \frac{\left(\frac{i}{n}\right)^{a_0+a_1+1-k} \left(1 - \frac{i}{n}\right)^k - \left(\frac{i-1}{n}\right)^{a_0+a_1+1-k} \left(1 - \frac{i-1}{n}\right)^k}{k!(a_0 + a_1 + 1 - k)!} \quad (3)$$

$$\mathbf{Y}_i = (b_0 + b_1 + 1)! \sum_{k=0}^{b_1} \frac{\left(\frac{i}{n}\right)^{b_0+b_1+1-k} \left(1 - \frac{i}{n}\right)^k - \left(\frac{i-1}{n}\right)^{b_0+b_1+1-k} \left(1 - \frac{i-1}{n}\right)^k}{k!(b_0 + b_1 + 1 - k)!} \quad (4)$$

for all  $i$  from  $i = 1$  to  $i = n$

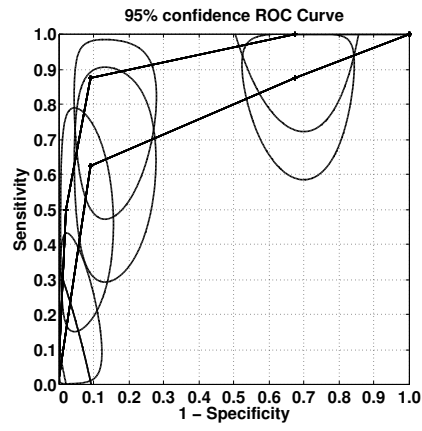


Fig. 4: Initial data  $n = 51$  for pancreatic data.

4. Compare confidence intervals for the chosen ROC parameter between models.
5. If the ROC parameters are significantly different, i.e. (5) or (6) is true, then the existing sample size is large enough and the algorithm stops. Otherwise, follow the next step.
6. Increase sample size according to preselected method on the step 2.
7. Repeat steps 3-5 until the condition in step 5 is true.

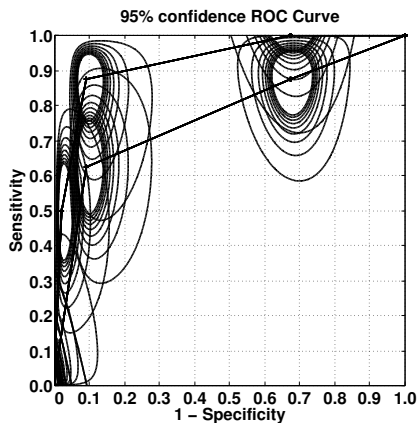
To illustrate our method of sample size determination we applied the methodology above to a pancreatic dataset described in [8]. The total number of subjects was 51. Initial ROC contours for ROC points is presented on 4. Sample size was incremented  $n$ -fold.

As can be seen, in 4 initial ROC contours overlaps greatly, so for the present number of cases it is not possible to separate ROC contours and the difference between two groups is not significant. We applied simple step-wise increment of elements in confusion matrix in order to identify sample size,  $n$ , satisfying (5).

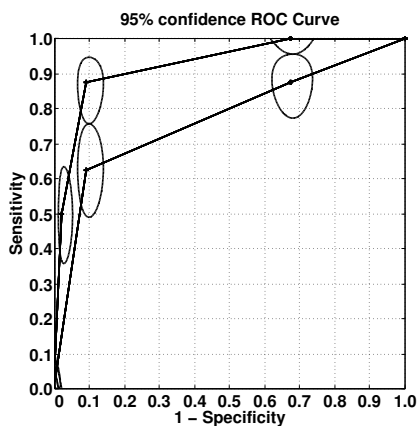
Fig. 5 illustrates a set of contours with the final results shown in 6 when the ROC contours for the two models do not overlap. Thus, for this example the total number of cases required to state that there is significant difference between two models is 459.

## CONCLUSION AND FUTURE WORK

The paper presents an extension of our work in the area of the performance evaluation. Here a new approach to sample size determination (SSD) is introduced which is



**Fig. 5:** Stepwise decrement of CI contours for pancreatic data.



**Fig. 6:** Significant difference is obtained for  $n = 459$  for pancreatic data.

based on nonparametric ROC analysis. Description of the method is illustrated by its application of it on a real clinical example.

The advantages of the method include ability to handle and compare models built on multidimensional multimodal data, no effect size and variability has to be specified. Potential impact of the method is in estimation of the sample size for prospective diagnostic clinical trials with adaptive design [9] when there is a possibility to adjust sample size during a trial along with other dynamical modifications of a trial course.

To disadvantages of the approach one can refer a high computational requirements, especially a method of SSD deploying AUC as a measure of systems performances. This drawback can be overcome by employing high-throughput computing technologies, such as GRID.

Part of our future work is to investigate possibility of analytical solution for the method as the current version is implemented as an iterative resources-consuming algorithm. We plan to develop MATLAB toolbox with GUI interface to enable real-life usage and testing of the method.

## ACKNOWLEDGMENTS

We acknowledge the financial support of the European Commission (The BIOPATTERN Project, Contract No. 508803) for this work.

## REFERENCES

- [1] J. Eng, "Sample size estimation: how many individuals should be studied?" *Radiology*, vol. 227, no. 2, pp. 309–13, May 2003.
- [2] C. Adcock, "Sample size determination: a review," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 46, no. 2, pp. 261–283, Jul 1997.
- [3] M. Pepe, R. Etzioni, Z. Feng, J. Potter, M. Thompson, M. Thornquist, M. Winget, and Y. Yasui, "Phases of biomarker development for early detection of cancer," *J Natl Cancer Inst*, vol. 93, no. 14, pp. 1054–61, Jul 2001.
- [4] M. Pepe, *The statistical evaluation of medical tests for classification and prediction*. New York: Oxford University Press, 2003, ch. Sample Size Calculations for Phase 2 Studies.
- [5] N. Obuchowski and D. McClish, "Sample size determination for diagnostic accuracy studies involving binormal roc curve indices," *Stat Med*, vol. 16, no. 13, pp. 1529–1542, Jul 1997.
- [6] J. Tilbury, P. Van Eetvelt, J. Garibaldi, J. Curnow, and E. Ifeachor, "Receiver operator characteristic analysis for intelligent medical systems - a new approach for finding confidence intervals," *IEEE Trans Biomed Eng*, vol. 47, no. 7, pp. 952–963, 2000.

- [7] J. Tilbury, P. Van-Eetvelt, J. Curnow, and E. Ifeachor, "Objective evaluation of intelligent medical systems using a Bayesian approach to analysis of ROC curves," *Proceedings of the 1st International Conference on Computational Intelligence in Medicine and Healthcare (CIMED'03)*, Sheffield, United Kingdom, July, 2003, 2003.
- [8] K. P. Adlassnig and W. Scheithauer, "Performance evaluation of medical expert systems using ROC curves," *Computers and Biomedical Research*, vol. 22, no. 4, pp. 297–313, 1989.
- [9] H. L. Golub, "The need for more efficient trial designs," *Statistics in Medicine*, vol. 25, no. 19, pp. 3231–3235, Oct 2006.