FREQUENTIST VERSUS BAYESIAN APPROACHES FOR AUC CONFIDENCE INTERVALS BOUNDS

Brahim Hamadicharef

Tiara #22-02
1 Kim Seng Walk
Singapore 239403

ABSTRACT

In this paper we first present two approaches, Frequentist and Bayesian, to calculate the Confidence Interval (CI) of Area Under the Curve (AUC). The goal of this study is to compare both approaches and find out if they reveal significant differences along the sample size.

We first generate a large number of hypothetical cases, based on True Negative (TN), True Positive (TP), False Positive (FP) and False Negative (FN), that lead to to specific AUC values (90, 85, 80, 75, etc.). We then use both Frequentist and Bayesian approach to calculate the AUC CI bounds, AUC_L and AUC_H, and plot them for visual comparison.

Results indicate that 1) for one sample size value the Bayesian approach can have multiple AUC CI bounds values, while the Frequentist has unique set of bounds, 2) for all sample size, the AUC_L and AUC_U values using the Frequentist approach are consistently under-estimated compared to the Bayesian ones, and 3) for very large sample size both approaches converge toward same values.

1. INTRODUCTION

In research fields such as machine learning, pattern recognition, data mining, medical diagnosis, etc. performance evaluation results are typically claimed in terms of sensitivity, sensibility and accuracy. However these measures are limited in the sense that they do not provide any sense of scale related to sample size. To indicate the reliability of such measure, Confidence Intervals (CI) need to be calculated based on sample size.

The choice between the Frequentist and Bayesian approach is an important aspect in performance evaluation. From their fundamental definition, they imply that the Frequentist approach is based on the assumption of large sample size while the Bayesian approach is more suitable for small sample size.

In this paper we are interested to compare the difference between both approaches as we increase the sample size from very small to very large. In practice, the majority of results presented in the scientific literature has often limited value, as based on small sample size. Therefore this makes the choice of an appropriate approach to calculate AUC CI bounds is of great importance.

Based on previous work [7], we first present both Frequentist and Bayesian approaches. The Bayesian approach in particular is based on Receiver Operating Characteristic (ROC) analysis, and was developed for the performance evaluation of intelligent medical systems [9].

The rest of the paper is organized as follow. In Section 2, we present the two main approaches to calculate AUC CI. In Section 3 we present results when comparing both approaches. Finally, we conclude the paper in Section 4.

2. APPROACHES FOR AUC CI

Considering a 2-class medical diagnosis, e.g. diagnosis of ovarian cancer [10], thus having four possible outcomes: True Positive (TP) when the tumor is malignant and diagnosed correctly, True Negative (TN) when the tumor is benign and diagnosed correctly, False Positive (FP) when the tumor is benign but diagnosed incorrectly as malignant, and False Negative (FN) when the tumor is malignant but diagnosed incorrectly as benign. Using the parameter set {TN,TP,FP,FN}, one can calculate the sensitivity (TP/(TP+FN)) and specificity (TN/(TN+FP)), and use them to plot points of the Receiver Operating Characteristic (ROC) curve (i.e. Sen vs 1-Spe) and calculate the Area Under the Curve (AUC) [3].

2.1. Frequentist Approach

Inspired from Wilson's score method [11][6], a Frequentist approach was proposed in [7] to calculate, for a specific confidence level (1 - alpha, with alpha 0.05 and 0.01, for respectively, 95% and 99%), the lower (AUC_L) and upper (AUC_U) AUC CI bounds:

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

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

where n is the sample size, and z has normal distribution with a value of 1.96 for 95% CI (2.577 for 99% CI). To

978-1-4244-7167-6/10/$26.00 ©2010 IEEE

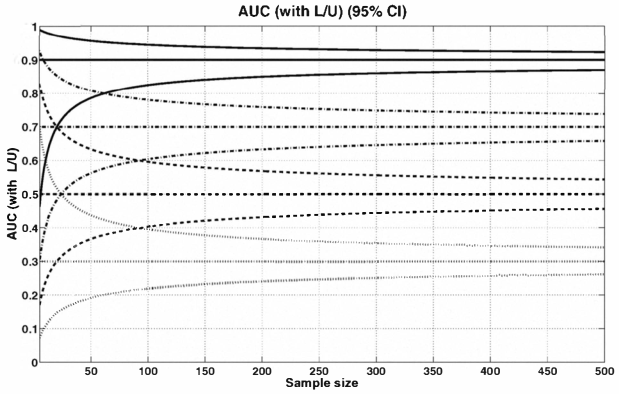


Fig. 1. AUC with AUC_L and AUC_U using the Frequentist approach

illustrate the variations of AUC_L and AUC_U using the Frequentist approach for an alpha level of 0.05 (i.e. 95% CI) and 0.01 (i.e. 99% CI), we plot examples for AUC (0.9, 0.7, 0.5, and 0.3) in Figure 1 for 95% CI.

It is worth noticing that by definition, the validity of the Frequentist approach should only be for large sample size. However, in practice, as only relatively small sample size are often used one should look for a Bayesian approach. Another limitation of the Frequentist approach is that formulae do not detail n=TP+TN-FP+FN, whereas as we will show with the Bayesian approach, there could be different values AUC_L and AUC_U for one value of n.

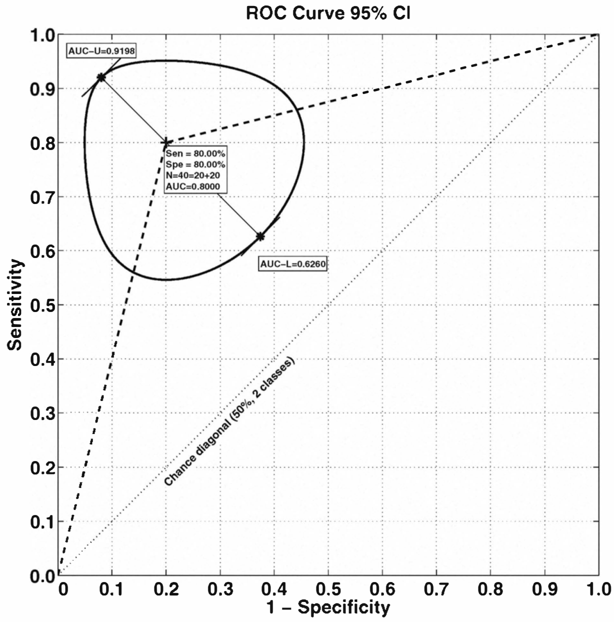
2.2. Bayesian Approach

Based on the original work by Tilbury [9], the Bayesian approach [8] is a methodology, derived from first principles, which calculates the probability density function (PDF) for each point on a ROC curve for any given sample size, and use them to plot the CI contour at a specified alpha level. The method, validated by Monte Carlos simulations, was shown to be accurate and robust, and most importantly not having any assumptions on the distribution. A graph search method was proposed in [9] to find values of AUC_L and AUC_U, and applied to the issue of Sample Size Determination (SSD) in [7]. However one major limitation of the graph search is its computation time.

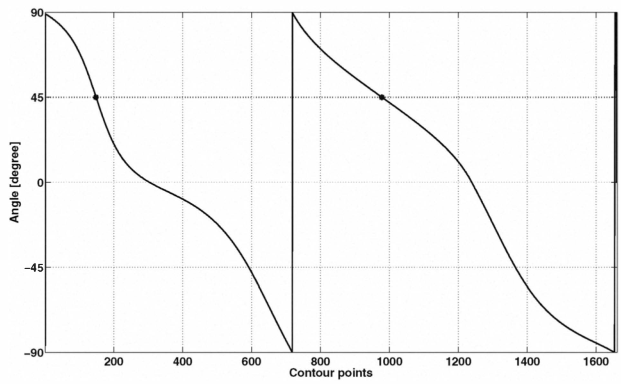
In this paper, results data from the contour graph are used to obtain AUC_L and AUC_U. Following the contour plot, we extract points coordinates and with a slope equation from a line passing via two points (x1 y1) and (x2 y2), we obtain the orientation angle, θ, as:

$$\theta = atan\left(\frac{y2 - y1}{x2 - x1}\right) \tag{3}$$

An example of the trajectory of the angle (slope of the tangent) along the contour is shown in Figure 2(b). A simple test for crossing the 45 degree line detects the contour points corresponding to the tangent of the contour. Even with a large square grid (2048^2) for the ROC PDF calculation resulting in a contour with a large number of points (more than 1500), the exact position of the two



(a)



(b)

Fig. 2. Example with a) ROC points with 95% CI contour, and b) Angle trajectory along the 95% CI contour at the ROC point

ROC points of interest are found very rapidly. As shown in Figure 2(b) the angle of the contour trajectory is cutting the 45 degree (oblique) twice, corresponding to the position of the two ROC points at AUC_L and AUC_U. The accuracy of the resulting AUC CI bounds depends on the grid size, however the calculation is faster than the method in our previous work [7]. At these ROC point coordinates (RocPtX and RocPtY) we use the trapezoidal rule to calculate the AUC.

3. RESULTS

We elaborated the following procedure. We enumerate the parameter set {TN,TP,FP,FN} that gives exact sensitivity (e.g. 90%) and specificity (e.g. 90%), thus exact AUC, i.e. 0.9. Using AUC and n, we calculate AUC_L and AUC_U using the Frequentist approach with Eq.(1) and Eq.(2). Using the same values of the {TN,TP,FP,FN} parameter set, we also calculate AUC_L and AUC_U using the Bayesian approach.

Our initial hypothesis was that there would be a clear difference between both approaches and that we could obtain an interval of $n \in [N_L \dots N_U]$ within which such difference would be very small, and also identify n_{sss} at which such difference was minimal, defining a *small sample size*.

However, as shown in the examples in Figure 3, the difference between both approach and for various AUC is constant. There is no value of sample size n_{sss} at which the difference is minimal. We can only assume that the Frequentist approach behaves better than expected for small sample size. It is also clear from Figure 3(a) to Figure 3(f) that with the Bayesian approach using the parameter set {TN,TP,FP,FN} we can obtain many different AUC CI bounds for each sample size. This issue gradually fades away for large sample size, converging to only one value of AUC_L and AUC_U.

We observe that AUC_L and AUC_U values using the Frequentist approach are consistently under-estimated in relation to the Bayesian ones. Furthermore, the Bayesian approach shows that for a sample size n there could be different AUC_L and AUC_U values, while the Frequentist approach provides one unique pair of bounds.

4. CONCLUSIONS

In this paper, we presented both Frequentist and Bayesian approaches to calculate AUC CI bounds, with an aim to compare them from very small to very large sample size. We defined a procedure to enumerate {TN,TP,FP,FN} parameter sets to obtain exact sensitivity and specificity, thus exact AUC values. These are then used to calculate and compare, using both Frequentist and Bayesian approaches, AUC_L and AUC_U the AUC CI bounds.

The Bayesian approach has the advantage to give exact AUC CI bounds for all possible {TN,TP,FP,FN} parameter sets at a specific sample size, this is an important aspect for medical diagnosis. We also observed from the results that the Frequentist approach consistently under-estimates the AUC CI bounds compared to the Bayesian ones. Finally, as expected both approaches converge towards the same CI bounds when the sample size become very large.

Future work will be re-evaluating the performance of medical systems, in particular when studies have small sample size, such as for EEG-based detection of Alzheimer Disease (AD) [1][2] and diagnosis models in gynecology and obstetrics [5][10]. A comparison with tailed Jeffreys prior interval [4] will also be investigated.

5. ACKNOWLEDGEMENTS

The contributions and assistance of Dr J. Tilbury and Dr V. Stabovskaya are gratefully acknowledged.

6. REFERENCES

[1] D. Abásolo, R. Hornero, C. Gómez, M. García, and M. López, "Analysis of EEG background activity in Alzheimer's disease patients with LempelZiv complexity and central tendency measure," *Medical Engineering and Physics*, vol. 28, no. 4, pp. 315–322, May 2006.

[2] B. Hamadicharef, C. Guan, E. Ifeachor, N. Hudson, and S. Wimalaratna, "Performance Evaluation and Fusion of Methods for Early Detection of Alzheimer Disease," *Proceedings of the International Conference on BioMedical Engineering and Informatics (BMEI2008), Sanya, Hainan, China, May 27–30, 2008*, pp. 347–351.

[3] J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, no. 1, pp. 29–36, April 1982.

[4] H. Jeffreys, "An Invariant Form for the Prior Probability in Estimation Problems," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.

[5] K. S. Khan, P. F. W. Chien, and L. S. Dwarakanath, "Logistic Regression Models in Obstetrics and Gynecology Literature," *Obstetrics and Gynecology*, vol. 93, no. 6, pp. 1014–1020, June 1999.

[6] R. G. Newcombe, "Interval estimation for the difference between independent proportions: comparison of eleven methods," *Statistics in Medicine*, vol. 17, no. 8, pp. 873–890, April 1998.

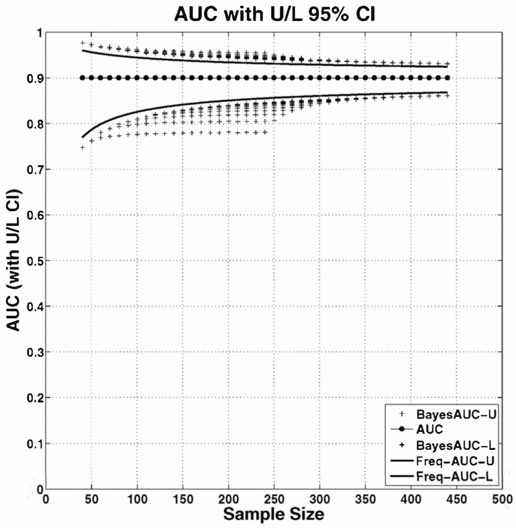
[7] V. Stalbovskaya, B. Hamadicharef, and E. C. Ifeachor, "Sample Size Determination using ROC Analysis," *Proceeding of the 3rd International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007), Plymouth, U.K., July 25–27, 2007.*

[8] 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*, pp. 85–90.

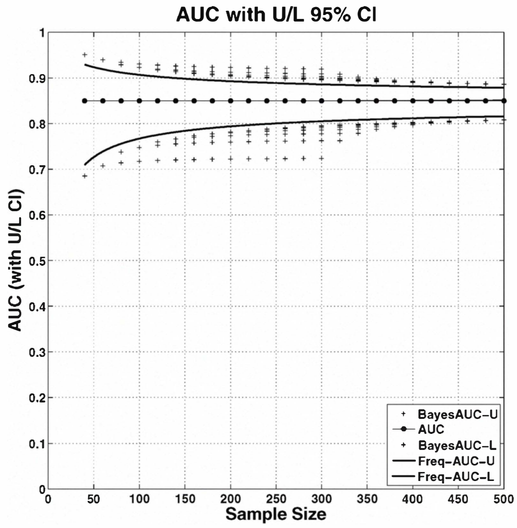
[9] J. B. Tilbury, "Evaluation of Intelligent Medical Systems," Ph.D. thesis, Department of Communications and Electronic Engineering (DCEE), University of Plymouth, Drake Circus, Plymouth PL4 8AA, Devon, United Kingdom, September 2002.

[10] D. Timmerman, T. Bourne, A. Tailor, W. P. Collins, H. Verrelst, K. Vandenberghe, and I. Vergote, "A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: the development of a new logistic regression model," *American Journal of Obstetrics and Gynecology*, vol. 181, no. 1, pp. 57–65, July 1999.

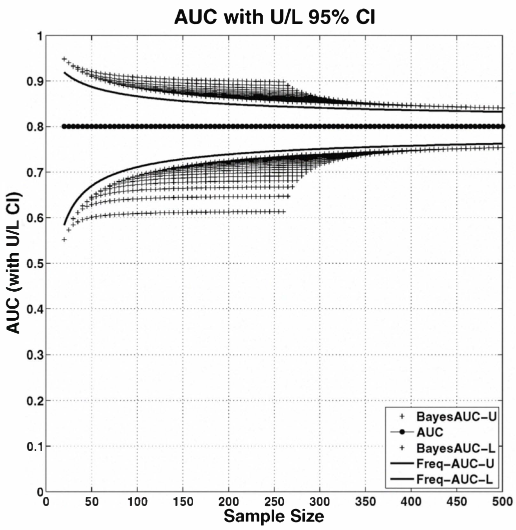
[11] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, pp. 209–212, 1927.



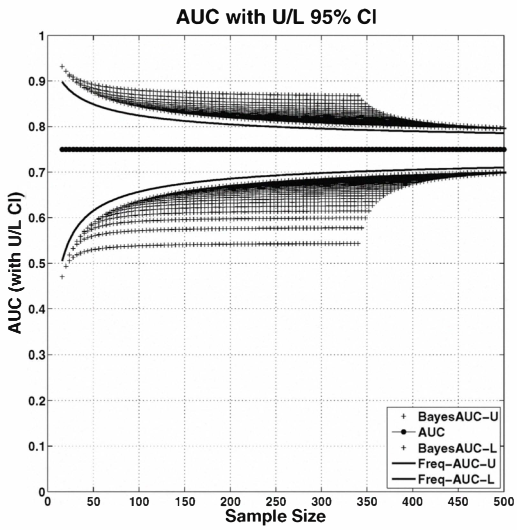
(a) AUC = 0.95



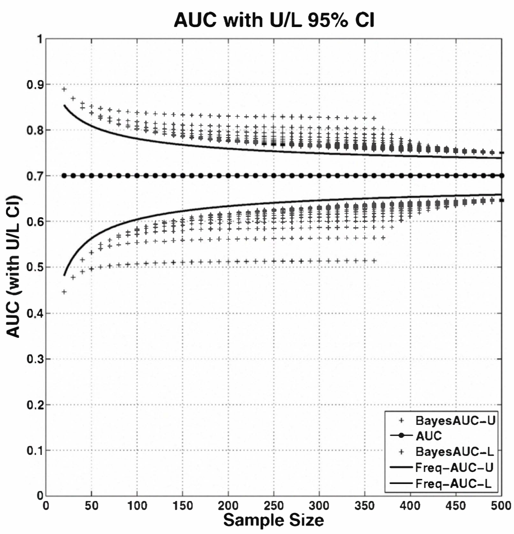
(b) AUC = 0.85



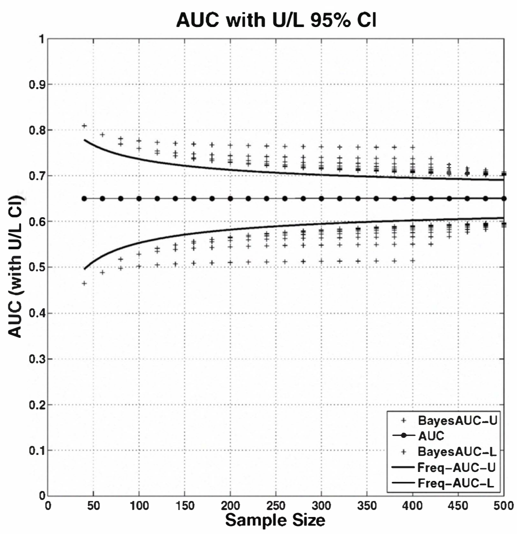
(c) AUC = 0.80



(d) AUC = 0.75



(e) AUC = 0.70



(f) AUC = 0.65

Fig. 3. Plot of AUC 95% CI using Frequentist and Bayesian approaches for various AUC