6-1-2021 2:00 PM

Sample Size Formulas For Estimating Areas Under the Receiver Operating Characteristic Curves With Precision and Assurance

Grace Lu, *The University of Western Ontario*

Supervisor: Choi, Yun-Hee, *The University of Western Ontario*
Joint Supervisor: Zou, Guangyong, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Epidemiology and Biostatistics

Abstract

The area under the receiver operating characteristic curve (AUC) is commonly used to quantify the discriminative ability of tests with ordinal or continuous test data. When planning a study to evaluate a new diagnostic test, it is important to determine a minimum sample size required to achieve a prespecified precision of estimating AUC. However, conventional sample size formulas do not consider the probability of achieving a prespecified precision, resulting in underestimation of sample sizes. To incorporate the assurance probability, asymptotic sample size formulas were derived using different variance estimators for AUC in this thesis. The precision of AUC estimations was quantified by either lower confidence limits or interval width. The performance of proposed sample size formulas was evaluated through simulation studies. Simulation results show that the formula based on lower limits with the nonparametric method performs best and can be used with both ordinal and continuous data. The methods are illustrated with examples from previously published data.

Keywords: assurance probability, confidence interval, variance

Summary for Lay Audience

The area under the receiver operating characteristic curve (AUC) is a tool used for describing the discriminative ability of diagnostic tests. Discriminative ability must be evaluated before adopting a test and using it in practice. An important factor to consider when planning an evaluation study is the minimum required sample size, as too small a sample size would make it difficult to see desired results, and too large a sample size may cause resources to be wasted. Typically, sample sizes are calculated using sample size formulas, however, existing sample size formulas tend to underestimate the required sample size because they do not consider the assurance probability of achieving a prespecified level of precision. In this thesis, we derived sample size formulas that incorporate this prespecified assurance probability. As sample size formulas require the variance of the AUC, we chose three different variance formulas to use. Simulation studies were conducted to evaluate the performance of sample size formulas. The results show that the formula based on lower limits with nonparametric method performed best and can be used with both ordinal and continuous data.

Acknowledgements

I would like to express my appreciation for everyone who has been a part of my journey to getting a Master's degree at Western University.

First and foremost, I would like to thank my supervisors, Dr. Guangyong Zou and Dr. Yun-Hee Choi, for supporting me throughout my study. Their kindness and guidance were vital in the completion of this thesis, and they motivated me to work hard and to never give up.

I would also like to thank the Department of Epidemiology and Biostatistics for providing financial support as well as an excellent environment for me to complete my courses and thesis. I have learned a lot over the past couple of years.

I am grateful for my friend and classmate, Zhenni Xue, for the moral support throughout my experience. It was great to be able to discuss courses and thesis with her, and she offered me good advice and suggestions.

Lastly, I would like to especially thank my parents and partner for their love and support. I'm thankful that they were by my side the entire time and have been nothing but encouraging.

Table of Contents

Abstract ii

Summary for Lay Audience iii

Acknowledgements iv

List of Tables vii

Chapter 1 Introduction **1**

1.1 The Receiver Operating Characteristic curve 1

1.2 Area under the ROC curve 3

1.3 Sample size estimation 6

1.4 Confidence intervals for AUC 8

1.5 Variance of AUC 8

1.6 Objective of thesis 9

1.7 Organization of thesis 10

Chapter 2 Literature Review **12**

2.1 History of ROC curve 12

2.2 AUC and Mann-Whitney statistic 14

2.3 Variance of AUC 17

2.4 Sample size estimation based on confidence intervals 19

Chapter 3 Methods **21**

3.1 Introduction 21

3.2 Existing variance formulas 22

3.3 Delta method for logit transformation 27

3.4 Sample size estimation 29

Chapter 4 Simulation 40

4.1 Achieving a prespecified lower limit 40

4.2 Achieving a prespecified confidence interval width 58

4.3 Nonparametric method using pilot data 81

4.4 Conclusion 86

Chapter 5 Illustration 88

Chapter 6 Discussion 92

Bibliography 96

VITA 99

List of Tables

4.1 a: Empirical assurance probabilities at the 50% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.9$. 49

4.1 b: Empirical assurance probabilities at the 50% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.8$. 50

4.1 c: Empirical assurance probabilities at the 50% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.7$. 51

4.2 a: Empirical assurance probabilities at the 80% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.9$. 52

4.2 b: Empirical assurance probabilities at the 80% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.8$. 53

4.2 c: Empirical assurance probabilities at the 80% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.7$. 54

4.3 a: Empirical assurance probabilities at the 90% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.9$. 55

4.3 b: Empirical assurance probabilities at the 90% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.8$. 56

4.3 c: Empirical assurance probabilities at the 90% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.7$. 57

4.4 a: Empirical assurance probabilities at the 50% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.9$. 62

4.4 b: Empirical assurance probabilities at the 50% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.8$. 63

4.4 c: Empirical assurance probabilities at the 50% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.7$. 64

4.5 a: Empirical assurance probabilities at the 80% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.9$. 65

4.5 b: Empirical assurance probabilities at the 80% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.8$. 66

4.5 c: Empirical assurance probabilities at the 80% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.7$. 67

4.6 a: Empirical assurance probabilities at the 90% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.9$. 68

4.6 b: Empirical assurance probabilities at the 90% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.8$. 69

4.6 c: Empirical assurance probabilities at the 90% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.7$. 70

4.7 a: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 50% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.9$. 72

4.7 b: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 50% assurance level such that the half-width of a two-sided 95%

confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.8$. 73

4.7 c: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 50% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.7$. 74

4.8 a: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 80% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.9$. 75

4.8 b: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 80% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.8$. 76

4.8 c: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 80% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.7$. 77

4.9 a: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 90% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.9$. 78

4.9 b: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 90% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.8$. 79

4.9 c: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 90% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.7$. 80

4.10 a: Empirical assurance probabilities at three assurance levels such that the lower bound of a two-sided 95% confidence interval for the AUC is not greater than the prespecified lower limit θ_0. Based on the AUC obtained from the CA 19-19 data by Wieand et al. 84

4.10 b: Empirical assurance probabilities at three assurance levels such that the lower bound of a two-sided 95% confidence interval for the AUC is not greater than the prespecified lower limit θ_0. Based on the AUC obtained from the CA125 data by Wieand et al. 85

Chapter 1 Introduction

1.1 The Receiver Operating Characteristic curve

Accurate diagnosis of a disease condition is crucial for successful treatment and management of the condition. Whether it may be for diagnosis or for screening, an accurate test may help patients with a disease get the treatment they need, and patients who do not have the disease will be free of unnecessary treatment. For a binary diagnostic test, there are two aspects of measures of accuracy: sensitivity and specificity. Sensitivity (or true positive rate) is the probability of correctly identifying an individual with the condition as positive, whereas specificity (or true negative rate) is the probability of correctly identifying an individual without the condition as negative (Yerushalmy, 1947). These two measures are negatively related. In other words, if a test were made to have very high sensitivity, it could potentially classify people free of the disease as positive because it would be much more conservative on classifying people as having the disease. Thus, it would be misleading to compare one measure without taking into consideration the other.

Additionally, there are situations that require tests with higher sensitivity and lower specificity or vise versa. For instance, a test for screening may require a higher specificity because screening is typically used on an asymptomatic population where only a small proportion of people have a condition. Since there is such a low probability of someone in this population having the condition, it may be more important that people without the disease are correctly classified as negative. On the other hand, for diagnosis, a test is usually applied to a symptomatic population where everyone may be at high risk for a particular disease. Since there is a high probability of having the disease amongst this population, it would be more important to have higher sensitivity

so that if someone does have the disease then it can be detected for certain. Screening and diagnosis may be done on different populations for different reasons, but they are evaluated using the same statistics.

True binary test results allow us to calculate sensitivity and specificity directly, but for test results that are continuous or ordinal, we must set various cut-off points in order to calculate the sensitivity and specificity. A tool that can visually display the discriminative ability of tests based on sensitivity and specificity through various cut-off points would be the Receiver Operating Characteristic (ROC) curve. The ROC curve is a useful tool because it plots sensitivity as a function of specificity based on all possible threshold values so the optimal trade-off between the two can be seen and adjusted. This allows us to evaluate the discriminative ability of tests as well as compare those of different tests.

The ROC graph goes from 0 to 1 on both axes, and the sensitivity (true positive rate) is on the y-axis, while '1 − specificity' (false positive rate) is on the x-axis. A test with ideal discrimination ability would have its ROC curve going from the origin to (0,1) to (1,1), forming two straight lines. On the other hand, a test with virtually no discrimination ability between typical (people without a condition of interest) and atypical (people with a condition of interest) populations would have a ROC curve going from the origin to (1,1) directly, as its true positive rate and false positive rate would be equal. Thus, the closer the ROC curve is to the top left corner of the graph, the better the test is at discriminating between two populations, and the more accurate it is. To plot an ROC curve, the sensitivity and specificity must be calculated for a series of different thresholds, and these sensitivity-specificity pairs are the coordinates for the point on the graph.

1.2 Area under the ROC curve

Moreover, test results can be continuous or ordinal. For example, high blood pressure is a condition that is not simply categorized as a positive and negative as it has many stages such as 'normal' and 'elevated', followed by different stages of hypertension. When a condition has many levels of severity or is measured on a continuous scale, we can no longer quantify test accuracy with estimates of sensitivity or specificity without having to dichotomize data with well-defined thresholds.

The area under the ROC curve (AUC) is useful in summarizing the accuracy of a test without needing to set various thresholds to dichotomize continuous or ordinal data. It quantifies the discriminative ability of tests and is commonly regarded as a global measure of accuracy. The AUC represents the probability that a randomly chosen value from one population would be greater than a randomly chosen value from another population, in other words, how well a test can discriminate between two different populations. The AUC of a test with ideal discrimination ability would be 1, as the area beneath this ROC curve would take up the entire plot. On the contrary, the ROC curve of a test without any discrimination ability would bisect the plot, making the area under this curve 0.5. As tests with better discriminative abilities are more favoured, the discriminative ability of a test can be interpreted on the basis of the AUC, using benchmarks such as those presented in Table 1.1 (El Khouli et al., 2010).

Table 1.1: Benchmarks to describe a test's discriminative ability based on different AUC values by El Khouli et al. (2010).

0.5 – 0.6	Failed
0.6 – 0.7	Poor

0.7 – 0.8	Fair
0.8 – 0.9	Good
0.9 – 1.0	Excellent

The area under the ROC curve has many applications in practice. Green and Swets (1966) found that there are two tasks equivalent to the AUC: the two-alternative forced choice task and the rating task. The two-alternative forced choice task is where two choices are presented, and an observer is forced to choose one of the choices. This is frequently used in signal detection theory and psychophysics, where a signal is presented in one of two observational intervals and noise is presented in both of the intervals. An observer must then determine which of the two intervals contains the signal (Green & Swets, 1966). The rating task has an observer rate a pair of randomly mixed stimuli on a scale, based on the strength of the stimulus. An example of this in the medical field would be when healthcare professionals evaluate a patient's condition based on the medical images of that patient. These images are often rated on a five-point scale from "definitely normal" to "definitely abnormal" (Hanley & McNeil, 1982). In both the two-alternative forced choice task and the rating task, the AUC represents the probability that a difference between two items is detected—in the forced choice task, the signal is recognized as different from noise, and in the rating task, an atypical patient's medical image is rated differently from a typical patient's image.

The area under the curve can be formally defined as the probability of correctly ranking a typical-atypical image pair. Let X and Y be the random variables denoting test values of typical and atypical subjects respectively, and θ denote the AUC, then

$$\theta = \Pr(X < Y)$$

where higher AUC values suggest subjects are likely to be atypical. Bamber (1975) observed that this probability $\Pr(X < Y)$ is the one-to-one function of the Mann-Whitney U statistic. This test

statistic is nonparametric and tests the null hypothesis that the cumulative distributions of two random variables are equal (Mann & Whitney, 1947). In other words, it tests whether the distributions of two populations are the same by determining whether a randomly chosen value from one population would not be greater than nor less than a randomly chosen value from another population. Accordingly, the null hypothesis of the Mann-Whitney test is $H_0: \Pr(X < Y) = 0.5$. Thus, the Mann-Whitney *U* statistic is the equivalent of the two-alternative forced choice task and the AUC, as all of them test whether two populations are different from one another.

We can express the area under the ROC curve as follows. When test results X and Y are assumed to be independent and normally distributed, with $X \sim N(\mu_X, {\sigma_X}^2)$ and $Y \sim N(\mu_Y, {\sigma_Y}^2)$, then

$$\begin{aligned}
\Pr(X < Y) &= \Pr(Y - X > 0) \\
&= \Pr\left(\frac{(Y - X) - (\mu_Y - \mu_X)}{\sqrt{{\sigma_Y}^2 + {\sigma_X}^2}} > \frac{-(\mu_Y - \mu_X)}{\sqrt{{\sigma_Y}^2 + {\sigma_X}^2}}\right) \\
&= \Pr\left(Z > \frac{-(\mu_Y - \mu_X)}{\sqrt{{\sigma_Y}^2 + {\sigma_X}^2}}\right) \\
&= \Pr\left(Z < \frac{\mu_Y - \mu_X}{\sqrt{{\sigma_Y}^2 + {\sigma_X}^2}}\right) \\
&= \Phi\left(\frac{\mu_Y - \mu_X}{\sqrt{{\sigma_Y}^2 + {\sigma_X}^2}}\right). \qquad (1.1)
\end{aligned}$$

The above definition can be extended to handle ties as

$$\theta = \Pr(X < Y) + 0.5 \Pr(X = Y).$$

1.3 Sample size estimation

Assessment of diagnostic tests or indices are important before applying them to practice. When evaluating new tests, it is vital to determine if a test's accuracy is adequate through a study regarding its AUC.

To design a research study evaluating test accuracy, three main factors need to be considered. First, the desired accuracy that is to be detected through the study needs to be determined. Next, determining a sample size is important. Larger sample sizes allow for the desired accuracy to be detected more easily, however, studies may have financial and time constraints so a large sample size may not always be plausible. The sample size also cannot be too small, as that may cause the desired accuracy to be difficult to detect. Finally, a level of power needs to be specified to ensure that the desired accuracy can be detected with a certain level of precision. Given any two of these factors, the third factor can easily be determined using a sample size formula.

To design a study focusing on estimation rather than hypothesis testing, the common approach is to base sample size on expected confidence interval width. A confidence interval is a range of values likely to contain the true value of an unknown parameter with a certain confidence. Confidence intervals can also display the result of a hypothesis test by including or excluding the null value within the interval, and the width of a confidence interval is determined by a degree of confidence and the variance of the estimate. Since the variance is affected by the sample size, it is clear that confidence intervals are related to the sample size of a study group.

The required sample size can actually be determined using the anticipated confidence interval width, along with a fixed percentage of confidence (Gordon, 1987). This is because sample size is closely related to confidence intervals, where greater sample sizes may be associated with smaller variances and narrower confidence intervals. However, if the expected width and

discriminatory power of the confidence interval are not simultaneously considered, then this method may underestimate the sample size (Greenland, 1988). Suppose a researcher would like to estimate the required sample size for a study to guarantee that if the true standard mortality ratio is 1, the expected upper confidence limit of a 95 percent confidence interval would be 1.2, and if the true standard mortality ratio is 0.7, there would be an expected upper confidence limit of 1 (Gordon, 1987). To distinguish between two values of a parameter such as in this example where one of the two values is correct, typically, there is only a 50 percent chance in achieving the desired confidence interval width as there is an equal probability that "the observed interval will exclude the incorrect one of two parameter values if the other of the two values is correct" (Greenland, 1988). In other words, if the true mortality ratio is 0.7, there would only be a 50 percent chance that the confidence interval excludes the incorrect value of 1, and if the true mortality ratio is 1, there is a 50 percent chance its confidence interval excludes 0.7. In order to increase this chance, we need to consider the probability of discriminating between these two values. This is actually equivalent to hypothesis testing: the power of a one-sided hypothesis test is equivalent to the probability that the confidence interval excludes the incorrect value. Therefore, we want to calculate a sample size such that there is a certain percent discriminatory power at a specific confidence level.

Thus, both interval width and discriminatory power need to be considered, and this is similar to the consideration of significance and test power in hypothesis testing. We define the assurance probability as the probability of achieving a prespecified criteria of a confidence interval, be it the lower bound or the width of it. This assurance probability is like the power in hypothesis testing and is also denoted by $1-\beta$. In order to ensure that sample sizes are not

underestimated, a prespecified assurance probability can be incorporated into sample size formulas.

1.4 Confidence intervals for AUC

The Wald confidence interval is the simplest and most commonly used confidence interval. It is symmetrical around the estimate and based on the variance of the estimate. For a two-sided confidence interval for the AUC, at the α significance level, we have:

$$\hat{\theta} \pm Z_{\alpha/2}\sqrt{var(\hat{\theta})} \tag{1.2}$$

where $\hat{\theta}$ is the estimator of the AUC.

For a one-sided confidence interval above or below the estimate at the α level, we have

$$\hat{\theta} + Z_{\alpha}\sqrt{var(\hat{\theta})}$$

$$\hat{\theta} - Z_{\alpha}\sqrt{var(\hat{\theta})}$$

where Z_{α} is the critical value corresponding to the desired α level.

There are many other types of confidence intervals, and they have different usages as well. For example, the Wilson's score confidence interval is commonly used for binomial proportions. In this thesis, we will use the Wald confidence interval when constructing confidence intervals for the AUC.

1.5 Variance of AUC

In order to construct confidence intervals for the AUC, we require the variance of its estimator. There are several different methods of estimating the variance and each method has its own

features which are explained in more depth in the Literature Review chapter. Some of the formulas may require actual data first, which would not be suitable for usage in planning studies. Thus, in this thesis, we focus on three main variance estimators that have components that can be easily determined during the planning stage.

The three variance estimators we consider are based on the exponential model by Hanley and McNeil (1982), the binormal model by Obuchowski (1994), and the probit model by Rosner and Glynn (2009). These methods are parametric, and we chose these variance estimators because they are functions of the anticipated value of the AUC and do not require the observed data, which would work well for planning studies. The variance estimator from the exponential model is based on the anticipated area under the curve and size ratio between typical and atypical groups and uses the same probabilities as in the Mann Whitney *U* test. The variance estimator from the binormal model is based on the anticipated area under the curve and the standard deviation ratio between the typical and atypical groups. The variance estimator from the probit model uses the probit transformation and the assumption that shifting the distribution of the typical group by a certain value can allow us to obtain the distribution of the atypical group.

1.6 Objective of thesis

When evaluating new tests, studies involving the AUC of the test must be conducted to determine whether the accuracy of the test is adequate. In order to plan these studies, a sufficient sample size is required. However, conventional sample sizes are determined using confidence interval based methods which tend to underestimate the required sample size as they do not consider the probability of achieving a prespecified criteria of a confidence interval, whether it is the lower bound or width. A way of ensuring that sample sizes are not underestimated would be to consider this probability, which we call the assurance probability.

In this thesis, we want to determine new sample size formulas that incorporate this assurance probability so that sample sizes are not underestimated. Then we proceed to evaluate the performance of them by comparing the empirical assurance probability to a prespecified assurance probability. We start by deriving asymptotic sample size formulas that incorporate a prespecified assurance probability of achieving a desired confidence interval lower bound or width. These sample size formulas are derived using three different variance estimators for AUC and then a simulation study is conducted to evaluate the performance of the proposed sample size formulas. The empirical assurance probability is defined as the frequency of the lower bound of a confidence interval around the estimated AUC being no lower than a preset lower limit, or the half-width of a confidence interval around the estimated AUC being no wider than a preset width. We then evaluate the performance of the proposed sample size formulas by comparing the empirical assurance probability to the prespecified assurance, where the closer the two values are, the better the performance. We follow up with a third method based on pilot data by Wieand et al. (1989), where the AUC and variances of the pilot dataset are first determined and then used in a new sample size formula to estimate sample size. The performance of this proposed method is also evaluated by comparing the empirical assurance probability to the prespecified assurance.

1.7 Organization of thesis

Chapter 2 reviews existing literature regarding the definitions of ROC and AUC, methods of estimating the variance of AUC, and confidence interval based methods of estimating sample size. Chapter 3 presents details about the three methods for estimating the variance of AUC, shows derivations of sample size formulas based on large sample theory, and then introduces the delta method and logit transformation. The simulation study evaluating the sample size formulas is

described in Chapter 4. Illustration of the sample size formula applied to existing data is shown in Chapter 5, and finally, the discussion concludes the thesis in Chapter 6.

Chapter 2 Literature Review

2.1 History of ROC curve

The receiver operating characteristic curve originated from World War II, where it was used as a way to display the accuracy of distinguishing signals from noise signals. It was believed that the ideal was to identify as many signals as possible, however it was soon realized that such a conservative approach in decision making also came with an increase in false positives. Thus, the ROC curve became a tool that displayed the false positive rate against the true positive rate, helping to find a trade-off between false positives and false negatives. Later the ROC curve became commonly used in psychophysics, where it displayed one's ability to distinguish various sensory stimuli such as physical and auditory stimuli. It is now commonly used to present the accuracy of diagnostic tools in the medical field.

In psychophysics, the ROC curve was used to express the result of the two-alternative forced choice test. This test was employed on participants and its purpose was to measure one's ability to detect whether a signal appeared or not when presented in the presence of noise. Green and Swets (1966) showed that the probability of correctly responding in this test is equivalent to the area under the ROC curve. It was shown that in these experiments, the stimulus and noise are unchanging while the only change that occurs is in the instructions, which may cause the observer's behaviour to change. The behaviour of the observer may be more or less conservative based on the goal that is instructed—if the observer is rewarded for all their correct 'yes' responses then they might increase their hit rate, but also increase their false positive rate. If the observer is rewarded for correctly identifying 'yes' and 'no' responses, then they may decrease their hits and

be more conservative when answering. These changes in behaviour correspond to various thresholds of true and false positive rates. When the resulting hits and misses are plotted, an ROC curve is formed (Green & Swets, 1966).

The rating task was also able to create ROC curves similarly to the two-alternative forced choice test. This task was used in medical imaging where images from patients with and without a condition would be presented to a rater who would then rate the images on a five-point scale from 'very normal' to 'very abnormal' (Hanley & McNeil, 1982). The images would be presented in typical-atypical pairs and the goal of this task was to correctly classify the two populations by rating the atypical image as more atypical than the typical image each time. The probability of correctly rating the images in this task was found to be equivalent to the area under the ROC curve as this probability is the same as the probability of correctly identifying a signal in the two-alternative forced choice task.

Green and Moses (1966) also conducted a recognition memory test as a variation of the rating task, where participants were given a list of material to memorize, and later given another list with a mix of old and new items. The participants were then asked to rate the items on the new list based on how confident they were about whether each item had been on the old list, on a six-point scale from +3 (very certain it was on the old list) to -3 (very certain it was not on the old list). Since the probability of correctly rating each item is fundamentally about being able to distinguish between the old and new items, Green and Moses (1966) verified that this probability is equivalent to the area under the ROC curve.

2.2 AUC and Mann-Whitney statistic

Bamber (1975) noted that the area under the ROC curve is similar to the area above the ordinal dominance (OD) graph. The OD graph plots the probability of random variable *X* being equal to or less than a constant *c* on the x-axis, and on the y-axis is the probability of random variable *Y* being equal to or less than *c* ($-\infty < c < +\infty$). Like the ROC curve, the OD graph ranges from 0 to 1 on both axes. If *X* and *Y* are continuous then the area $A(X,Y)$ above the OD graph can be defined as:

$$A(X,Y) = \int_0^1 P(X \leq c)\, dP\,(Y \leq c)$$

$$= \int_{-\infty}^{+\infty} P(X \leq c)\, f_Y(c)\, dc$$

$$= P(X \leq Y)$$

where f_Y is the probability density function of *Y*. The first integration from 0 to 1 is over the probability of $Y \leq c$, which is on the Y-axis of the OD graph. The second integration is over the constant c for all values of c ($-\infty < c < +\infty$), which is the OD curve that bounds the other side of this area.

It was shown that since the area above the OD graph is equivalent to the probability that *X* is less than or equal to *Y*, this area can be used as a measure of the probability of discriminating between two populations, such as a population with a condition and a population without a condition. Since the signal detection task is also used to discriminate between two items—a signal and a noise—the probabilities in these two tasks should be the same, and thus the area under the ROC curve should be equivalent to the area above the OD graph (Bamber, 1975).

Bamber then related the probability of correctly ranking the typical-atypical paired images to the Mann-Whitney test statistic. The Mann-Whitney test was proposed in 1947 based on the test

by Wilcoxon (1945). The method is commonly known as the Wilcoxon Mann-Whitney test today, as both the Wilcoxon test and the Mann-Whitney test are non-parametric for two group comparisons that test the null hypothesis that two distributions are the same (Mann & Whitney, 1947; Wilcoxon, 1945). However, the explicit methods for calculating the Wilcoxon statistic and the Mann-Whitney statistic are slightly different. Additionally, Wilcoxon also proposed a signed-rank test for correlated samples in the same paper (Wilcoxon, 1945).

The null hypothesis of the Wilcoxon-Mann Whitney test is

$$H_0: \Pr(X > Y) = \Pr(X < Y) = 0.5$$

where *X* and *Y* are two random variables that are independent and continuous. The Mann-Whitney *U* test then tests if the probability that a randomly chosen value from *X* will be greater than a randomly chosen value from *Y* is different from the probability that a randomly chosen value from X will be smaller than a randomly chosen value from Y (Mann & Whitney, 1947). The test statistic *U* is calculated using ranks, where all the values of *X* and *Y* must be arranged in order and then the number of times a *Y* value comes before an *X* value for each population is counted as

$$U_1 = mn + \frac{m(m+1)}{2} - R_1$$

$$U_2 = mn + \frac{n(n+1)}{2} - R_2 \qquad (2.1)$$

where U_1 and U_2 are the Mann Whitney *U* statistics for the atypical and typical populations respectively, *n* is the number of *X* samples, *m* is the number of *Y* samples, and R_1 and R_2 are the sums of the ranks of *X* and *Y* values, respectively. The smaller value of U_1 and U_2 is compared against a critical value corresponding to a certain level of significance and sample size. If this *U* is larger than the critical value the null hypothesis would be supported, and if *U* is smaller than the critical value then the null hypothesis would be rejected.

Since the Mann-Whitney test is for distinguishing between two populations, Bamber (1975) noticed that the image rating task must be closely related to it. The rating task is ultimately about being able to distinguish that one population is higher ranked (more likely to have a disease) than the other, and thus, the probability of distinguishing between two images should be the same as what the Mann-Whitney statistic measures (Bamber, 1975).

Hanley and McNeil (1982) expanded on this connection and added in the equivalence of a third factor. A connection was shown between the AUC representing the probability of correctly ranking a typical-atypical pair of images, and the Wilcoxon statistic also measuring this probability. This was clear as the explicit method of calculating the Wilcoxon statistic already makes the same comparisons as the experimental tasks. The Wilcoxon statistic is formed by comparing each possible combination of *X*-*Y* pairs and scoring them based on how their values compare:

$$S(X,Y) = \begin{cases} 1 & if\ X > Y \\ 1/2 & if\ X = Y \\ 0 & if\ X < Y \end{cases}$$

where *S* is the score. Then these values are averaged to receive the Wilcoxon statistic *W*

$$W = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} S\left(X_i, Y_j\right) \tag{2.2}$$

where *n* and *m* are the number of people in the atypical and typical populations respectively, and X_i and Y_j are the test scores of the atypical and typical populations, respectively. *W* is the proportion of how frequent the values from *X* are greater than values from *Y,* and is ranged between 0 and 1. Thus, Hanley and McNeil (1982) showed that the Wilcoxon statistic is equivalent to both the area under the curve and the image rating task, giving us the AUC formula that is commonly used today

$$\hat{\theta} = \frac{U_1}{nm}.$$

2.3 Variance of AUC

There are several methods to calculate the variance of an AUC estimator. Bamber (1975) not only related the AUC to the OD graph and to the Mann-Whitney statistic but also derived a variance formula for the AUC. This variance formula is based on the formula by Noether (1967), which does not assume that patient scores X and Y are continuous. Two elements of the formula, B_{XXY} and B_{YYX} must be estimated: B_{YYX} is the probability that two randomly chosen values from the Y distribution would be both greater than or less than a randomly chosen value from the X distribution, minus the probability that the value from X is in between the two values from Y. B_{XXY} is defined similarly but with two values from the X distribution, and one from the Y distribution.

In addition to the discovery of the three-way equivalence, Hanley and McNeil (1982) illustrated that the variance of the AUC would be an important factor in making comparisons of independent AUCs. Thus, a variance estimate of the AUC was developed based on the negative exponential distribution.

Hanley and McNeil (1983) also extended the previous result to include a different situation: comparing multiple AUCs generated from the same patient. Since these AUCs would be correlated, the standard error of the difference between two AUCs cannot be calculated by a summation of their individual standard errors. A formula for the correlation between AUCs was developed based on the average correlation between ratings of typical and atypical groups and the average AUCs. The standard error of the difference between correlated AUCs was then derived using this correlation, and its performance was evaluated through an experiment using phantoms. This formula performed well when the data was continuous and Gaussian distributed.

DeLong et al. (1988) developed a nonparametric method for comparing the areas under correlated ROC curves. This nonparametric method incorporated the method of structural

components by Sen (1960) to estimate a covariance matrix of *U*-statistics and used this to develop a statistical test that is asymptotically chi-squared distributed. The validity of the nonparametric method rests on the asymptotic normality established by Hoeffding (1948) and the consistency of variance-covariance estimators proved by Sen (1960). This nonparametric method has become widely used for comparing ROC curves and is implemented in common software such as SAS.

Obuchowski (1994) evaluated the variance estimator based on the exponential distribution by Hanley and McNeil (1982) and investigated how this variance may be affected by extreme standard deviation ratios between typical and atypical patient groups. The exponential model was tested for standard deviation ratios more extreme than 1:0.71, as well as different numbers of rating categories. The results revealed that the variance estimated from the exponential model was not conservative when the standard deviation ratio is very small.

Obuchowski (1994) proposed another method for calculating the variance of AUC that yielded more conservative results than the exponential model. The proposed method used a two-parameter binormal distribution that considers both the anticipated area under the curve as well as the standard deviation ratio between study groups. The results of this method tended to be close to the empirical variances but were conservative when the standard deviation ratio is 1, meaning that the standard deviations of typical and atypical patient groups are the same. As the standard deviation ratio decreases (more variability in the atypical group compared to the typical group), the variance obtained from the binormal model would underestimate the empirical variance.

Rosner and Glynn (2009) derived a new variance formula for the AUC using the probit transformation. First the Wilcoxon-Mann Whitney *U* test was compared to the shift model alternative hypothesis described by Lehmann (1956), which defines the null and alternative hypotheses of comparing two tests as:

$$H_0: F_X = F_Y$$

$$H_1: F_Y(y) = F_X(y - \Delta)$$

for all x, where *X* and *Y* are scores of patients from continuous distributions F_X and F_Y. This implies that the distribution F_X can be obtained by altering F_Y by the amount Δ (Lehmann, 1956). Rosner and Glynn (2009) evaluated the asymptotic relative efficiency (ARE) of both of those methods and then developed a different type of shift alternative for estimation of the Wilcoxon-Mann Whitney *U* test when sample sizes are small. This new shift alternative uses the probit transformation:

$$H_0: H_X = H_Y = H_{Y_c}$$

$$H_1: H_Y = H_{Y_c} + \mu$$

where Y_c is a counterfactual variable where a study group receives a control treatment, $H_X = \Phi^{-1}(F_X)$ where H_X is the probit corresponding to *X* and Φ is the cumulative distribution function (cdf) of a standard normal distribution. An explicit expression for the power of the Wilcoxon-Mann Whitney *U* test under this shift alternative was developed, as well as its variance estimate.

2.4 Sample size estimation based on confidence intervals

Traditionally, the required sample size can be estimated based on confidence intervals when planning a study. However, Greenland (1988) pointed out that using confidence intervals would often lead to an underestimation of the sample size. This is because the probability of achieving the desired confidence interval is not considered. Suppose one intends to construct a confidence interval with an upper bound below a prespecified upper limit. This confidence interval would be centered around an expected value, but its width is not considered which means there is only a 50 percent chance of the interval's upper bound excluding that prespecified upper limit. If a study

incorporated an 80 percent assurance probability, the constructed confidence interval would have an 80 percent chance at excluding the prespecified value. A sample size would be considered adequate if its observed confidence interval excludes the prespecified value (Greenland, 1988). Kupper and Hafner (1989) also evaluated popular sample size formulas based on large sample theory and found that these formulas would often underestimate the sample size in small sample situations.

Although confidence interval based methods tend to underestimate the required sample size, Daly (1991) showed that confidence interval based methods could still be used if they are interpreted differently. Since there is a relationship between significant tests and confidence intervals, the power of a significance test should be translated to the probability of achieving a desired confidence interval given that it excludes the null value. This would ensure that sample sizes are not underestimated.

Zou (2012) improved the idea of considering confidence interval widths when calculating sample size estimations by deriving sample size formulas that directly incorporated a prespecified probability of achieving the desired confidence interval width or lower bound. These formulas were utilized to attain a large enough sample size for studies involving the intraclass correlation coefficient (ICC). The prespecified assurance probability was paired with a prespecified confidence interval width or lower bound, and simulation studies were conducted to evaluate the performance of the sample size formulas.

Chapter 3 Methods

3.1 Introduction

When planning studies for estimating the area under the characteristic operating curve (AUC), using traditional confidence interval based methods to estimate a required sample size often leads to inadequate sample sizes. This is because the methods only consider the expected interval width without incorporating the probability of actually achieving that width (Daly, 1991). Thus, there is only a 50 percent chance of the expected confidence interval excluding a prespecified value that is to be excluded (Greenland, 1988). The underestimated sample size may lead to studies being unable to reach study objectives. Therefore, in order to plan studies with sufficient sample sizes, the probability of achieving the desired confidence interval (assurance probability) must be considered.

In this section, we incorporate the assurance probability into sample size formulas to address two research questions—what is the required sample size if we want a confidence interval's lower bound to be above a certain preset lower limit? And what is the required sample size if we want a confidence interval's width to be narrower than a certain preset width? We also introduce the three variance estimators which are needed to form confidence intervals around the AUC.

3.2 Existing variance formulas

Several variance formulas for the AUC have been proposed in the literature for the purpose of sample size estimation. However, these methods are mostly concerning hypothesis testing and comparisons of multiple AUCs. Hanley and McNeil (1982) proposed a variance estimator based on the exponential distribution and calculated required sample sizes for detecting differences between various pairs of AUCs. Based on the bivariate normal distribution, Obuchowski (1994) derived a variance estimator that is more conservative than the exponential based variance. Lastly, the variance estimator based on the probit transformation by Rosner and Glynn (2009) gives an explicit expression for the shift alternative variance of AUC.

3.2.1 The variance estimator based on the exponential distribution

After the discovery of three-way equivalence between the forced choice task, rating experiments, and the Wilcoxon Mann-Whitney *U* test by Hanley and McNeil (1982), the importance of the area under the curve became apparent and a formula for calculating its variance was developed. The method by Hanley and McNeil (1982) assumes test data follow a negative exponential distribution, and this method was shown to provide more conservative variance estimates than the method based on the Gaussian distribution. This formula was purely based on the anticipated area under the curve θ, and is shown as:

$$\mathrm{var}\left(\hat{\theta}\right) = \frac{\theta(1-\theta) + (n-1)(Q_1 - \theta^2) + (m-1)(Q_2 - \theta^2)}{mn} \tag{3.1}$$

where

$$Q_1 = \frac{\theta}{2-\theta}$$

$$Q_2 = \frac{2\theta^2}{1+\theta}.$$

Q_1 and Q_2 are similar to the terms used in the variance formula by Bamber (1975): Q_1 is the probability that two randomly chosen values from the atypical population will be ranked higher than a random chosen value from the typical population, $P(X_i < Y_{j1} \text{ and } X_i < Y_{j2})$. Q_2 is the probability that one randomly chosen value from the atypical population will be ranked higher than two randomly chosen values from the typical population, $P(X_{i1} < Y_j \text{ and } X_{i2} < Y_j)$.

As we can see, this variance formula only requires the anticipated area under the curve θ, the number of patients with the condition *n*, and the number of patients without the condition *m*. It is not based on observed data which makes this method useful when planning studies or when there is not much information about the data and only an anticipated area under the curve is known.

We simplified all the variance formulas first in order to derive the sample size formula in the next section, as the derivation requires the variance formula to have an *N* term factored out. We started by expanding the original formula

$$\text{var}(\hat{\theta}) = \frac{1}{nm}[\theta(1-\theta) + (n-1)(Q_1 - \theta^2) + (m-1)(Q_2 - \theta^2)]$$

$$= \frac{1}{nm}[\theta(1-\theta) + n(Q_1 - \theta^2) + m(Q_2 - \theta^2) - (Q_1 + Q_2 - 2\theta^2)]\,.$$

Then we make use of the relationships between total sample size *N* and the individual group sizes *n* and *m*, as well as the relationship between group sizes $m = rn$ to give us

$$n = \frac{1}{1+r}N$$

$$m = \frac{r}{1+r}N\,.$$

Substituting these into the variance formula gives us

$$\text{var}(\hat{\theta}) = \frac{(1+r)}{rN}\left[\frac{Q_1}{r} + Q_2 - \theta^2\left(\frac{1}{r} + 1\right)\right].$$

Substituting in Q_1 and Q_2 and cancelling out the *N* term gives us the variance component that we will use in the next section. Let us call this $f(\theta)$

$$f(\theta) = (1 + \mathrm{r})\left(\frac{\theta}{r(2-\theta)} + \frac{2\theta^2}{1+\theta} - \theta^2\left(\frac{1}{r} + 1\right)\right).$$

3.2.2 The variance estimator based on the bivariate normal distribution

Obuchowski (1994) wanted to improve on the exponential model of variance as it had not been evaluated on typical to atypical group standard deviation ratios (*B*) more extreme than 1:0.71. A new estimate of variance was developed based on normal data where data of typical subjects are normally distributed and have mean μ_m and variance ${\sigma_m}^2$, and data of atypical subjects follow a normal distribution with mean μ_n and variance ${\sigma_n}^2$. Using the fact that

$$\theta = \Phi\left(\frac{\mu_n - \mu_m}{\sqrt{\sigma_n^2 + \sigma_m^2}}\right)$$

Obuchowski applied the delta method to derive the variance for $\hat{\theta}$. This binormal based variance estimate was developed on the basis of considering the standard deviation ratio between study groups *B*, as well as the anticipated area under the curve.

The formula for calculating the variance estimate based on the binormal model is

$$\mathrm{var}\left(\hat{\theta}\right) = (2\pi)^{-1}\, e^{-\frac{A^2}{W}}\left(\frac{V_1}{W} + (AB)^2\frac{V_2}{W^3}\right) \tag{3.2}$$

where

$$A = \sqrt{(1 + B^2)}\,\Phi^{-1}(\theta)$$

$$B = \frac{\sigma_m}{\sigma_n}$$

$$W = 1 + B^2$$

$$V_1 = \frac{1}{n} + \frac{B^2}{m} + \frac{A^2}{2n}$$

$$V_2 = \frac{B^2}{2}\left(\frac{1}{n} + \frac{1}{m}\right).$$

The binormal based variance formula was also simplified similarly, substituting in

$$n = \frac{1}{1+r}N$$

$$m = \frac{r}{1+r}N$$

and factoring out and cancelling the *N* term to get the variance component that will be used for sample size calculation in the next section. Again, let us call this $f(\theta)$

$$f(\theta) = (1+r)\frac{1}{2\pi}e^{-\frac{A^2}{W}}\left[\frac{V_1}{W} + (AB)^2\left(\frac{V_2}{W^3}\right)\right]$$

where the individual terms are the same except

$$V_1 = 1 + \frac{B^2}{r} + \frac{A^2}{2}$$

$$V_2 = \frac{B^2}{2}\left(1 + \frac{1}{r}\right).$$

3.2.3 The variance estimator based on the probit transformation

Rosner and Glynn (2009) developed a method for calculating the variance of the AUC which was based on Lehmann's (1975) shift alternative but uses the probit transformation. This is a location shift of distributions for typical and atypical subjects, which is rank preserving. By using the property that the transformed values are normal, Rosner and Glynn showed that

$$P\left(X_i < Y_{j1} \text{ and } X_i < Y_{j2}\right)$$

$$= P\left(X_{i1} < Y_j \text{ and } X_{i2} < Y_j\right)$$

$$= \Phi_2\left\{\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{1}{2}\right\}$$

where Φ is the cdf of a standard normal distribution, and Φ_2 is a two-dimensional normal cdf given by:

$$\Phi_2(c_1, c_2, \rho) = \Pr\left(Z_1 \leq c_1 \text{ and } Z_2 \leq c_2 | (Z_1, Z_2) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)\right)$$

and where $P\left(X_i < Y_{j1} \text{ and } X_i < Y_{j2}\right)$ is the probability that two randomly chosen values from the atypical population will be ranked higher than a random chosen value from the typical population, and $P\left(X_{i1} < Y_j \text{ and } X_{i2} < Y_j\right)$ is the probability that one randomly chosen value from the atypical population will be ranked higher than two randomly chosen values from the typical population.

The variance formula based on the probit transformation was derived by

$$\text{var}\left(\hat{\theta}\right) = \frac{\theta(1-\theta) + (m+n-2)\left[\Phi_2\left\{\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{1}{2}\right\} - \theta^2\right]}{mn}. \qquad (3.3)$$

We also simplified this variance formula by substituting in

$$n = \frac{1}{1+r}N$$

$$m = \frac{r}{1+r}N$$

to give us

$$\text{var}\left(\hat{\theta}\right) = \frac{1}{\left(\frac{rN}{1+r}\right)\left(\frac{N}{1+r}\right)}\left(\theta(1+\theta) + \left(\frac{rN}{1+r} + \frac{N}{1+r} - 2\right)\left[\Phi_2\left\{\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{1}{2}\right\}\right]\right).$$

Then after factoring out and cancelling the *N* term, we get the final form $f(\theta)$:

$$f(\theta) = \frac{(r+1)^2}{r}\left[\Phi_2\left\{\Phi^{-1}(\theta), \Phi^{-1}(\theta), \frac{1}{2}\right\} - \theta^2\right].$$

3.3 Delta method for logit transformation

In this thesis, we used the Wald confidence interval. Since the calculation of the Wald interval is formed using the AUC estimate and the variance of the AUC estimate, it is possible to get a range that lands outside the domain of our variable. In this case, we did not want a confidence interval for the AUC to be less than 0 or greater than 1 because it is not possible for the AUC to be outside of that domain, as it is a probability measure. Hence, we used the logit transformation to construct confidence intervals as the sampling distribution for the logit transformed estimate approaches normality faster than on the raw scale. Then it can be transformed back onto the probability scale. Since the logit transformation is monotonic, probabilities are preserved. By doing this, we can ensure that the confidence interval we get is not outside the desired range of (0,1).

To approximate the asymptotic distributions of the logit transformed variables, the use of the delta method is required. The delta method is a way of finding asymptotic variances based on the Taylor series and is often used to approximate means and variances.

Suppose g is a function that has a derivative g'. Then for random variable X with mean μ, we can approximate the function $g(X)$ at μ using the first order Taylor series expansion:

$$g(X) = g(\mu) + g'(\mu)(X - \mu)$$

then $g(X)$ asymptotically follows the normal distribution with the mean and variance

$$E[g(X)] = g(\mu)$$

$$\text{var}[g(X)] = \left(g'(\mu)\right)^2 \text{var}(X)\,.$$

The delta method was used to apply the logit transformation to the variances calculated using each of the three methods. Using the variance of the AUC estimate, $\text{var}(\hat{\theta})$, calculated using each respective method, the variance of the logit transformed AUC can be obtained with:

$$g(\theta) = \ln\frac{\theta}{1-\theta}$$

$$g'(\theta) = -\frac{1}{\theta(1-\theta)}$$

$$\widehat{\text{var}}\left(\ln\frac{\hat{\theta}}{1-\hat{\theta}}\right) = \frac{\widehat{\text{var}}(\hat{\theta})}{\hat{\theta}^2(1-\hat{\theta})^2}.$$

Based on the above, a typical logit transformed Wald confidence interval would be constructed as such

$$\left(\ln\frac{\hat{\theta}}{1-\hat{\theta}} - Z_{\alpha/2}\sqrt{\frac{\widehat{\text{var}}(\hat{\theta})}{\hat{\theta}^2\left(1-\hat{\theta}\right)^2}}, \qquad \ln\frac{\hat{\theta}}{1-\hat{\theta}} + Z_{\alpha/2}\sqrt{\frac{\widehat{\text{var}}(\hat{\theta})}{\hat{\theta}^2\left(1-\hat{\theta}\right)^2}}\right).$$

Then substituting in the variance forms that are used in this thesis, we get the confidence interval

$$\left(\ln\frac{\hat{\theta}}{1-\hat{\theta}} - Z_{\alpha/2}\frac{\sqrt{f(\hat{\theta})}}{\hat{\theta}(1-\hat{\theta})\sqrt{N}}, \qquad \ln\frac{\hat{\theta}}{1-\hat{\theta}} + Z_{\alpha/2}\frac{\sqrt{f(\hat{\theta})}}{\hat{\theta}(1-\hat{\theta})\sqrt{N}}\right)$$

where the variance of the AUC estimator is estimated by

$$\frac{f(\hat{\theta})}{N}.$$

To transform the interval back onto the raw scale, the back logit transformation

$$\frac{\exp(x)}{1+\exp(x)}$$

is applied to the lower and upper bounds. After performing this transformation, the confidence interval for the AUC is obtained as

$$\left(\frac{\exp\left(\ln\frac{\hat{\theta}}{1-\hat{\theta}} - Z_{\alpha/2}\frac{\sqrt{f(\hat{\theta})}}{\hat{\theta}(1-\hat{\theta})\sqrt{N}} \right)}{1+\exp\left(\ln\frac{\hat{\theta}}{1-\hat{\theta}} - Z_{\alpha/2}\frac{\sqrt{f(\hat{\theta})}}{\hat{\theta}(1-\hat{\theta})\sqrt{N}} \right)}, \quad \frac{\exp\left(\ln\frac{\hat{\theta}}{1-\hat{\theta}} + Z_{\alpha/2}\frac{\sqrt{f(\hat{\theta})}}{\hat{\theta}(1-\hat{\theta})\sqrt{N}} \right)}{1+\exp\left(\ln\frac{\hat{\theta}}{1-\hat{\theta}} + Z_{\alpha/2}\frac{\sqrt{f(\hat{\theta})}}{\hat{\theta}(1-\hat{\theta})\sqrt{N}} \right)} \right).$$

The transformed variances and AUCs are used in the simulation study in the Chapter 4.

The delta method was also used in deriving the sample size formulas in the next section.

3.4 Sample size estimation

When planning studies, we need to determine the sample size, the desired effect to be detected, and the amount of confidence that the effect will be detected. In order to do this, we will need a sample size formula that ties these three factors together. On top of those, we incorporate a prespecified assurance probability of achieving a prespecified lower bound and a prespecified confidence interval half-width to derive two formulas that would allow us to examine the assurance probability in those two conditions in the subsequent sections. Additionally, we perform a logit transformation on the AUC estimate and its variance so that the confidence interval around the AUC does not exceed the possible range of (0,1).

3.4.1 Sample size formula with a prespecified lower limit

To derive a sample size formula with a prespecified lower limit based on the logit transformation, which we denote by lgt θ, the logit transformation is

$$\text{lgt}\,\theta = \ln\frac{\theta}{1-\theta}$$

we first start with the assurance probability

$$1-\beta = \Pr\left(\hat{\theta}_L \geq \theta_0\right)$$

$$= \Pr\left(\operatorname{lgt}\hat{\theta} - Z_{\alpha/2}\sqrt{\operatorname{var}\left(\operatorname{lgt}(\hat{\theta})\right)} \geq \operatorname{lgt}\theta_0\right)$$

$$= \Pr\left(\ln\frac{\hat{\theta}}{1-\hat{\theta}} - Z_{\alpha/2}\sqrt{\operatorname{var}\left(\ln\frac{\hat{\theta}}{1-\hat{\theta}}\right)} \geq \ln\frac{\theta_0}{1-\theta_0}\right)$$

where $\hat{\theta}_L$ is the lower bound of the two-sided confidence interval for the AUC, θ_0 is the prespecified lower bound, and $1-\beta$ is the prespecified assurance probability. Note that the logit transformation is strictly monotonic, so the probabilities are preserved.

After cancelling out the *N* terms from the three variance estimators in Section 3.2, we can substitute those into our sample size formula derivation. For the variance estimator based on the exponential distribution, we have the simplified form $f(\theta)$:

$$f(\theta) = (1+\mathrm{r})\left(\frac{\theta}{r(2-\theta)} + \frac{2\theta^2}{1+\theta} - \theta^2\left(\frac{1}{r}+1\right)\right).$$

If we substitute in $f(\theta)$ for the variance in our derivation, we get

$$1-\beta = \Pr\left(\operatorname{lgt}\hat{\theta} - Z_{\alpha/2}\frac{\sqrt{\frac{f(\theta)}{N}}}{\theta(1-\theta)} \geq \operatorname{lgt}\theta_0\right)$$

$$= \Pr\left(\operatorname{lgt}\hat{\theta} \geq \operatorname{lgt}\theta_0 + Z_{\alpha/2}\frac{\sqrt{f(\theta)}}{\theta(1-\theta)\sqrt{N}}\right).$$

Using the delta method, we can get the mean and variance of $\operatorname{lgt}\hat{\theta}$

$$\mathrm{E}\left[\operatorname{lgt}\hat{\theta}\right] = \operatorname{lgt}\theta$$

$$\operatorname{var}\left(\operatorname{lgt}\hat{\theta}\right) \approx \left(\operatorname{lgt}\theta\right)'^2 \operatorname{var}\left(\hat{\theta}\right)$$

$$= \left(\ln\frac{\theta}{1-\theta}\right)'^2 \frac{f(\theta)}{N}$$

$$= \frac{f(\theta)}{\theta^2(1-\theta)^2 N}.$$

Then, by the central limit theorem,

$$\operatorname{lgt} \hat{\theta} \sim N\left(\operatorname{lgt} \theta, \frac{f(\theta)}{\theta^2(1-\theta)^2 N}\right).$$

We can substitute this information into the equation to standardize it

$$1-\beta = \Pr\left(\frac{\operatorname{lgt} \hat{\theta} - \operatorname{lgt} \theta}{\frac{\sqrt{f(\theta)}}{\theta(1-\theta)\sqrt{N}}} \geq \frac{\operatorname{lgt} \theta_0 - \operatorname{lgt} \theta + Z_{\alpha/2} \frac{\sqrt{f(\theta)}}{\theta(1-\theta)\sqrt{N}}}{\frac{\sqrt{f(\theta)}}{\theta(1-\theta)\sqrt{N}}}\right).$$

This gives us the expression for assurance probability:

$$Z_\beta = \frac{-\operatorname{lgt} \theta_0 + \operatorname{lgt} \theta - Z_{\alpha/2} \frac{\sqrt{f(\theta)}}{\theta(1-\theta)\sqrt{N}}}{\frac{\sqrt{f(\theta)}}{\theta(1-\theta)\sqrt{N}}}$$

where Z_β is the upper β quantile of the standard normal distribution. After isolating *N*, we get the sample size formula

$$N = \left(\frac{Z_\beta + Z_{\alpha/2}}{\operatorname{lgt} \theta - \operatorname{lgt} \theta_0}\right)^2 \frac{f(\theta)}{\theta^2(1-\theta)^2}. \quad (3.4)$$

3.4.2 Sample size formula with prespecified interval width

To derive a sample size formula with a prespecified confidence interval half-width ω based on the logit transformation, we must first find the corresponding half-width for the confidence interval on the logit scale, denoted ω^*. We start by letting the prespecified width (2ω) be equal to the confidence interval width calculated from the back transformation of the confidence interval in the logit scale. As we construct the confidence interval in the logit scale as

$$(\text{lgt}\,\hat{\theta} - \omega^*, \text{lgt}\,\hat{\theta} + \omega^*)$$

with the inverse logit transformation

$$\frac{\exp(x)}{1 + \exp(x)}$$

then 2ω, the total width of the confidence interval for θ in the original scale is expressed as

$$2\omega = \frac{\frac{\theta \exp(\omega^*)}{1-\theta}}{\frac{\theta \exp(\omega^*)}{1-\theta} + 1} - \frac{\frac{\theta}{(1-\theta)\exp(\omega^*)}}{\frac{\theta}{(1-\theta)\exp(\omega^*)} + 1}.$$

Solving for ω^* gives us

$$\omega^* = \ln\left(\frac{2\omega(1-\theta)^2 + \theta^2}{(1-2\omega)\theta(1-\theta)} + \sqrt{\left(\frac{2\omega(1-\theta)^2 + \theta^2}{(1-2\omega)\theta(1-\theta)}\right)^2 - \frac{4(2\omega+1)}{2\omega-1}}\right) - \ln 2\,.$$

We then express the assurance probability in terms of the confidence interval half-width in the logit scale with the specified half-width in the logit scale, ω^* as follows:

$$1 - \beta = \Pr\left(Z_{\alpha/2}\sqrt{\widehat{\text{var}}(\text{lgt}\,\hat{\theta})} \leq \omega^*\right)$$

and replacing the variance with $f(\theta)$, we get

$$1 - \beta = \Pr\left(Z_{\alpha/2}\frac{\sqrt{f(\hat{\theta})}}{\hat{\theta}(1-\hat{\theta})\sqrt{N}} \leq \omega^*\right)$$

$$= \Pr\left(\frac{\sqrt{f(\hat{\theta})}}{\hat{\theta}(1-\hat{\theta})} \leq \frac{\omega^*\sqrt{N}}{Z_{\alpha/2}}\right).$$

Then we use the delta method to find the asymptotic distribution of

$$\frac{\sqrt{f(\hat{\theta})}}{\hat{\theta}(1-\hat{\theta})}.$$

By the delta method,

$$\mathrm{E}\left[\frac{\sqrt{f(\hat{\theta})}}{\hat{\theta}(1-\hat{\theta})}\right] = \frac{\sqrt{f(\theta)}}{\theta(1-\theta)}$$

$$\mathrm{var}\left(\frac{\sqrt{f(\hat{\theta})}}{\hat{\theta}(1-\hat{\theta})}\right) \approx \left(\frac{\sqrt{f(\theta)}}{\theta(1-\theta)}\right)'^{2} \mathrm{var}(\hat{\theta})$$

$$= \left(\frac{f'(\theta)}{2\sqrt{f(\theta)}\theta(1-\theta)} - \frac{\sqrt{f(\theta)}(1-2\theta)}{\theta^2(1-\theta)^2}\right)\frac{\sqrt{f(\theta)}}{\sqrt{N}}$$

$$= \frac{1}{2\sqrt{N}}\left(\frac{f'(\theta)}{\theta(1-\theta)} - \frac{2f(\theta)(1-2\theta)}{\theta^2(1-\theta)^2}\right)$$

$$= \frac{f^*}{2\sqrt{N}}$$

where

$$f^* = \frac{f'(\theta)}{\theta(1-\theta)} - \frac{2f(\theta)(1-2\theta)}{\theta^2(1-\theta)^2}.$$

Thus, by the central limit theorem,

$$\frac{\sqrt{f(\hat{\theta})}}{\hat{\theta}(1-\hat{\theta})} \sim \mathrm{N}\left(\frac{\sqrt{f(\theta)}}{\theta(1-\theta)}, \left(\frac{f^*}{2\sqrt{N}}\right)^2\right).$$

Substituting this in, we can continue with the derivation:

$$1-\beta = \Pr\left(\frac{\frac{\sqrt{f(\hat{\theta})}}{\hat{\theta}(1-\hat{\theta})} - \frac{\sqrt{f(\theta)}}{\theta(1-\theta)}}{\frac{f^*}{2\sqrt{N}}} \leq \frac{\frac{\omega^*\sqrt{N}}{Z_{\alpha/2}} - \frac{\sqrt{f(\theta)}}{\theta(1-\theta)}}{\frac{f^*}{2\sqrt{N}}}\right).$$

Thus we get for the assurance level formula

$$Z_\beta = \frac{\frac{\omega^*\sqrt{N}}{Z_{\alpha/2}} - \frac{\sqrt{f(\theta)}}{\theta(1-\theta)}}{\frac{f^*}{2\sqrt{N}}}.$$

And after solving for *N*, we get the sample size formula

$$N = \left(\frac{\frac{\sqrt{f(\theta)}}{\theta(1-\theta)} + \sqrt{\frac{f(\theta)}{\theta^2(1-\theta)^2} + \frac{2\omega^* Z_\beta f^*}{Z_{\alpha/2}}}}{\frac{2\omega^*}{Z_{\alpha/2}}} \right)^2. \tag{3.5}$$

The sample size formula (3.5) can be used with each of the three variance estimators, with $f(\theta)$ being the respective variance component without the *N* term, and f^* being the expression that requires the first derivative of $f(\theta)$ with respect to θ, denoted by $f'(\theta)$. In what follows, we present the first derivative of the variance component derived for each variance estimator.

The first derivative of the variance estimator based on the exponential model is

$$f'(\theta) = (1+r)\left(\frac{Q_1'}{r} + Q_2' - 2\theta\left(\frac{1}{r}+1\right)\right)$$

where Q_1' and Q_2' represent the first derivative with respect to θ of Q_1 and Q_2, respectively, as follows

$$Q_1' = \frac{(1+Q_1)}{2-\theta} = \frac{2}{(2-\theta)^2}$$

$$Q_2' = \frac{4\theta - Q_2}{1+\theta} = \frac{4\theta + 2\theta^2}{(1+\theta)^2}.$$

The first derivative of the variance estimator based on the binormal model is

$$f'(\theta) = \left(\frac{1+r}{2\pi}\right)\frac{-2AA'}{W}\exp\left(-\frac{A^2}{W}\right)\left(\frac{V_1}{W} + (AB)^2\frac{V_2}{W^3}\right)$$

$$+ \frac{1+r}{2\pi}\exp\left(-\frac{A^2}{W}\right)\left(\frac{V_1{}'}{W} + 2\,A'B^2A\,\frac{V_2}{W^3}\right)$$

where A' and V_1' represent the first derivatives with respect to θ of A and V_1, respectively, and are expressed with the standard normal density function $\phi(x)$ as follows

$$A' = \frac{\sqrt{1+B^2}}{\phi\left(\Phi^{-1}(\theta)\right)}$$

$$V_1{}' = A'A\,.$$

Lastly, the first derivative of the variance estimator based on the probit model is derived as

$$f'(\theta) = \frac{(r+1)^2}{r}\left(\frac{\partial}{\partial\theta}\Phi_2(\Phi^{-1}(\theta), \Phi^{-1}(\theta), 0.5) - 2\theta\right).$$

Letting $x = y = \Phi^{-1}(\theta)$ and using the chain rule, we can obtain

$$f'(\theta) = \frac{(r+1)^2}{r}\left(\frac{\partial\Phi_2(x,\ y,\ 0.5)}{\partial x}\frac{\partial x}{\partial\theta} + \frac{\partial\Phi_2(x,\ y,\ 0.5)}{\partial y}\frac{\partial y}{\partial\theta} - 2\theta\right)$$

$$= \frac{(r+1)^2}{r}\left(2\frac{\partial\Phi_2(x,\ y,\ 0.5)}{\partial x}\frac{\partial x}{\partial\theta} - 2\theta\right)$$

$= \frac{(r+1)^2}{r}(2F(\Phi^{-1}(\theta)|\Phi^{-1}(\theta)) - 2\theta)$ .Here the first derivative of the bivariate normal distribution function is obtained as the product of the conditional distribution and the standard normal density function as follows

$$\frac{\partial\Phi_2(x,y,0.5)}{\partial x} = F(y|x;0.5)\ \phi(x)$$

where $F(y|x;0.5)$ is the conditional distribution function of Y given $X = x$ with $Y|X = x \sim N(\rho x, 1-\rho^2)$ when X and Y follow a bivariate normal distribution with the distribution function $\Phi_2(x,y;\ \rho)$. $\phi(x)$ is the standard normal density function and

$$\frac{\partial x}{\partial\theta} = \frac{\partial\Phi^{-1}(\theta)}{\partial\theta} = \frac{1}{\phi(x)}.$$

3.4.3 Sample size estimation based on pilot data

The three variance formulas presented in previous sections are based on parametric assumptions and only use summary statistics. We now propose a nonparametric method for estimating the sample size based on available pilot data from which we can use to estimate the variance needed for sample size estimation. We first describe the method then apply it to the data presented by Wieand et al. (1989) involving two biomarkers for detecting pancreatic cancer. This data set contains 91 subjects with pancreatic cancer and 50 subjects without the disease.

Consider a study with a total of *N* observations, with *m* denoting the number of subjects without disease and *n* denoting the number of subjects with disease. Let X_i $(i = 1, 2, \ldots, m)$ be the test results for the sample of *m* subjects without disease and Y_j $(j = 1, 2, \ldots, n)$ be the test results for the sample of *n* subjects with disease. Recall the area under the receiver operating characteristic curve is defined as

$$\text{AUC} = \Pr\,(X < Y)$$

which can be modified to handle ties as

$$\text{AUC} = \Pr(X < Y) + 0.5\,\Pr(X = Y)\,.$$

We can estimate AUC by considering the placement value for each observation. The placement value for an observation $Y_j, j = 1, 2, \ldots, n$, is the percentage of $X_i, i = 1, 2, \ldots, m$, observations that it exceeds (Hanley & Hajian-Tilaki, 1997; Zou, 2021). In other words, we consider the percentile in the sample *X* that observation Y_j takes up. The mean of the placement values is the nonparametric estimator for AUC. We can easily compute placement values using ranks.

The overall rank of a given observation Y_j is denoted by R_j in the combined sample of both *X* and *Y*. One less than the overall rank is the number of times Y_j is no less than the rest of the $m +$

$n-1$ observations. But the interest here is the number of times Y_j is no less than the m observations for subjects without disease. This can be achieved by $R_j^Y - 1$ where R_j^Y is the rank of Y_j in the sample Y with size n. Finally, the placement value for a single observation Y_j is given by

$$p_j = \frac{(R_j - 1) - (R_j^Y - 1)}{m} = \frac{R_j - R_j^Y}{m}, \qquad j = 1, 2, \dots, n\,.$$

Similarly, we can obtain the placement values for X_i as

$$q_i = \frac{R_i - R_i^X}{n}, \qquad i = 1,2, \dots, m$$

where R_i is the rank of X_i in the combined sample of X and Y, and R_i^X is its rank in the sample of X with size m. Note that the mean of p_i is the AUC, and the mean of q_i is $1 - \text{AUC}$. The variance of the AUC estimate can be estimated by

$$\widehat{\text{var}}\left(\hat{\theta}\right) = \frac{\text{var}(q_i)}{m} + \frac{\text{var}(p_j)}{n}$$

where

$$\text{var}(p_j) = \frac{\sum_{j=1}^{n}\left(p_j - \bar{p}\right)^2}{n-1}$$

and

$$\text{var}(q_i) = \frac{\sum_{i=1}^{m}(q_i - \bar{q})^2}{m-1}$$

where $\bar{p}$ and $\bar{q}$ are the means of the placement values p_j and q_i, respectively. Note that this variance estimator is identical to the one proposed by DeLong et al. (1988).

If we define $r = \frac{m}{n}$ then we have, $m = rn$ and $N = m + n$, and thus

$$m = \frac{r}{1+r}N$$

$$n = \frac{1}{1+r}N\,.$$

Substituting these into our variance formula gives

$$\operatorname{var}(\hat{\theta}) = \frac{1}{N}\left[\frac{1+r}{r}\left(\operatorname{var}(q_i) + r\operatorname{var}(p_j)\right)\right].$$

Now let $f(\theta)$ be the variance without the $1/N$ term

$$f(\theta) = \frac{1+r}{r}\left(\operatorname{var}(q_i) + r\operatorname{var}(p_j)\right).$$

To derive the sample size formula with a prespecified lower bound and assurance probability using the logit transformation, we first start with the equivalence

$$\begin{aligned} 1-\beta &= \Pr(\operatorname{lgt}\theta_L \geq \operatorname{lgt}\theta_0) \\ &= \Pr\left(\operatorname{lgt}\hat{\theta} - Z_{\frac{\alpha}{2}}\sqrt{\operatorname{var}\left(\operatorname{lgt}(\hat{\theta})\right)} \geq \operatorname{lgt}\theta_0\right) \\ &= \Pr\left(\operatorname{lgt}\hat{\theta} \geq \operatorname{lgt}\theta_0 + Z_{\frac{\alpha}{2}}\frac{\sqrt{f(\theta)}}{\sqrt{N}\theta(1-\theta)}\right). \end{aligned}$$

After standardizing, we get the formula for assurance probability:

$$Z_\beta = \frac{-\operatorname{lgt}\theta_0 + \operatorname{lgt}\theta - Z_{\alpha/2}\frac{\sqrt{f(\theta)}}{\sqrt{N}\theta(1-\theta)}}{\frac{\sqrt{f(\theta)}}{\sqrt{N}\theta(1-\theta)}}.$$

Solving for N and then substituting $f(\theta)$ with the individual variances gives us the total sample size formula

$$N = \frac{(r+1)}{r}\left[\frac{Z_{\alpha/2} + Z_\beta}{\operatorname{lgt}\theta - \operatorname{lgt}\theta_0}\right]^2\left[\frac{r\operatorname{var}(p_j) + \operatorname{var}(q_i)}{\theta^2(1-\theta)^2}\right]. \tag{3.6}$$

We only considered the case of estimating sample size with a prespecified lower limit because a larger upper bound is always better as this means a test or tool would be more accurate. Thus, limiting the upper bound is generally not a concern, so we decided to focus on obtaining a sample size estimate with precision when a lower limit is prespecified.

Chapter 4 Simulation

The sample size formulas in Chapter 3 were derived based on large sample theory, and in this chapter, we evaluated their performance in finite samples via simulation studies. Simulation studies allow us to compare the empirical assurance probability of each variance model to a prespecified assurance probability. Sample size estimates were first determined using the proposed sample size formulas, using either a prespecified lower bound or confidence interval half-width. Then data was then generated based on this sample size and the AUC and its variance for the generated data were estimated, and confidence intervals were be constructed around this AUC estimate. The empirical assurance probability was obtained based on how many of those confidence intervals exclude the prespecified lower bound or are narrower than the prespecified width. This empirical assurance probability was then compared to the prespecified assurance probability of 50, 80, and 90 percent. SAS 9.4 was used to perform the simulation study.

4.1 Achieving a prespecified lower limit

The first part of our simulation study was to evaluate the performance of the proposed sample size formula when a prespecified lower bound is incorporated, and when three different variance estimators are used. We determined a sample size with prespecified assurance such that the lower bound of a confidence interval around the AUC is no less than a certain preset limit.

4.1.1 Study design

To resemble the most realistic scenarios in the real world, we used a ratio of typical patients to atypical patients (r) of 0.5 to 2, as there may be more or less atypical patients than typical patients depending on where the sample may be acquired. We set true values of AUC (θ) in the range of 0.7 to 0.9 with increments of 0.1, while the required lower limits (θ_0) between 0.5 and 0.85. The ratio of standard deviations (B) between the typical and atypical groups was set to be 0.5 to 1, which is under the assumption that the variability amongst the atypical group may be larger than that of the typical group. Lastly, we used a prespecified assurance probability of 0.5, 0.8, and 0.9.

The empirical assurance probability (EAP) was defined as the number of times the lower bound of a 95 percent one-sided confidence interval around AUC estimate $\hat{\theta}$ would be no less than a prespecified lower bound θ_0. The empirical coverage percentage was also considered and defined as the number of times the true AUC θ is within the 95 percent confidence interval around the estimate $\hat{\theta}$. In order to be considered as good performance, the coverage must be close to 95 percent and the EAPs must be as close to the desired assurance probability as possible or greater than it.

The 50 percent assurance level would be our control group, as this is the amount of assurance we would get from a traditional confidence interval based methods. This would mean that there is no assurance of achieving the desired confidence interval, only that it is assumed to be correct.

4.1.2 Data generation

First, we determined a minimum sample size to generate data. Thus, we took various combinations of θ, θ_0, r, and B and used each of the three variance formulas to calculate the variance component $f(\theta)$ based on these combinations. They were then used in our sample size formula (Equation 3.4):

$$N = \left(\frac{Z_\beta + Z_{\alpha/2}}{\operatorname{lgt}\theta - \operatorname{lgt}\theta_0}\right)^2 \frac{f(\theta)}{\theta^2(1-\theta)^2}.$$

After N was determined using the sample size formulas derived in Chapter 3, we used this sample size to generate 10000 data sets. The data in the typical group were generated from a normal distribution with a mean of 0 and a variance of 1, while the data in the atypical group were from a normal distribution with a mean of μ_n and a variance of $\frac{1}{B^2}$.

The mean and variance of the two groups can be arbitrary because it is their relation to each other that creates a ROC curve. Any mean can be used as a starting point as long as the AUC that is created from these two populations would be based off of a desired AUC of θ. The mean of the atypical group μ_n was calculated to achieve the desired AUC of θ as follows:

$$\Pr(X < Y) = \Phi\left(\frac{\mu_m - \mu_n}{\sqrt{{\sigma_n}^2 + {\sigma_m}^2}}\right)$$

$$\theta = \Phi\left(\frac{\mu_n}{\sqrt{\frac{1}{B^2} + 1}}\right)$$

$$\Phi^{-1}(\theta) = \frac{\mu_n}{\sqrt{\frac{1}{B^2} + 1}}$$

$$\Rightarrow \mu_n = \sqrt{1 + \frac{1}{B^2}}\ \Phi^{-1}(\theta)$$

where μ_m and μ_n are the means of the typical and atypical groups, respectively, and σ_m and σ_n are their standard deviations.

Datasets of size *N* were generated, and the simulation was run 10000 times for each sample size. Afterwards, we estimated the area under the ROC curve $\hat{\theta}$, and its variance, from the generated data using the nonparametric method by DeLong et al. (1988). The nonparametric formula for the variance of this AUC is

$$\widehat{\text{var}}(\hat{\theta}) = \frac{\text{var}(q_i)}{m} + \frac{\text{var}(p_j)}{n}$$

where

$$\text{var}(p_j) = \frac{\sum_{j=1}^{n}(p_j - \bar{p})^2}{n-1}$$

$$\text{var}(q_i) = \frac{\sum_{i=1}^{m}(q_j - \bar{q})^2}{m-1}$$

similar to the method described in Section 3.3.3. Then logit transformed Wald confidence intervals were constructed around the AUC estimate $\hat{\theta}$ so that the confidence intervals would not be beyond the AUC range of (0,1). The lower bound of a logit transformed two-sided 95 percent confidence interval is

$$\text{lgt}\,\hat{\theta} - Z_{\alpha/2}\sqrt{\widehat{\text{var}}(\text{lgt}\,\hat{\theta})}$$

where

$$\widehat{\text{var}}(\text{lgt}\,\hat{\theta}) = \frac{\widehat{\text{var}}(\hat{\theta})}{\hat{\theta}^2(1-\hat{\theta})^2}$$

by the delta method.

Finally, the confidence interval was transformed back onto the raw scale in order to be compared to the prespecified lower bound, θ_0. The empirical assurance probability was calculated

as the number of times this lower bound was greater than the prespecified lower bound, θ_0. The coverage probability was also calculated as the number of times the true AUC value θ was within the 95 percent confidence interval around $\hat{\theta}$.

4.1.3 Results

The simulation results for the sample sizes for achieving a prespecified lower limit based on three variance formulas are summarized in Tables 4.1a to 4.3c. Tables 4.1a to 4.1c display the results of the simulation at the 50 percent assurance level, Tables 4.2a to 4.2c are at the 80 percent assurance level, and Tables 4.3a to 4.3c are at the 90 percent assurance level. The first table of each assurance level (a) contains results when $\theta = 0.9$, the second (b) contains results when $\theta = 0.8$, and the third (c) contains results when $\theta = 0.7$. The coverage probabilities for all sections are very close to 95 percent, indicating that these confident intervals were properly constructed.

There are a few general trends that can be seen within the sample size and the empirical assurance probability. In terms of sample size, it is noticeable that the required sample size tends to be larger when the difference between θ and lower limit θ_0 is small. For example, this can be seen in Table 4.1a, where $N = 197$ when $\theta = 0.9$ and $\theta_0 = 0.85$, compared to $N = 64$ when $\theta = 0.9$ and $\theta_0 = 0.8$, holding both $r = 0.5$ and $B = 0.5$ constant. This pattern is expected because in general, when a clinical difference is larger and easier to detect, the required sample size is smaller.

Tables 4.1a to 4.3c show that the ratio standard deviations (B) of the typical to atypical groups has no effect on the sample sizes based on the exponential model or the probit model. This is because the exponential and probit models do not incorporate the standard deviation ratio in their respective variance formulas. However, we can see that the standard deviation ratio does affect the sample size of the binormal model's variance formula. As the standard deviation ratio B

increases, the sample size calculated using the binormal based variance formula tends to increase as well. In all of our simulation runs, this sample size turned out to be greatest when $B = 1$ and smallest when $B = 0.5$. For instance, this can be seen in Table 4.1b when $\theta = 0.8$, $\theta_0 = 0.75$, $r = 0.5$ are held constant and B is the only variable changing, the value of N increases from 341 to 395 to 452 as B increases from 0.5 to 0.75 to 1, respectively.

The group size ratio r also has an effect on the required sample size. As r increases, the minimum sample size N tends to increase for the exponential and binormal variance methods (see Tables 4.1a to 4.3c). For example, when $\theta = 0.9$, $\theta_0 = 0.85$, $B = 0.5$, and $r = 0.5$ (Table 4.2a), the N of the exponential based variance starts at 401 but then increases to 461, 545, and 635 as group size ratio r increases from 0.5 to 1 to 2. For the variance estimator based on the probit model, the sample size is actually smallest when group size ratio $r = 1$ and tends to be greater when r is not 1. This exception can also be seen in the exponential based variance model only when $\theta = 0.7$, as the sample size calculated using the binormal based variance becomes smallest when $r = 1$ and greatest when r is far from 1 (Tables 4.1c, 4.2c, and 4.3c).

Sample sizes also decrease as the prespecified assurance probability decreases—the required sample sizes were smallest for the 50 percent assurance condition (see Tables 4.1a to 4.1c), and greatest for the 90 percent assurance condition (see Tables 4.3a to 4.3c). This is to be expected as a greater sample size is needed for greater precision when conducting studies so that a desired outcome can be found with more assurance. The results also show that for the same difference of 0.1 between θ and θ_0, the required sample sizes are small when θ is closer to 1, and large when θ is closer to 0.5. This can be seen in Tables 4.1a and 4.1b where the exponential based variance requires sample sizes between 64 to 102 when $\theta = 0.9$ and $\theta_0 = 0.8$, but increase to 97 to 131 when $\theta = 0.8$ and $\theta_0 = 0.7$. This pattern matches the findings by Hanley and McNeil

(1982) where it was noticed that "a difference of 10% is more easily detected if it is a difference between 80% and 90% than if it is a difference between 70% and 80%" (Hanley & McNeil, 1982). Although this finding was for comparison of multiple AUCs, we can see that it also applies to a single AUC estimation as well.

In terms of trends in the EAPs, the standard deviation ratio B seems to have an impact on the empirical assurance probability for all three variance formulas. As the standard deviation ratio increases, the EAP tends to increase as well such that it is greatest when $B = 1$ and smallest when $B = 0.5$. The exception to this pattern is seen in the exponential and probit based variance formulas when $r = 0.5$. When group size ratio $r = 0.5$, the EAPs for these two methods are actually largest when standard deviation ratio $B = 0.5$ and are smallest when $B = 1$. For example, this can be seen in Table 4.1c—the EAP of the exponential based variance starts at 53.46 percent when group size ratio $r = 0.5$ and standard deviation ratio $B = 0.5$ but drops to 46.29 percent when $r = 0.5$ and $B = 1$. For all other values of group size ratio r, the usual pattern is seen where the EAP is greatest when $B = 1$, and smallest when $B = 0.5$. This may be because the sample size N is not changing with respect to B for the exponential and probit variance models, and thus B would only affect the standard deviations of the generated data. With a skewed standard deviation ratio, the variance calculated using a nonparametric method may be larger, and thus cause the EAP to be smaller.

The 50 percent assurance level is the same amount of assurance that traditional confidence interval based methods use. Therefore, we simulated results for this level in order to act as a control that would allow us to more easily compare the performance of the three different variance formulas. First, we examined the empirical assurance probabilities of the three models at the most neutral condition, when group size ratio $r = 1$ and standard deviation ratio $B = 1$. Under these conditions, it is clear that the EAP calculated from the sample size formula using the binormal

based variance is noticeably greater than the others, especially when θ is large. In fact, when θ is 0.9 and 0.8 and θ_0 is 0.05 less than θ, the binormal model's EAP tends to surpass the 50 percent mark, as seen in Table 4.1a and 4.1b. However, as θ decreases, the binormal model's EAP drops slightly and becomes closer to 50 percent (see Table 4.1c). On the other hand, the exponential based variance and probit based variances tend to be fairly consistent across all values of θ. The EAPs of all three variance models tend to be smallest when B is small, and largest when B is large.

At the 80 percent assurance level, all three methods generally perform decently since their EAPs are all very close to the prespecified assurance probability of 80 percent. The binormal model appears to have EAPs overshoot the 80 percent mark quite frequently when θ is large, with its EAP going above 80 percent and sometimes into the 90 percent range, as shown in Table 4.2a and 4.2b. This could be an indication that the binormal based estimate of variance may be too conservative. Notably, the binormal based variance formula tends to make the EAP overshoot when standard deviation ratio $B \neq 0.5$. This goes along with the findings by Obuchowski (1994) that the binormal based variance estimate is conservative when the standard deviation ratio is 1, however it decreases as B decreases. In our simulation study, the binormal based variance causes EAPs to overshoot and performs quite conservatively when standard deviation ratio $B = 1$ and $B = 0.75$, making the required sample size as well as EAP increase. On the other hand when $B = 0.5$, a non-conservative variance means the required sample size is smaller and thus EAP also tends to be slightly smaller.

Similarly, the EAPs from the exponential and probit based variance formulas surpass 80 percent, as shown in Tables 4.2a to 4.2c. When group size ratio $r = 0.5$ and standard deviation ratio $B = 0.5$ and 0.75, the probit model's EAP appears steady at 80 percent and higher. On the other

hand, the EAP of the exponential model appears to surpass the 80 percent mark when the standard deviation ratio is larger, like $B = 1$ and sometimes $B = 0.75$.

At the 90 percent assurance level, we can see some of the same patterns as with the other assurance levels. The binormal model's EAPs overshoot a lot more than at 50 and 80 percent assurance, especially when θ is large, as seen in Table 4.3a and 4.3b. As θ decreases, as shown in Table 4.3c, the binormal model's EAPs decrease as well and tend to go beyond 90 percent only when standard deviation ratio $B \neq 0.5$, once again matching what was illustrated by Obuchowski (1994). In general, the cases that surpassed 80 percent in the previous condition also tend to surpass 90 percent in this condition. However, there are even more EAPs at the 90 percent assurance level that surpass 90 percent (see Table 4.3a). Additionally, the required sample sizes are much larger here than for the previous assurance levels (see Tables 4.3a to 4.3c). This is expected since the 90 percent assurance level is the most precise and would require a larger sample size to ensure that small clinical differences are detected with confidence.

Overall, the sample size formula with three different variance models perform similarly. The binormal based variance seems to perform slightly better than the other two as it provides more conservative estimates. Although it tends to overestimate the prespecified assurance probability more frequently than the exponential and probit based variances, it is also the most conservative variance estimator as the sample sizes are larger than the ones from the other two methods.

Table 4.1 a: Empirical assurance probabilities at the 50% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.9$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	θ_0	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.9	0.85	0.5	0.5	197	49.97	95.29	218	54.39	94.73	224	54.79	94.72
			0.75		49.00	94.85	256	60.67	95.13		54.45	95.31
			1		45.80	95.05	301	63.88	95.25		50.83	94.97
		1	0.5	226	48.83	94.96	258	53.70	94.62	199	42.94	95.08
			0.75		54.16	94.90	278	64.28	95.24		48.68	95.08
			1		56.14	95.06	305	70.23	95.14		51.01	95.03
		1.5	0.5	267	48.59	94.94	309	56.16	94.85	207	38.44	94.71
			0.75		56.62	94.88	321	66.40	95.20		45.92	94.90
			1		61.76	94.99	341	74.68	94.95		50.31	94.95
		2	0.5	311	49.05	94.81	363	56.01	94.70	224	36.03	94.95
			0.75		58.97	94.76	369	67.42	94.89		44.68	95.19
			1		66.11	95.14	385	75.93	95.64		51.56	94.90
	0.8	0.5	0.5	64	46.31	95.44	71	53.20	95.39	73	55.06	95.43
			0.75		47.07	95.20	84	60.82	95.26		54.34	95.41
			1		43.99	95.11	98	64.00	95.15		50.07	94.85
		1	0.5	74	47.19	94.35	84	54.34	94.82	65	41.29	94.43
			0.75		54.20	95.10	91	65.98	94.68		47.11	94.82
			1		55.14	95.56	100	72.18	95.62		49.10	95.23
		1.5	0.5	87	48.23	93.84	101	55.58	94.23	68	34.50	93.39
			0.75		57.47	94.61	105	66.60	94.52		43.90	94.22
			1		63.58	95.55	111	75.99	95.16		48.85	95.25
		2	0.5	102	46.91	93.48	118	55.76	93.70	73	33.41	93.07
			0.75		59.15	94.12	120	68.97	94.33		43.94	94.45
			1		66.47	94.86	126	76.91	95.28		50.22	94.68

Note: Empirical assurance probability (EAP) is the frequency of times the lower bound of the 95 percent CI is above the prespecified lower bound of θ_0. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

Table 4.1 b: Empirical assurance probabilities at the 50% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.8$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	θ_0	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.8	0.75	0.5	0.5	339	51.62	95.37	341	50.97	95.50	385	56.06	95.38
			0.75		49.41	94.83	395	55.28	95.29		54.42	95.35
			1		45.50	95.48	452	57.59	94.93		50.84	95.21
		1	0.5	355	46.76	95.51	393	51.94	95.21	342	45.39	95.10
			0.75		51.51	95.31	411	56.54	95.00		49.75	95.31
			1		52.21	94.61	436	61.08	95.07		49.53	94.91
		1.5	0.5	403	46.04	95.20	466	51.40	95.17	357	40.88	94.75
			0.75		51.98	95.28	466	57.70	94.89		46.36	94.87
			1		54.65	95.34	475	62.34	95.25		50.86	95.68
		2	0.5	460	44.60	94.82	544	50.78	95.01	385	38.00	95.05
			0.75		51.74	95.13	530	58.53	94.8		44.92	94.94
			1		57.96	94.74	528	63.50	94.93		50.17	94.80
	0.7	0.5	0.5	97	50.43	95.62	97	49.91	95.69	110	55.27	95.57
			0.75		47.89	95.07	113	54.67	95.07		52.93	95.68
			1		43.24	94.98	129	56.92	95.83		50.05	95.20
		1	0.5	102	45.68	95.61	112	51.14	95.35	98	44.68	95.19
			0.75		49.85	95.78	118	56.88	95.05		48.28	95.39
			1		51.87	95.33	124	59.85	95.35		48.87	95.12
		1.5	0.5	115	43.67	94.85	133	51.54	95.10	102	39.32	94.94
			0.75		51.32	95.24	133	58.53	95.38		45.09	95.06
			1		53.83	95.59	136	63.25	95.35		49.68	95.22
		2	0.5	131	43.23	94.65	155	49.99	95.31	110	36.87	95.17
			0.75		51.73	95.07	151	58.96	94.77		44.03	94.60
			1		56.41	95.40	151	65.46	95.17		49.58	95.30

Note: Empirical assurance probability (EAP) is the frequency of times the lower bound of the 95 percent CI is above the prespecified lower bound of θ_0. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

Table 4.1 c: Empirical assurance probabilities at the 50% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.7$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	θ_0	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.7	0.65	0.5	0.5	460	53.46	94.71	411	48.50	95.11	503	57.21	95.09
			0.75		49.94	95.22	471	51.45	95.01		53.27	94.88
			1		46.29	95.19	528	52.10	95.11		49.99	95.10
		1	0.5	452	46.39	95.26	465	48.48	95.25	447	45.74	95.09
			0.75		49.58	95.16	475	51.94	95.11		49.69	95.06
			1		50.63	94.90	488	53.25	94.40		50.24	95.12
		1.5	0.5	497	44.60	95.22	546	48.41	94.84	466	42.26	94.96
			0.75		49.16	94.83	530	52.03	95.32		46.71	94.85
			1		53.02	94.92	520	53.55	94.70		49.87	95.30
		2	0.5	556	43.59	95.15	635	47.51	94.76	503	38.97	95.05
			0.75		49.35	94.92	597	51.32	95.45		45.63	95.16
			1		53.64	95.38	570	54.74	95.36		50.45	94.90
	0.6	0.5	0.5	123	51.75	95.41	110	47.44	95.10	135	55.82	95.71
			0.75		49.19	95.41	126	50.15	95.35		52.94	95.42
			1		45.96	95.48	141	52.18	94.74		50.00	95.24
		1	0.5	121	45.89	95.13	124	46.65	95.19	120	45.75	95.14
			0.75		48.81	95.19	127	51.87	95.14		48.45	95.21
			1		50.29	95.62	131	53.21	95.04		48.81	95.47
		1.5	0.5	133	43.76	94.96	146	46.73	94.90	125	41.64	95.39
			0.75		48.70	95.61	142	50.08	94.94		45.56	95.09
			1		53.01	95.41	139	53.88	95.00		48.80	95.16
		2	0.5	149	42.51	95.34	170	47.28	95.42	135	38.03	95.33
			0.75		48.84	95.29	160	52.04	95.37		44.50	95.41
			1		53.16	95.35	152	54.49	95.23		48.44	95.16

Note: Empirical assurance probability (EAP) is the frequency of times the lower bound of the 95 percent CI is above the prespecified lower bound of θ_0. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

Table 4.2 a: Empirical assurance probabilities at the 80% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.9$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	θ_0	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.9	0.85	0.5	0.5	401	82.55	94.94	444	86.50	95.04	457	87.41	94.91
			0.75		81.06	94.79	524	90.89	95.05		86.99	95.17
			1		77.98	95.22	615	92.63	94.95		83.78	95.25
		1	0.5	461	80.96	95.30	527	86.62	94.98	406	76.86	94.59
			0.75		86.53	95.02	568	92.84	94.90		81.85	95.08
			1		88.12	95.21	623	95.33	94.88		83.73	94.93
		1.5	0.5	545	80.99	94.92	631	86.54	94.56	423	68.91	94.90
			0.75		88.34	94.95	655	93.97	95.26		79.35	94.88
			1		91.89	94.79	697	96.89	95.11		83.48	95.15
		2	0.5	635	81.09	95.12	741	86.91	94.65	457	66.35	95.03
			0.75		89.50	95.42	754	93.77	94.79		77.07	95.13
			1		93.28	94.86	787	97.48	94.98		83.30	94.93
	0.8	0.5	0.5	131	84.64	95.31	145	88.39	95.20	149	89.38	95.56
			0.75		83.32	95.01	171	93.00	95.06		88.39	95.15
			1		79.04	95.02	200	93.91	94.91		85.09	94.77
		1	0.5	150	83.56	95.08	172	88.69	95.06	133	78.56	95.41
			0.75		87.46	95.00	185	94.52	95.12		83.55	94.90
			1		89.22	95.12	203	96.77	95.04		85.07	95.23
		1.5	0.5	178	82.43	94.45	206	87.97	94.43	138	71.43	94.30
			0.75		89.67	94.91	214	95.22	95.21		80.17	94.79
			1		93.61	95.25	227	98.05	95.47		85.75	95.20
		2	0.5	207	83.06	94.71	242	89.13	94.64	149	68.37	94.54
			0.75		91.01	94.80	246	95.18	95.13		78.25	94.34
			1		94.48	95.00	256	98.06	94.75		84.56	94.77

Note: Empirical assurance probability (EAP) is the frequency of times the lower bound of the 95 percent CI is above the prespecified lower bound of θ_0. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

Table 4.2 b: Empirical assurance probabilities at the 80% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.8$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	θ_0	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.8	0.75	0.5	0.5	692	82.58	95.28	696	82.72	94.98	787	86.46	94.89
			0.75		79.46	95.05	807	85.42	94.96		85.47	95.22
			1		75.46	95.13	924	87.92	95.12		81.05	94.92
		1	0.5	725	77.83	95.29	803	81.95	94.91	699	76.47	94.88
			0.75		81.69	95.15	840	87.52	94.97		80.07	94.96
			1		82.77	95.11	890	89.71	94.92		81.02	95.05
		1.5	0.5	824	77.28	95.07	952	82.87	95.06	728	72.16	95.13
			0.75		83.00	95.11	952	88.25	94.85		78.14	94.74
			1		86.41	95.06	970	91.10	95.39		81.48	95.15
		2	0.5	939	75.95	95.06	1111	82.57	94.98	787	67.37	95.20
			0.75		83.15	95.02	1083	88.70	95.22		76.23	95.17
			1		88.32	95.10	1078	91.80	95.07		81.55	95.07
	0.7	0.5	0.5	197	83.04	94.84	199	83.83	95.26	224	88.36	94.91
			0.75		80.10	95.23	230	86.74	95.32		86.01	94.92
			1		77.20	95.15	264	87.64	94.83		82.43	95.28
		1	0.5	207	79.27	95.08	229	83.84	95.07	200	77.73	95.69
			0.75		82.74	95.25	240	88.52	95.04		81.10	95.49
			1		84.03	95.03	254	90.85	95.11		82.81	95.28
		1.5	0.5	235	77.61	95.16	272	83.97	95.22	208	70.94	95.03
			0.75		83.99	95.56	272	89.17	94.96		78.74	94.73
			1		87.11	95.14	277	91.75	95.09		82.65	95.57
		2	0.5	268	76.25	94.84	317	84.46	95.09	224	68.22	95.17
			0.75		84.21	94.98	309	89.74	95.16		77.39	95.27
			1		88.83	95.19	307	92.93	95.19		82.10	95.08

Note: Empirical assurance probability (EAP) is the frequency of times the lower bound of the 95 percent CI is above the prespecified lower bound of θ_0. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

Table 4.2 c: Empirical assurance probabilities at the 80% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.7$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	θ_0	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.7	0.65	0.5	0.5	940	83.96	94.94	839	78.47	94.78	1027	86.13	95.25
			0.75		80.99	94.97	961	81.90	94.97		84.34	94.76
			1		77.28	94.47	1079	82.89	95.16		81.44	95.42
		1	0.5	922	77.89	95.08	949	78.40	95.03	913	76.73	94.61
			0.75		80.80	94.59	969	81.76	95.18		79.82	94.96
			1		80.94	94.87	997	84.53	95.02		80.74	94.90
		1.5	0.5	1015	74.77	94.94	1115	78.44	95.12	951	71.77	94.74
			0.75		80.26	95.01	1082	82.46	94.76		77.45	94.90
			1		83.31	95.22	1062	84.54	95.28		81.47	94.94
		2	0.5	1135	72.80	94.76	1296	78.69	95.12	1027	69.03	94.88
			0.75		79.68	94.79	1220	82.80	94.92		76.56	95.04
			1		84.85	94.90	1164	85.10	94.82		81.39	95.12
	0.6	0.5	0.5	251	84.05	94.94	224	79.86	94.89	274	87.27	95.04
			0.75		81.02	95.31	257	82.14	95.21		84.55	95.52
			1		77.39	95.56	288	82.65	95.02		80.75	95.01
		1	0.5	247	78.14	95.55	254	78.96	95.21	244	76.47	94.97
			0.75		81.61	95.31	259	82.57	95.11		79.98	94.95
			1		81.34	94.73	266	84.77	95.56		81.36	94.72
		1.5	0.5	271	75.11	95.36	298	79.20	95.02	254	72.45	94.79
			0.75		80.19	95.15	289	82.56	95.05		77.93	95.07
			1		83.35	95.19	284	85.20	95.24		81.28	95.23
		2	0.5	303	73.25	95.29	346	79.01	95.27	274	68.39	95.24
			0.75		80.49	95.30	326	82.71	95.20		76.50	94.99
			1		84.64	95.06	311	85.56	95.04		80.74	94.83

Note: Empirical assurance probability (EAP) is the frequency of times the lower bound of the 95 percent CI is above the prespecified lower bound of θ_0. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

Table 4.3 a: Empirical assurance probabilities at the 90% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.9$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	θ_0	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.9	0.85	0.5	0.5	537	92.64	95.16	594	94.56	95.18	612	94.61	94.69
			0.75		91.71	94.99	701	96.99	94.95		94.91	95.00
			1		88.82	95.01	823	97.71	94.76		92.21	94.94
		1	0.5	616	91.39	95.01	705	94.96	95.26	544	87.94	95.19
			0.75		94.92	95.24	760	98.16	95.28		91.31	94.88
			1		95.46	95.20	834	99.00	95.09		92.57	95.04
		1.5	0.5	729	91.32	94.91	845	95.31	94.7	566	82.86	94.94
			0.75		95.79	95.24	877	98.09	94.94		89.74	94.96
			1		97.43	95.20	932	99.39	94.95		92.6	95.05
		2	0.5	850	91.37	95.07	992	94.72	94.97	612	79.06	95.04
			0.75		96.34	95.06	1009	98.14	94.99		88.33	95.09
			1		98.03	94.93	1053	99.39	94.60		92.54	95.31
	0.8	0.5	0.5	175	93.48	94.62	194	95.94	95.58	199	96.82	95.59
			0.75		93.55	95.27	228	98.01	95.54		96.2	95.20
			1		90.76	95.01	268	98.72	95.23		94.11	94.96
		1	0.5	201	93.45	94.91	230	96.38	95.17	177	89.51	95.00
			0.75		96.13	95.21	248	98.65	95.43		93.39	95.19
			1		96.29	95.04	272	99.49	95.26		93.73	94.58
		1.5	0.5	238	93.08	94.67	275	96.09	94.52	185	84.87	94.93
			0.75		96.71	94.92	286	98.92	95.21		91.15	94.84
			1		98.15	95.19	304	99.66	95.38		93.99	95.13
		2	0.5	277	92.65	94.87	323	96.26	95.10	199	80.57	94.28
			0.75		97.27	94.63	329	99.15	95.18		90.20	95.09
			1		99.00	95.20	343	99.72	95.34		94.02	95.26

Note: Empirical assurance probability (EAP) is the frequency of times the lower bound of the 95 percent CI is above the prespecified lower bound of θ_0. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations

Table 4.3 b: Empirical assurance probabilities at the 90% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.8$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	θ_0	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.8	0.75	0.5	0.5	926	91.28	95.10	931	91.78	94.99	1053	94.41	94.46
			0.75		90.49	94.90	1080	93.97	95.10		93.56	95.17
			1		87.55	94.72	1236	95.03	95.39		91.35	95.17
		1	0.5	970	88.80	94.89	1075	91.69	94.94	936	88.10	94.83
			0.75		91.55	94.91	1125	95.03	95.13		90.47	95.13
			1		91.93	95.11	1191	96.33	95.06		91.14	94.97
		1.5	0.5	1103	88.06	94.99	1275	91.75	94.81	975	83.12	95.06
			0.75		92.37	95.00	1274	95.39	94.74		88.70	95.06
			1		94.19	94.98	1298	97.15	95.46		91.37	95.05
		2	0.5	1257	86.11	95.05	1488	92.01	95.17	1053	80.59	94.85
			0.75		92.56	95.13	1450	95.59	95.01		87.12	94.96
			1		95.59	95.25	1443	97.43	95.42		91.44	95.15
	0.7	0.5	0.5	264	92.67	95.17	266	92.84	95.04	300	95.22	94.87
			0.75		91.17	94.90	308	94.88	95.42		94.10	94.94
			1		87.80	95.34	353	95.63	95.17		92.05	94.81
		1	0.5	277	89.76	94.84	307	93.33	95.01	267	88.63	94.94
			0.75		92.95	95.50	321	95.38	94.75		91.50	95.05
			1		93.37	95.35	340	97.17	95.41		92.27	95.03
		1.5	0.5	314	88.51	94.74	363	93.07	94.92	278	84.58	95.22
			0.75		93.08	95.05	363	96.11	95.28		89.78	94.89
			1		95.11	95.53	370	97.52	94.83		92.01	95.07
		2	0.5	358	87.90	95.07	424	92.78	95.08	300	81.83	94.85
			0.75		93.86	95.09	413	96.46	95.24		87.74	95.24
			1		95.76	95.15	411	97.95	95.38		92.02	94.97

Note: Empirical assurance probability (EAP) is the frequency of times the lower bound of the 95 percent CI is above the prespecified lower bound of θ_0. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. r is the ratio between the size of the typical and atypical populations, B is the ratio between their standard deviations, and N is the total sample size from the two populations.

Table 4.3 c: Empirical assurance probabilities at the 90% assurance level for three variance formulas such that the lower bound of a two-sided 95% confidence interval for the AUC is not below the prespecified lower limit θ_0 when the true AUC $\theta = 0.7$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	θ_0	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.7	0.65	0.5	0.5	1258	92.44	94.94	1123	89.61	94.93	1374	94.40	94.79
			0.75		90.82	95.01	1287	91.26	95.06		92.96	94.87
			1		88.31	95.21	1444	92.31	95.12		90.36	95.09
		1	0.5	1235	88.71	94.64	1270	89.15	95.26	1222	87.91	94.83
			0.75		90.68	94.89	1298	92.43	95.3		90.17	95.17
			1		90.99	95.49	1334	92.76	94.82		90.49	95.03
		1.5	0.5	1358	86.53	95.02	1493	89.04	94.87	1273	83.68	94.49
			0.75		90.24	95.00	1448	92.04	95.02		88.68	95.28
			1		92.13	95.09	1421	93.18	94.90		90.52	94.56
		2	0.5	1519	84.74	94.55	1735	89.19	94.97	1374	80.50	95.32
			0.75		89.97	94.71	1633	92.51	95.09		87.34	94.37
			1		93.12	94.93	1558	93.86	95.18		90.58	94.91
	0.6	0.5	0.5	336	93.11	95.28	300	90.34	95.32	367	94.71	94.94
			0.75		91.02	94.94	344	91.63	94.64		93.19	94.99
			1		88.17	95.36	386	92.02	94.81		90.88	95.14
		1	0.5	330	89.14	95.51	339	89.12	95.06	326	88.25	95.27
			0.75		90.66	95.41	347	92.35	94.99		90.76	94.84
			1		91.47	95.05	356	93.40	95.37		90.40	95.18
		1.5	0.5	363	86.21	95.15	399	89.72	95.13	340	84.61	95.25
			0.75		90.46	94.83	387	92.30	94.81		89.05	95.16
			1		92.63	95.04	380	93.84	95.22		91.11	95.25
		2	0.5	406	85.56	95.06	463	89.76	95.04	367	81.82	94.98
			0.75		90.68	94.97	436	92.55	95.09		87.35	94.98
			1		93.73	95.36	416	93.75	95.14		90.89	95.13

Note: Empirical assurance probability (EAP) is the frequency of times the lower bound of the 95 percent CI is above the prespecified lower bound of θ_0. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

4.2 Achieving a prespecified confidence interval width

In the second simulation study, we investigated the performance of the sample size formula for achieving a prespecified confidence interval half-width with prespecified assurance. The same three variance estimators were used in the sample size formula. We determined a sample size with prespecified assurance such that the half-width of a confidence interval for the AUC is no wider than a certain preset half-width.

4.2.1 Study design

The study design for this simulation was very similar to the previous simulation. We used the same values and definitions of r, θ, and B. This time instead of a lower bound, we used a preset confidence interval half-width ω, and to keep things consistent with the first simulation, ω was set to be 0.1 or 0.05. The assurance probabilities were also kept at 50 percent, 80 percent, and 90 percent. The empirical assurance probability was defined as the number of times the half-width of a two-sided 95 percent confidence interval around $\hat{\theta}$ is smaller than a prespecified half-width ω. The empirical coverage percentage was also obtained from the number of times the true AUC θ is within the 95 percent confidence interval around the estimate $\hat{\theta}$. In order to be considered as good performance, the coverage must be close to 95 percent and the EAPs must be as close to the desired assurance probability as possible or greater than it.

4.2.2 Data generation

First we used the sample size formula (Equation 3.5) to calculate the required sample size based on the values of all the variables. Recall Equation 3.5:

$$N = \left(\frac{\frac{\sqrt{f(\theta)}}{\theta(1-\theta)} + \sqrt{\frac{f(\theta)}{\theta^2(1-\theta)^2} + \frac{2\omega^* Z_\beta f^*}{Z_{\alpha/2}}}}{\frac{2\omega^*}{Z_{\alpha/2}}} \right)^2 .$$

The three different variance estimators were used in this sample size formula. After determining *N*, we generated 10000 data sets with *N* as the sample size. The typical group had the same mean of 0 and variance of 1, while the atypical group had a mean of μ_n and a variance of $\frac{1}{B^2}$ just like the previous simulation. The nonparametric method by DeLong et al. (1988) was used to estimate the AUC and its variance, then logit transformed 95 percent confidence intervals were constructed. The logit transformed intervals were then transformed back onto the raw scale to be compared with the prespecified half-width ω, and the empirical assurance probability (EAP) was then calculated as the proportion of times the confidence interval half-width was smaller or equal to the prespecified width ω.

4.2.3 Results

Tables 4.4a to 4.6c show that the results for this simulation do not seem to follow the trends like the previous section did, especially within the empirical assurance probabilities. The coverage appears to be within range of 95 percent, which indicate that the confidence intervals were properly constructed. Just like the previous simulation, the sample sizes calculated using the exponential and binormal variance formulas seem to increase as the group size ratio *r* increases, however the sample size based on the probit variance is smallest when group size ratio $r = 1$ and greatest when *r* is far from 1. However, in terms of the EAPs, there are many instances where the EAP jumps up much higher than the prespecified assurance probability or is drastically lower than it. When

the standard deviation ratio $B = 0.5$, the EAPs tends to be much lower than when $B \neq 0.5$, typically when $r \neq 0.5$ (see Tables 4.4a to 4.6c). The dramatic drops in EAPs tend to occur more within the exponential and probit models of variance, while the binormal model's EAP has extreme overshooting. The extreme overshooting occurs mainly when $B \neq 0.5$, and this occurs in the exponential model's EAPs too although they are not as extreme as in the other two variance models.

When the prespecified assurance level is 50 percent, the EAPs for all three variance formulas have dramatic jumps, ranging from as high as 99.84 (Table 4.4a) to as low as 0.16 (Table 4.4c). These extreme EAPs tend to occur more when θ is small (Table 4.4c) but also appear when θ is large (see Table 4.4a). Even at the control conditions of group size ratio $r = 1$ and standard deviation ratio $B = 1$, there are some extreme EAPs regardless of the value of θ. For example, in Table 4.4a, the binormal based variance's EAP reaches 98.70 when $\theta = 0.9$, $\omega = 0.05$, $r = 1$, and $B = 1$.

At the 80 percent assurance level, there are many cases when the empirical assurance probability is much lower than the prespecified assurance probability, with numbers as low as 0.78 (Table 4.5c). The dramatic drops occur when θ is small and especially when $B = 0.5$. It is interesting that for the same conditions that produce such low EAPs for one variance formula, a different variance formula may get an EAP as high as 100.

At the 90 percent assurance level, the required sample sizes are larger than at the 80 percent level, which is expected (see Tables 4.6a to 4.6c). However, as with the 80 percent assurance level, there are also atypical EAPs especially in the probit variance model when θ is small, as seen in Table 4.6c. The same pattern as found in the previous condition is found here—when standard deviation ratio $B = 0.5$ the EAPs often end up being abnormally low or very high. There are quite

a few cases that ended up with an EAP of 100, especially in the binormal variance model's EAPs, as shown in Table 4.6a to 4.6c.

The results were not expected though there may be some explanations as to why this sample size formula performed poorly in this simulation. One reason could be that for some cases, the atypical group may be the smaller group when it has a greater variance. When $B = 0.5$, this means that the standard deviation of the atypical group is twice the size of the standard deviation of the typical group. When $r = 2$, this means that the typical group is twice the size of the atypical group. If we put these together, that would create a large amount of variability amongst the atypical group, thus making the simulated group's distribution very wide and unpredictable. This may lead to confidence intervals that are much larger than the prespecified confidence interval half-width, which would in turn lead to very small or negligible EAPs.

Another possibility lies within the variance formula that was used to calculate the variance of the simulated data. Our simulation study generated data that was normally distributed, then we used the nonparametric method by DeLong et al. (1988) to calculate the AUC estimate and variance estimate. Nonparametric methods may work well on all types of data, but parametric methods may work even better when we know that the data is normally distributed. Thus, using the nonparametric method on the generated data could potentially cause the variance may be larger, and the confidence intervals to be wider than expected, and thus lead to very small EAPs.

Table 4.4 a: Empirical assurance probabilities at the 50% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.9$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.9	0.05	0.5	0.5	143	42.88	95.42	158	58.50	95.25	162	63.63	94.80
			0.75		43.25	95.18	186	81.81	94.83		60.98	94.88
			1		32.28	95.31	218	88.49	95.12		46.96	94.87
		1	0.5	164	40.37	95.15	187	61.13	94.85	144	26.89	94.58
			0.75		60.54	95.21	202	90.36	94.73		41.82	94.98
			1		66.66	95.06	221	98.70	95.47		46.43	95.24
		1.5	0.5	193	40.77	94.94	224	62.89	94.61	150	18.64	94.69
			0.75		68.78	94.82	233	92.39	95.11		34.17	94.65
			1		85.51	95.39	247	99.59	95.18		46.30	94.85
		2	0.5	226	41.68	94.37	263	61.99	94.87	162	15.73	94.42
			0.75		74.55	94.90	268	93.36	95.08		30.62	94.36
			1		92.87	95.49	279	99.84	94.98		47.99	95.09
	0.1	0.5	0.5	40	38.17	94.95	44	47.05	94.65	45	49.91	94.73
			0.75		40.95	95.00	52	69.10	95.31		49.80	95.07
			1		38.55	94.93	61	74.73	95.25		46.19	94.90
		1	0.5	46	39.64	93.64	52	51.08	93.75	40	31.37	93.31
			0.75		50.48	94.78	56	74.13	95.27		37.76	94.58
			1		53.94	95.29	62	88.00	95.23		40.43	94.90
		1.5	0.5	54	42.32	92.54	62	53.22	93.44	42	31.19	91.89
			0.75		59.81	94.58	65	77.73	94.45		38.76	94.57
			1		70.24	95.12	69	93.26	95.26		41.88	95.43
		2	0.5	63	42.25	91.42	73	56.18	93.40	45	30.58	90.68
			0.75		60.87	93.76	74	78.53	93.86		39.32	93.90
			1		75.66	94.97	78	93.89	95.02		45.83	94.52

Note: Empirical assurance probability (EAP) is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. r is the ratio between the size of the typical and atypical populations, B is the ratio between their standard deviations, and N is the total sample size from the two populations.

Table 4.4 b: Empirical assurance probabilities at the 50% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.8$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.8	0.05	0.5	0.5	288	48.86	95.17	289	51.13	95.23	327	91.10	95.17
			0.75		34.94	95.27	335	83.19	95.11		76.09	95.12
			1		16.37	95.08	384	90.01	94.70		45.25	95.20
		1	0.5	301	24.39	94.86	334	53.99	94.81	291	17.68	95.38
			0.75		50.08	95.43	349	93.40	95.09		39.20	95.28
			1		57.00	95.25	370	99.31	95.00		46.21	95.12
		1.5	0.5	342	17.99	95.00	396	55.91	94.81	303	5.08	95.08
			0.75		57.57	95.54	396	94.88	94.92		23.89	94.66
			1		84.06	95.28	403	99.82	95.17		44.61	95.49
		2	0.5	390	14.99	94.98	462	55.11	95.15	327	3.04	95.33
			0.75		60.60	94.77	450	94.31	94.83		17.82	94.88
			1		93.01	95.11	448	99.93	95.15		46.20	95.33
	0.1	0.5	0.5	72	40.17	95.38	73	45.67	95.71	82	68.91	95.25
			0.75		34.09	95.41	84	60.32	95.14		58.55	95.58
			1		24.58	95.32	97	72.51	94.79		43.80	95.44
		1	0.5	76	30.60	95.32	84	44.97	95.70	73	26.88	95.86
			0.75		41.51	95.29	88	70.64	94.99		37.20	95.19
			1		44.76	95.40	93	87.58	95.34		40.45	95.71
		1.5	0.5	86	28.91	95.21	100	47.44	95.24	76	17.76	95.19
			0.75		50.28	95.45	99	74.88	95.04		31.13	95.35
			1		64.15	95.18	101	91.58	95.00		41.80	95.20
		2	0.5	98	26.91	94.91	116	48.69	94.99	82	14.44	95.10
			0.75		50.31	95.36	113	77.08	95.38		28.00	95.13
			1		73.96	94.91	113	92.79	95.34		44.09	95.51

Note: Empirical assurance probability (EAP) is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

Table 4.4 c: Empirical assurance probabilities at the 50% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.7$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.7	0.05	0.5	0.5	421	87.26	95.16	376	20.43	95.19	460	99.97	95.48
			0.75		47.89	94.99	431	60.86	94.92		93.17	95.00
			1		10.62	95.28	483	70.47	94.86		46.04	94.90
		1	0.5	413	11.13	94.79	425	18.39	95.35	409	8.43	95.36
			0.75		40.24	94.71	434	69.57	95.21		34.05	95.53
			1		50.71	95.21	446	92.97	95.34		43.73	94.97
		1.5	0.5	455	2.64	95.35	500	18.41	95.36	426	0.54	95.36
			0.75		36.41	95.23	484	74.96	95.13		12.47	95.06
			1		82.31	94.87	476	97.04	94.64		44.35	95.25
		2	0.5	508	1.36	95.01	580	20.75	95.01	460	0.16	95.39
			0.75		35.99	94.76	546	73.45	94.93		6.72	95.05
			1		93.16	95.10	521	97.46	95.23		45.77	95.22
	0.1	0.5	0.5	104	61.89	95.35	93	22.39	95.34	114	91.37	95.04
			0.75		40.56	95.27	106	46.39	95.14		67.75	95.39
			1		21.14	95.04	119	52.87	95.47		38.74	95.30
		1	0.5	102	19.39	95.35	105	27.28	95.24	101	19.57	95.59
			0.75		33.38	95.47	107	51.10	95.15		33.75	95.20
			1		37.79	95.24	110	64.51	95.22		37.81	95.27
		1.5	0.5	112	12.81	95.25	123	28.73	95.40	105	7.87	95.47
			0.75		34.59	95.30	120	53.54	95.28		20.59	95.03
			1		58.76	95.12	117	73.77	95.25		36.24	95.46
		2	0.5	126	10.56	95.18	143	28.29	95.29	114	5.43	95.18
			0.75		35.73	95.06	135	54.70	95.55		17.78	95.02
			1		69.43	95.52	129	76.06	95.46		38.66	95.12

Note: Empirical assurance probability (EAP) is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

Table 4.5 a: Empirical assurance probabilities at the 80% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.9$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.9	0.05	0.5	0.5	162	63.63	94.80	180	81.68	94.82	218	98.39	95.15
			0.75		60.98	94.88	213	96.26	95.00		96.96	95.20
			1		46.96	94.87	252	98.17	95.11		88.81	94.86
		1	0.5	190	63.48	94.82	215	83.39	94.87	194	67.09	95.24
			0.75		83.81	95.03	233	98.98	95.53		86.29	95.31
			1		88.15	95.02	257	99.97	95.05		90.33	95.13
		1.5	0.5	226	63.67	94.76	258	83.80	94.75	202	47.15	94.77
			0.75		90.37	95.14	270	99.25	94.97		75.24	95.08
			1		97.46	94.71	289	100.00	95.60		89.85	95.06
		2	0.5	265	63.64	94.57	303	84.40	94.60	218	36.40	94.86
			0.75		92.87	95.25	311	99.31	94.74		69.17	94.76
			1		99.34	94.93	327	100.00	95.01		89.34	95.02
	0.1	0.5	0.5	50	63.28	95.23	56	80.55	95.18	73	98.83	95.40
			0.75		62.65	95.37	66	93.00	95.49		97.25	95.72
			1		55.00	95.21	78	95.57	94.76		91.04	94.84
		1	0.5	59	67.64	94.19	66	82.58	94.08	65	79.31	93.82
			0.75		82.02	95.05	72	97.13	95.19		90.92	95.12
			1		84.79	95.10	80	99.75	95.00		93.27	95.17
		1.5	0.5	71	69.94	93.34	80	84.44	94.02	68	61.45	92.91
			0.75		87.63	94.93	84	97.74	94.72		81.73	94.69
			1		95.00	94.96	90	99.85	95.18		91.19	95.01
		2	0.5	83	68.32	93.01	94	82.77	92.98	73	55.52	92.30
			0.75		89.23	93.81	97	97.52	94.13		78.09	93.68
			1		96.88	94.75	102	99.95	94.97		91.00	94.77

Note: Empirical assurance probability (EAP) is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. r is the ratio between the size of the typical and atypical populations, B is the ratio between their standard deviations, and N is the total sample size from the two populations.

Table 4.5 b: Empirical assurance probabilities at the 80% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.8$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.8	0.05	0.5	0.5	302	68.36	95.03	311	78.36	95.21	376	99.96	95.14
			0.75		50.27	95.04	362	96.62	95.11		99.02	95.37
			1		26.30	95.04	416	98.82	95.51		87.32	95.02
		1	0.5	321	42.09	94.96	361	80.76	95.30	335	55.65	94.78
			0.75		71.88	95.13	379	99.48	95.34		84.85	95.40
			1		79.30	95.61	404	100.00	95.11		89.95	95.32
		1.5	0.5	367	32.62	95.08	428	80.88	95.13	348	21.42	94.81
			0.75		78.86	95.10	431	99.68	95.32		63.22	94.78
			1		96.01	94.54	442	100.00	95.26		89.11	95.17
		2	0.5	420	28.13	94.80	501	82.69	95.09	376	11.27	94.80
			0.75		82.12	95.17	491	99.54	94.91		51.10	94.56
			1		98.90	94.72	493	100.00	95.29		86.68	95.15
	0.1	0.5	0.5	80	61.61	95.60	84	77.21	95.26	106	99.69	95.01
			0.75		52.38	95.45	97	88.87	95.66		97.04	94.87
			1		37.79	95.23	112	93.57	95.20		86.94	95.43
		1	0.5	86	49.56	95.02	97	76.41	95.30	95	71.40	94.89
			0.75		66.68	95.43	103	96.35	95.49		86.8	95.34
			1		69.86	95.12	110	99.83	95.57		89.97	95.36
		1.5	0.5	98	47.33	94.99	116	78.64	94.92	98	46.38	95.02
			0.75		74.87	95.20	117	96.26	95.35		74.32	95.35
			1		87.79	95.49	121	99.92	95.46		87.84	95.02
		2	0.5	113	44.19	95.28	135	78.42	94.86	106	36.37	95.25
			0.75		75.75	95.26	133	96.76	95.47		66.05	95.18
			1		93.94	95.04	135	99.87	95.33		87.13	95.43

Note: Empirical assurance probability (EAP) is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

Table 4.5 c: Empirical assurance probabilities at the 80% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.7$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.7	0.05	0.5	0.5	430	95.01	95.32	392	42.16	95.30	495	100.00	95.13
			0.75		61.04	95.15	450	86.81	94.74		99.93	94.73
			1		17.32	95.10	507	93.72	95.26		83.75	94.90
		1	0.5	427	20.44	95.33	445	41.30	95.20	440	33.88	95.33
			0.75		60.59	94.91	456	94.59	95.14		78.43	95.44
			1		72.21	95.03	472	99.88	95.03		87.53	95.16
		1.5	0.5	472	6.09	94.90	524	42.44	95.09	458	3.27	94.77
			0.75		57.71	94.55	511	95.13	94.56		41.35	94.91
			1		95.70	94.83	505	99.98	95.00		87.00	95.21
		2	0.5	530	3.17	95.19	609	43.61	94.79	495	0.78	95.21
			0.75		58.23	94.66	577	95.01	95.06		23.7	95.01
			1		98.74	94.80	554	99.90	94.72		83.6	95.05
	0.1	0.5	0.5	108	73.20	95.42	101	51.02	95.63	131	100.00	95.28
			0.75		48.46	95.48	116	76.24	95.50		98.39	95.86
			1		26.96	95.20	131	82.51	95.78		81.05	95.19
		1	0.5	109	36.26	94.77	115	53.69	95.24	116	54.30	95.55
			0.75		57.50	95.04	118	88.60	95.23		77.85	95.38
			1		64.32	95.17	123	97.94	95.24		84.06	95.33
		1.5	0.5	121	24.74	95.19	135	55.70	95.21	121	25.06	94.86
			0.75		60.65	95.14	132	85.35	95.50		60.01	95.43
			1		85.44	95.82	132	98.72	95.05		85.10	95.15
		2	0.5	136	20.58	95.46	157	55.21	95.39	131	15.21	94.91
			0.75		60.80	95.20	150	89.34	95.27		47.60	95.31
			1		91.66	95.76	145	98.24	95.22		82.03	95.21

Note: Empirical assurance probability (EAP) is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

Table 4.6 a: Empirical assurance probabilities at the 90% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.9$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.9	0.05	0.5	0.5	172	74.27	95.09	191	89.21	95.08	245	99.85	95.25
			0.75		72.13	95.21	227	98.26	95.03		99.54	95.36
			1		56.67	94.96	268	99.22	94.94		97.16	94.92
		1	0.5	203	74.60	94.73	229	90.71	94.99	218	84.51	95.08
			0.75		91.68	94.91	249	99.81	95.12		96.16	94.78
			1		94.72	95.33	275	100.00	95.39		98.18	95.00
		1.5	0.5	242	74.06	95.00	275	91.20	94.64	227	64.91	94.79
			0.75		95.28	94.79	288	99.89	94.95		89.94	94.78
			1		99.15	95.09	310	100.00	95.18		97.52	95.14
		2	0.5	284	73.64	95.09	323	90.50	94.47	245	51.84	95.17
			0.75		96.84	95.32	333	99.89	95.08		84.56	94.79
			1		99.87	94.90	351	100.00	95.48		97.14	95.02
	0.1	0.5	0.5	55	77.93	95.42	61	89.10	95.52	87	99.95	94.78
			0.75		75.15	95.33	73	97.21	95.55		99.84	95.01
			1		65.90	94.71	86	98.01	94.86		97.88	94.84
		1	0.5	66	79.93	94.22	73	90.81	94.37	77	94.12	94.43
			0.75		90.28	94.96	80	99.48	95.13		98.74	94.93
			1		93.22	95.30	89	99.98	95.54		99.47	95.36
		1.5	0.5	79	81.13	93.83	88	91.50	93.95	80	80.53	93.37
			0.75		94.62	94.59	93	99.49	94.98		95.02	94.85
			1		98.45	95.24	100	100.00	95.15		98.68	95.16
		2	0.5	92	79.80	93.15	104	91.41	93.72	87	72.40	92.72
			0.75		95.35	94.60	107	99.21	94.54		91.14	94.41
			1		99.33	94.76	114	100.00	94.81		97.96	94.84

Note: Empirical assurance probability (EAP) is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. r is the ratio between the size of the typical and atypical populations, B is the ratio between their standard deviations, and N is the total sample size from the two populations.

Table 4.6 b: Empirical assurance probabilities at the 90% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.8$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.8	0.05	0.5	0.5	309	75.58	95.01	322	87.65	94.70	401	100.00	95.01
			0.75		57.59	95.13	375	98.95	95.17		99.91	94.83
			1		30.57	95.01	432	99.58	94.82		96.63	95.07
		1	0.5	331	51.24	95.03	375	90.08	95.25	356	75.29	95.51
			0.75		81.51	95.05	395	99.92	94.84		95.31	95.31
			1		87.08	95.33	422	100.00	94.97		97.36	94.61
		1.5	0.5	380	42.23	95.31	445	90.17	94.78	371	35.91	95.32
			0.75		87.32	94.82	449	99.93	94.98		82.39	95.15
			1		98.59	95.11	462	100.00	94.99		97.21	95.24
		2	0.5	436	38.29	94.78	520	89.00	94.84	401	19.62	95.32
			0.75		90.73	94.89	512	99.95	94.98		69.60	95.04
			1		99.80	95.29	516	100.00	94.43		96.30	94.96
	0.1	0.5	0.5	83	71.65	95.58	89	85.81	95.61	118	100.00	95.28
			0.75		60.78	95.40	104	95.96	95.29		99.82	95.42
			1		44.10	95.39	120	97.88	94.94		96.85	95.06
		1	0.5	91	62.87	95.06	104	89.09	95.08	105	89.21	95.51
			0.75		80.16	95.54	110	99.27	95.03		97.34	95.04
			1		83.61	95.71	118	99.97	95.37		98.60	95.14
		1.5	0.5	105	56.02	95.01	124	88.97	95.59	109	65.33	95.22
			0.75		83.76	95.09	126	99.38	95.02		90.14	95.28
			1		94.52	95.53	130	99.99	95.52		97.58	95.23
		2	0.5	120	53.55	95.01	145	89.03	95.39	118	53.08	95.20
			0.75		85.27	95.07	144	99.38	95.16		84.69	95.31
			1		97.05	95.15	146	100.00	95.29		97.43	95.30

Note: Empirical assurance probability (EAP) is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

Table 4.6 c: Empirical assurance probabilities at the 90% assurance level for three variance formulas such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.7$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.7	0.05	0.5	0.5	434	96.83	95.11	401	57.21	95.20	512	100.00	94.50
			0.75		66.14	94.71	460	93.37	95.25		99.99	95.16
			1		18.82	94.93	519	97.42	95.29		94.78	95.29
		1	0.5	433	27.48	95.21	455	56.23	94.74	456	55.63	95.23
			0.75		70.50	94.68	468	98.86	95.03		94.02	94.98
			1		80.28	95.14	485	100.00	95.08		97.70	95.13
		1.5	0.5	481	9.49	95.43	536	56.46	95.46	475	7.00	95.17
			0.75		70.71	94.99	524	98.46	94.82		61.89	95.01
			1		98.40	95.21	519	100.00	94.95		96.72	95.04
		2	0.5	541	5.39	95.16	624	57.05	95.21	512	1.69	95.19
			0.75		70.46	94.66	592	98.27	94.64		39.89	94.95
			1		99.69	94.96	571	100.00	95.10		94.53	95.00
	0.1	0.5	0.5	110	82.99	94.99	105	69.36	95.52	139	100.00	95.32
			0.75		58.58	95.42	121	88.63	95.77		99.90	95.14
			1		32.22	95.29	137	91.92	95.21		94.83	95.43
		1	0.5	112	41.35	95.11	120	72.51	95.26	124	78.09	95.38
			0.75		65.18	95.34	124	97.54	94.95		95.78	95.64
			1		71.52	95.12	129	99.97	95.47		98.03	95.34
		1.5	0.5	125	28.88	95.47	141	69.11	95.12	129	38.74	95.57
			0.75		66.19	94.86	139	95.84	95.36		79.60	95.40
			1		90.90	95.43	139	99.92	95.18		96.38	95.32
		2	0.5	142	28.21	95.47	164	67.28	94.97	139	24.34	95.73
			0.75		75.00	94.92	157	95.30	94.89		67.35	95.22
			1		97.23	95.18	153	99.76	95.37		94.88	95.36

Note: Empirical assurance probability (EAP) is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

4.2.4 Test of robustness

A source of error for the simulation involving confidence interval width may have been due to use of the nonparametric method by DeLong et al. (1988). As using a nonparametric method on a normally distributed data set may cause the method to lose power, we conducted an alternate simulation study where instead of using the nonparametric method to calculate the variance of the AUC estimate of the generated data, the three parametric variance estimators that were used to estimate the sample size were used again to construct confidence intervals. To ensure consistency, the same variance estimator that was used to determine the sample size would also be the one that is used in the second step. This way, we can test the robustness of the proposed sample size formulas.

This simulation study used the same settings as the previous simulation study for confidence interval width, with the only difference being the variance estimator that was used to calculate the variance of the AUC estimate, $\hat{\theta}$. The results of this method are in summarized in Tables 4.7a to 4.9c. Based on these results, it is clear that using the same parametric variance formula to calculate the variances and confidence intervals of the generated data provides better empirical assurance probabilities. While the nonparametric formula resulted in many EAPs that were very high or very low, using the parametric formulas ensures that the EAPs are much more consistent, and they are generally close to the prespecified assurance probability. For instance, when $\theta = 0.7$, $\omega = 0.05$, $r = 1.5$, $B = 0.5$, using the probit based variance gives an EAP of 50.28 (Table 4.7c) compared to using the nonparametric method to calculate the variance which gives an EAP of 0.54 (Table 4.4c). There is still some overshooting of the EAPs when the binormal and probit based variance estimators are used, but these EAPs are much more consistent than when using the nonparametric formula. Overall, the results are better than when using the nonparametric method to calculate the variance of the AUC estimate. This aligns with our hypothesis that using a nonparametric method on normally distributed data may cause it to lose power, thus resulting in inconsistent EAPs.

Table 4.7 d: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 50% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.9$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.9	0.05	0.5	0.5	143	51.52	95.22	158	52.35	95.38	162	49.60	96.15
			0.75		51.41	94.98	186	53.17	96.98		49.69	95.96
			1		51.07	93.66	218	51.43	97.48		50.24	95.52
		1	0.5	164	50.97	94.85	187	51.47	95.66	144	48.88	92.96
			0.75		50.60	96.44	202	53.39	97.62		49.71	94.34
			1		51.77	96.94	221	51.71	98.40		50.42	95.74
		1.5	0.5	193	49.72	94.66	224	51.58	95.67	150	50.38	91.04
			0.75		50.24	96.66	233	52.62	98.17		49.83	93.63
			1		50.49	97.69	247	51.99	98.66		50.69	94.64
		2	0.5	226	51.59	94.38	263	50.71	95.95	162	51.50	89.27
			0.75		50.63	97.05	268	52.94	98.08		50.32	93.36
			1		50.48	98.04	279	52.00	98.83		50.96	95.10
	0.1	0.5	0.5	40	48.55	97.05	44	54.60	95.05	45	49.74	96.63
			0.75		50.00	97.33	52	56.24	96.89		48.99	97.04
			1		49.37	96.68	61	58.01	97.13		49.45	96.05
		1	0.5	46	51.68	96.85	52	55.78	95.54	40	49.48	94.49
			0.75		51.55	97.28	56	54.80	97.33		48.93	96.04
			1		51.52	97.64	62	57.21	98.21		48.76	96.23
		1.5	0.5	54	48.99	96.24	62	50.80	95.52	42	48.82	92.15
			0.75		48.90	97.70	65	55.52	97.68		49.21	95.59
			1		48.19	97.92	69	53.42	98.66		47.91	96.37
		2	0.5	63	53.05	96.07	73	54.02	95.58	45	51.67	91.24
			0.75		53.09	97.42	74	50.70	97.66		50.72	94.43
			1		52.72	98.18	78	56.85	98.72		49.70	96.86

Note: Empirical assurance probability (EAP) is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 10000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

Table 4.7 e: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 50% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.8$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.8	0.05	0.5	0.5	288	49.48	95.25	289	51.46	95.21	327	50.38	96.48
			0.75		50.17	95.10	335	52.45	96.26		49.87	95.64
			1		49.99	93.68	384	52.90	96.56		49.89	94.97
		1	0.5	301	48.73	94.00	334	51.39	95.36	291	50.18	93.82
			0.75		49.55	95.34	349	50.59	96.68		49.92	94.91
			1		48.89	95.67	370	52.56	97.07		50.48	95.25
		1.5	0.5	342	47.62	93.61	396	50.80	95.57	303	50.45	91.85
			0.75		48.04	95.48	396	52.21	96.57		50.65	94.37
			1		47.48	96.53	403	52.08	97.56		49.52	94.81
		2	0.5	390	50.72	93.23	462	51.52	95.51	327	49.94	90.84
			0.75		49.90	95.62	450	52.14	96.72		50.25	93.33
			1		49.65	96.86	448	51.31	97.77		50.25	94.93
	0.1	0.5	0.5	72	47.80	95.86	73	53.56	95.63	82	51.56	96.76
			0.75		48.61	95.47	84	53.22	96.02		51.66	95.77
			1		48.76	94.08	97	58.91	96.74		51.40	95.04
		1	0.5	76	49.03	94.48	84	53.27	95.55	73	48.36	93.39
			0.75		48.98	95.40	88	54.01	96.49		47.75	94.80
			1		51.38	95.65	93	53.02	97.04		48.67	94.97
		1.5	0.5	86	47.79	93.94	100	54.45	95.40	76	48.13	91.93
			0.75		47.77	95.66	99	49.65	96.95		48.58	93.92
			1		46.94	96.91	101	50.97	97.44		47.89	94.78
		2	0.5	98	46.56	93.55	116	49.12	95.32	82	48.40	90.61
			0.75		46.59	95.81	113	49.35	96.82		49.16	93.61
			1		46.35	97.27	113	52.12	97.53		48.54	95.20

Note: EAP is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. ECP is the coverage of the 95 percent CI based on a 10000 run simulation. r is the ratio between the size of the typical and atypical populations, B is the ratio between their standard deviations, and N is the total sample size from the two populations.

Table 4.7 f: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 50% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.7$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.7	0.05	0.5	0.5	421	50.26	95.84	376	52.91	94.05	460	50.40	96.65
			0.75		49.62	94.73	431	53.85	95.39		50.66	96.10
			1		50.23	94.17	483	52.06	95.59		51.47	95.02
		1	0.5	413	48.45	94.08	425	48.99	94.46	409	49.30	94.10
			0.75		49.68	95.04	434	51.86	95.37		49.40	95.01
			1		47.66	95.14	446	50.64	95.74		50.04	95.00
		1.5	0.5	455	49.78	93.29	500	53.22	94.11	426	50.28	91.74
			0.75		50.82	94.87	484	49.03	95.72		49.12	94.21
			1		49.90	95.86	476	53.35	96.58		49.12	94.81
		2	0.5	508	48.62	92.20	580	48.95	94.14	460	48.73	91.46
			0.75		47.53	94.85	546	52.89	95.93		49.28	93.30
			1		47.74	96.05	521	51.30	96.53		50.19	94.92
	0.1	0.5	0.5	104	49.46	95.81	93	54.32	94.22	114	49.59	96.36
			0.75		50.06	94.97	106	56.24	95.45		49.85	96.05
			1		50.44	94.08	119	56.10	95.84		49.95	94.94
		1	0.5	102	49.26	93.99	105	50.02	94.42	101	49.36	93.89
			0.75		49.06	95.14	107	50.9	95.46		48.63	94.99
			1		48.35	95.52	110	52.77	96.02		48.83	94.87
		1.5	0.5	112	45.19	93.46	123	49.37	94.57	105	48.50	92.60
			0.75		43.75	94.89	120	57.04	95.36		49.22	94.22
			1		44.37	96.02	117	49.28	96.10		49.09	94.84
		2	0.5	126	50.18	92.96	143	46.10	94.16	114	50.38	91.37
			0.75		50.06	94.77	135	58.19	95.87		50.33	93.32
			1		51.59	96.51	129	61.65	96.46		50.44	95.11

Note: EAP is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. ECP is the coverage of the 95 percent CI based on a 10000 run simulation. r is the ratio between the size of the typical and atypical populations, B is the ratio between their standard deviations, and N is the total sample size from the two populations.

Table 4.8 d: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 80% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.9$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.9	0.05	0.5	0.5	162	71.23	95.04	180	76.02	95.39	218	93.69	96.17
			0.75		71.04	94.67	213	79.22	97.13		93.53	96.18
			1		70.26	94.18	252	81.68	97.66		92.55	95.07
		1	0.5	190	77.98	94.93	215	77.64	95.25	194	90.71	93.13
			0.75		79.12	96.17	233	82.48	97.52		92.14	94.60
			1		79.41	96.80	257	85.79	98.30		92.81	94.91
		1.5	0.5	226	80.04	94.63	258	77.56	95.65	202	88.61	90.85
			0.75		82.08	96.58	270	83.30	97.85		91.53	94.17
			1		82.85	97.58	289	87.95	98.71		92.38	95.23
		2	0.5	265	81.21	94.79	303	78.86	96.10	218	87.66	89.14
			0.75		84.13	96.95	311	83.79	98.00		90.97	93.40
			1		86.73	98.19	327	89.71	98.75		92.38	95.12
	0.1	0.5	0.5	50	76.36	97.41	56	83.53	95.20	73	96.88	95.88
			0.75		76.60	97.14	66	86.32	96.86		97.02	96.35
			1		74.75	96.22	78	88.62	97.51		96.40	95.52
		1	0.5	59	81.93	96.76	66	85.07	95.59	65	94.47	93.21
			0.75		83.48	97.37	72	89.53	97.67		95.75	94.69
			1		83.77	97.56	80	93.50	98.13		95.81	95.04
		1.5	0.5	71	86.66	96.38	80	86.83	95.97	68	94.16	91.36
			0.75		88.21	97.38	84	90.73	98.00		95.33	94.27
			1		89.57	97.83	90	95.75	98.49		96.12	95.51
		2	0.5	83	86.68	96.15	94	85.44	95.72	73	92.54	89.21
			0.75		89.36	97.27	97	91.90	98.12		94.39	93.08
			1		90.95	98.35	102	96.78	98.71		95.39	95.22

Note: EAP is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. ECP is the coverage of the 95 percent CI based on a 10000 run simulation. r is the ratio between the size of the typical and atypical populations, B is the ratio between their standard deviations, and N is the total sample size from the two populations.

Table 4.8 e: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 80% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.8$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.8	0.05	0.5	0.5	302	64.86	95.36	311	80.81	94.91	376	92.96	96.20
			0.75		65.37	94.75	362	84.07	96.08		93.35	96.12
			1		64.96	93.89	416	87.45	96.51		92.02	95.18
		1	0.5	321	71.67	93.92	361	84.58	95.35	335	90.02	93.57
			0.75		72.99	95.35	379	89.57	96.95		91.67	95.17
			1		73.20	95.61	404	95.29	97.27		91.51	94.86
		1.5	0.5	367	75.29	93.34	428	85.00	94.95	348	88.48	91.51
			0.75		77.21	95.63	431	91.50	96.93		90.09	94.13
			1		77.65	96.12	442	97.15	97.57		91.37	95.03
		2	0.5	420	79.25	93.58	501	86.05	95.36	376	87.16	90.73
			0.75		80.73	95.48	491	91.16	96.76		89.73	93.33
			1		82.56	96.93	493	97.87	97.77		91.06	95.20
	0.1	0.5	0.5	80	67.34	95.27	84	85.99	95.22	106	95.74	96.60
			0.75		66.41	95.05	97	87.62	96.11		94.90	95.90
			1		66.32	94.07	112	91.58	96.48		93.98	94.77
		1	0.5	86	74.42	94.53	97	87.53	95.53	95	93.32	94.09
			0.75		75.57	95.72	103	94.42	96.43		93.77	94.79
			1		76.86	95.91	110	99.11	97.32		93.76	95.09
		1.5	0.5	98	77.57	94.03	116	89.85	95.20	98	90.79	91.73
			0.75		78.94	95.94	117	94.91	96.78		92.16	93.86
			1		80.36	97.02	121	99.70	97.55		92.81	95.19
		2	0.5	113	79.34	93.69	135	90.26	94.99	106	89.44	90.71
			0.75		81.16	96.06	133	95.83	97.04		91.43	93.60
			1		83.33	97.24	135	99.71	97.69		92.87	95.32

Note: EAP is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. ECP is the coverage of the 95 percent CI based on a 10000 run simulation. r is the ratio between the size of the typical and atypical populations, B is the ratio between their standard deviations, and N is the total sample size from the two populations.

Table 4.5 f: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 80% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.7$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.7	0.05	0.5	0.5	430	61.99	96.07	392	84.46	94.81	495	93.11	96.67
			0.75		61.61	95.39	450	86.09	95.50		92.18	95.75
			1		61.02	94.07	507	91.02	95.67		91.26	94.89
		1	0.5	427	70.79	94.02	445	89.75	94.22	440	90.20	94.03
			0.75		71.70	94.92	456	96.16	95.25		91.10	95.05
			1		73.01	95.37	472	99.83	95.80		91.02	94.98
		1.5	0.5	472	75.81	93.00	524	91.28	94.58	458	87.98	92.06
			0.75		77.58	94.81	511	97.48	95.33		90.19	93.83
			1		78.21	96.02	505	99.91	96.08		90.51	94.92
		2	0.5	530	81.01	92.62	609	91.69	94.20	495	88.69	91.39
			0.75		82.41	94.42	577	97.20	95.73		89.79	93.68
			1		83.69	95.98	554	99.60	96.01		91.27	94.63
	0.1	0.5	0.5	108	60.03	95.58	101	88.70	94.81	131	95.54	96.36
			0.75		60.44	95.18	116	90.62	95.25		95.91	96.17
			1		59.03	93.58	131	94.36	95.74		95.38	95.69
		1	0.5	109	71.24	93.81	115	94.30	94.17	116	92.66	93.65
			0.75		72.79	94.67	118	99.84	95.70		93.55	95.22
			1		72.43	95.42	123	100.00	96.00		93.43	95.41
		1.5	0.5	121	77.96	93.00	135	96.16	94.45	121	91.17	92.09
			0.75		80.02	95.31	132	98.47	96.18		91.96	93.87
			1		80.87	95.81	132	99.99	96.19		92.98	95.07
		2	0.5	136	82.86	92.57	157	94.31	94.74	131	90.00	91.62
			0.75		85.32	95.26	150	99.18	95.46		91.28	93.56
			1		86.65	95.90	145	99.89	96.48		92.96	95.14

Note: EAP is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. ECP is the coverage of the 95 percent CI based on a 10000 run simulation. r is the ratio between the size of the typical and atypical populations, B is the ratio between their standard deviations, and N is the total sample size from the two populations.

Table 4.9 d: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 90% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.9$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.9	0.05	0.5	0.5	172	80.75	94.92	191	85.25	95.58	245	98.92	95.79
			0.75		80.60	95.01	227	88.14	96.76		99.21	96.29
			1		80.03	94.55	268	90.35	97.68		98.50	94.89
		1	0.5	203	87.51	94.62	229	87.86	95.91	218	97.98	92.79
			0.75		88.92	96.38	249	91.55	97.69		98.23	94.51
			1		89.54	96.81	275	94.70	98.30		98.61	94.89
		1.5	0.5	242	88.73	94.53	275	88.55	95.92	227	97.27	90.93
			0.75		91.69	96.91	288	92.47	97.83		98.15	93.95
			1		92.31	97.59	310	96.71	98.90		98.36	94.87
		2	0.5	284	90.61	94.46	323	87.95	95.92	245	96.45	89.63
			0.75		93.70	97.14	333	93.50	97.84		97.87	92.85
			1		94.38	98.10	351	97.57	99.02		98.54	95.12
	0.1	0.5	0.5	55	85.44	97.07	61	91.78	95.41	87	99.77	96.30
			0.75		85.58	97.08	73	94.93	96.74		99.78	95.90
			1		84.34	96.42	86	96.58	97.75		99.65	95.10
		1	0.5	66	93.13	96.63	73	93.30	95.94	77	99.40	92.57
			0.75		94.68	97.47	80	97.46	97.69		99.73	94.87
			1		94.86	97.55	89	98.98	98.18		99.81	95.38
		1.5	0.5	79	94.58	96.35	88	94.09	95.47	80	98.90	90.61
			0.75		96.49	97.70	93	98.24	97.75		99.42	93.84
			1		96.83	97.99	100	99.70	98.70		99.58	95.19
		2	0.5	92	95.04	95.98	104	94.11	95.94	87	99.17	89.90
			0.75		97.62	97.98	107	97.96	97.95		99.58	93.45
			1		97.89	98.39	114	99.83	98.87		99.69	95.39

Note: EAP is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. ECP is the coverage of the 95 percent CI based on a 10000 run simulation. r is the ratio between the size of the typical and atypical populations, B is the ratio between their standard deviations, and N is the total sample size from the two populations.

Table 4.9 e: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 90% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.8$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.8	0.05	0.5	0.5	309	72.57	95.24	322	90.54	95.25	401	99.13	96.82
			0.75		71.48	94.75	375	92.60	96.08		98.70	95.95
			1		71.52	93.65	432	95.39	96.53		98.62	95.25
		1	0.5	331	81.58	94.47	375	94.39	95.56	356	97.66	93.71
			0.75		82.95	95.19	395	97.99	96.78		98.10	94.64
			1		82.82	95.62	422	99.79	97.29		98.51	94.74
		1.5	0.5	380	87.15	93.45	445	94.47	95.03	371	96.85	91.69
			0.75		88.84	95.31	449	98.50	96.52		97.70	93.56
			1		90.06	96.57	462	99.91	97.48		98.20	95.05
		2	0.5	436	89.31	93.27	520	94.78	95.12	401	96.13	89.72
			0.75		91.78	95.73	512	98.84	97.08		97.60	93.34
			1		92.89	96.94	516	99.96	97.82		98.29	94.88
	0.1	0.5	0.5	83	73.80	95.47	89	94.19	94.85	118	99.70	96.35
			0.75		74.17	95.11	104	96.39	95.87		99.62	96.21
			1		72.75	93.70	120	98.32	96.65		99.56	95.33
		1	0.5	91	84.53	94.40	104	97.83	95.06	105	98.90	93.53
			0.75		85.60	95.42	110	99.74	97.12		99.35	95.17
			1		86.05	95.76	118	100.00	97.24		99.34	94.85
		1.5	0.5	105	91.02	93.86	124	97.97	95.51	109	98.54	92.12
			0.75		92.27	95.77	126	99.89	96.91		99.14	94.05
			1		92.83	96.87	130	100.00	97.72		99.27	95.23
		2	0.5	120	92.53	93.79	145	97.89	95.00	118	98.33	90.90
			0.75		94.22	95.94	144	99.88	96.95		99.14	93.82
			1		95.23	97.10	146	100.00	97.92		99.27	95.19

Note: EAP is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. ECP is the coverage of the 95 percent CI based on a 10000 run simulation. r is the ratio between the size of the typical and atypical populations, B is the ratio between their standard deviations, and N is the total sample size from the two populations.

Table 4.9 f: Comparisons of empirical assurance probabilities using the nonparametric (NP) and parametric methods at the 90% assurance level such that the half-width of a two-sided 95% confidence interval for the AUC is not greater than the prespecified width ω when the true AUC $\theta = 0.7$.

				Exponential Based Variance			Binormal Based Variance			Probit Based Variance		
θ	ω	r	B	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.7	0.05	0.5	0.5	434	66.56	95.51	401	93.94	94.35	512	98.89	96.52
			0.75		66.00	95.26	460	95.44	95.27		98.83	96.04
			1		66.22	93.87	519	97.78	95.20		98.52	95.17
		1	0.5	433	79.80	94.13	455	98.26	94.74	456	98.21	93.86
			0.75		79.60	95.15	468	99.88	95.33		98.64	94.77
			1		80.09	95.18	485	100.00	96.18		98.75	95.01
		1.5	0.5	481	88.02	93.51	536	98.31	94.16	475	97.63	92.11
			0.75		88.70	94.41	524	99.94	95.60		98.39	94.27
			1		90.28	95.91	519	100.00	95.96		98.62	95.29
		2	0.5	541	92.88	92.79	624	99.00	94.61	512	96.50	91.02
			0.75		94.59	94.98	592	99.83	95.45		97.80	93.70
			1		95.08	95.88	571	99.99	96.64		98.22	94.96
	0.1	0.5	0.5	110	66.10	95.88	105	96.55	94.45	139	99.98	96.68
			0.75		65.05	94.99	121	97.72	95.11		99.95	95.71
			1		65.21	93.93	137	99.21	95.73		99.94	94.81
		1	0.5	112	82.15	93.98	120	100.00	94.68	124	99.81	93.66
			0.75		82.32	95.19	124	100.00	95.47		99.81	94.66
			1		83.22	95.20	129	100.00	96.08		99.79	95.18
		1.5	0.5	125	90.95	92.91	141	99.92	94.54	129	99.45	91.92
			0.75		92.38	94.99	139	100.00	95.39		99.60	94.00
			1		93.85	95.96	139	100.00	96.03		99.79	94.93
		2	0.5	142	97.78	92.37	164	99.58	94.57	139	99.12	91.00
			0.75		98.58	95.28	157	99.99	95.57		99.62	93.95
			1		99.17	96.35	153	100.00	96.39		99.75	95.29

Note: EAP is the frequency of times the half-width of the 95 percent CI is smaller the prespecified width ω. ECP is the coverage of the 95 percent CI based on a 10000 run simulation. r is the ratio between the size of the typical and atypical populations, B is the ratio between their standard deviations, and N is the total sample size from the two populations.

4.3 Nonparametric method using pilot data

Similar to the first simulation study that was designed to estimate sample sizes with precision when achieving a prespecified lower bound, we conducted a third simulation study using the method based on pilot data. This method used pilot data from Wieand et al. (1989), which is a data set taken from the Mayo clinic where a total of 141 patients were evaluated for their disease status and levels of CA 19-9 and CA 125 markers, two biological markers that are often associated with pancreatic cancer. Higher values for biomarkers indicate higher probability of the disease.

We first estimated the AUC and variance of AUC for these two markers using the ranking method explained in Section 3.3.3. These values were then used in the proposed sample size formula to estimate required sample sizes based on the group size ratio *r*, the prespecified assurance probability $1-\beta$, and the lower bound θ_0. Data was generated based on this sample size by sampling the pilot data with replacement, and then a ROC test was performed in order to estimate the AUC, its variance, and confidence intervals. The simulation study was performed for 1000 runs with similar settings as used in the previous simulation studies.

We kept the range of parameters similar to the first simulation to ensure consistency. The ratio of typical to atypical patients, *r*, was set to be 0.5 to 2. The lower bound was slightly different from the first simulation study because the AUC is fixed in this simulation, so we decided to use various lower bounds that are within the applicable range of an AUC. We used lower bounds between 0.55 and 0.65 for the simulation study based on the CA 19-9 biomarker, which has an AUC of 0.70, and we used lower bounds between 0.7 and 0.8 for the simulation study based on the CA 125 biomarker, which has an AUC of 0.86. We used a prespecified assurance probability of 0.5, 0.8, and 0.9.

4.3.1 Results

The simulation results are summarized in Tables 4.10a and 4.10b, respectively for the simulation based on the CA 19-9 biomarker which has an AUC of 0.70, and the simulation based on the CA 125 biomarker whose AUC is 0.86. The nonparametric method performed better than the previous methods. The general trends observed from the simulation study for the parametric methods can also be observed here.

As the prespecified assurance probability increases, the estimated sample size also increases, as seen in Table 4.7a, and as the lower bound decreases, the sample size also decreases, which matches what is expected as a bigger difference between the AUC and lower bound leads to a smaller required sample size. The sample size also changes according to the group size ratio r, where the smallest sample size estimates occur when $r = 1$ and $r = 1.5$, but the largest sample sizes occur when $r = 0.5$ and $r = 2$. This means that fewer subjects are needed for studies that have fairly even study group sizes as compared to very uneven or skewed study group sizes.

With the true AUC of 0.70, the empirical assurance probabilities are generally very close to the prespecified assurance probabilities. When the prespecified assurance probability is 50 percent, the EAPs are within the range of 47 to 50 percent, with the majority of them being around 49. When the prespecified assurance probabilities are 80 and 90 percent, the EAPs tend to overshoot more such that as the prespecified assurance increases, the EAPs are in the range of 80 to 83 percent, and between 91 and 93 percent, respectively.

The second part of this simulation was based on the CA 125 biomarker, which has an AUC of 0.86. The same trends could be seen here with some slight differences. We can see that the sample sizes in Table 4.7b are similar to those estimated from data based on the CA 19-9 biomarker in Table 4.7a. However, as group size ratio r increases, the sample size also increases. For the

previous simulation, the sample sizes were highest when *r* was farthest away from $r = 1$, implying that fewer subjects would be needed in a study if the study group sizes are fairly even. This trend is not seen for this simulation, as it appears that as the ratio *r* increases, more subjects would be required, with the smallest sample size resulting from when $r = 0.5$.

The empirical assurance probabilities are also within a greater range than that seen in the previous simulation. When the prespecified assurance probability is 50 percent, the EAPs are generally close to 50 percent, in the range of 49 to 53. When the prespecified assurance is 80 percent, the EAPs are ranged between 80.6 and 85.3 percent. EAPs tend to slightly overshoot when the lower bound is 0.7. This pattern also appears at the 90 percent prespecified assurance level, where EAPs are ranged between 91.3 to 95.2 percent. The largest EAPs around 94 to 95 percent occur when the lower bound is 0.7.

Table 4.10 a: Empirical assurance probabilities at three assurance levels such that the lower bound of a two-sided 95% confidence interval for the AUC is not greater than the prespecified lower limit θ_0. Based on the AUC obtained from the CA 19-9 data by Wieand et al.

			50% Assurance			80% Assurance			90% Assurance		
θ	θ_0	r	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.7	0.65	0.5	443	48.4	95.9	905	82.3	96.7	1211	92.2	96.0
	0.60		131	47.0	95.1	268	83.2	96.0	359	92.2	96.0
	0.55		64	50.7	93.6	130	82.1	95.3	174	91.4	95.9
	0.65	1	382	49.8	95.2	779	80.7	97.0	1043	91.0	96.6
	0.60		113	49.0	95.6	231	82.4	96.4	309	91.5	97.0
	0.55		55	48.6	92.6	112	81.1	95.8	150	91.3	96.2
	0.65	1.5	390	49.0	95.4	796	80.7	96.4	1066	91.6	94.9
	0.60		116	49.6	95.7	236	81.8	96.4	316	92.1	95.7
	0.55		56	48.9	92.6	115	81.7	95.5	153	92.0	96.0
	0.65	2	416	48.9	96.1	849	80.0	95.9	1136	91.8	96.0
	0.60		123	49.3	94.7	252	81.0	96.5	337	93.0	94.7
	0.55		60	48.0	91.8	122	81.7	94.8	163	91.4	95.0

Note: Empirical assurance probability (EAP) is the frequency of times the lower bound of the 95 percent CI is smaller the prespecified lower limit θ_0. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 1000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

Table 4.10 b: Empirical assurance probabilities at three assurance levels such that the lower bound of a two-sided 95% confidence interval for the AUC is not greater than the prespecified lower limit θ_0. Based on the AUC obtained from the CA125 data by Wieand et al.

			50% Assurance			80% Assurance			90% Assurance		
θ	θ_0	r	N	EAP	ECP	N	EAP	ECP	N	EAP	ECP
0.86	0.80	0.5	180	51.9	94.3	367	81.4	93.8	492	91.3	94.8
	0.75		66	50.9	94.9	135	84.0	94.4	180	93.8	94.3
	0.70		37	49.5	94.9	75	84.4	95.0	100	94.2	95.1
	0.80	1	215	50.1	94.6	439	80.6	92.7	587	91.6	94.2
	0.75		79	50.8	94.9	161	84.4	94.9	215	94.2	94.6
	0.70		44	50.8	92.3	89	85.1	95.9	119	94.9	95.6
	0.80	1.5	258	51.2	94.4	527	81.0	94.2	705	91.8	94.0
	0.75		95	51.3	94.5	193	84.3	94.6	259	94.2	94.4
	0.70		53	52.7	93.2	107	85.5	95.0	143	95.2	94.7
	0.80	2	303	50.4	94.5	619	81.6	93.6	829	91.3	94.2
	0.75		111	51.6	94.5	227	84.5	94.9	304	93.6	94.3
	0.70		62	52.3	92.6	126	85.3	94.4	168	94.4	94.2

Note: Empirical assurance probability (EAP) is the frequency of times the lower bound of the 95 percent CI is smaller the prespecified lower limit θ_0. Empirical coverage probability (ECP) is the coverage of the 95 percent CI based on a 1000 run simulation. *r* is the ratio between the size of the typical and atypical populations, *B* is the ratio between their standard deviations, and *N* is the total sample size from the two populations.

4.4 Conclusion

In this simulation study, we investigated the performance of three sample size formulas with an incorporated prespecified assurance probability of achieving a certain lower limit or confidence interval half-width. The sample size calculations required variances, and we used the variance estimators based on the exponential distribution, the binormal distribution, and the probit transformation, proposed by Hanley and McNeil (1982), Obuchowski (1994), and Rosner and Glynn (2009), respectively. The required sample sizes were then calculated, and data was generated according to these sample sizes and desired AUC. The AUC estimate and its variance were then calculated using the nonparametric method by DeLong et al. (1988), though three parametric variance formulas were used to calculate these in the test of robustness in Section 4.2.4. Confidence intervals for the AUC were constructed with this variance using the logit transformation to ensure that the confidence intervals did not go outside the bounds of (0,1). For the simulation study about achieving a prespecified lower bound, the empirical assurance probability was calculated by determining how many lower bounds of the 95 percent two-sided intervals were greater than the prespecified lower bound, θ_0. For the confidence interval half-width simulation study, the EAP was calculated based on how many half-widths of two-sided 95 percent confidence intervals were smaller than or equal to the prespecified half-width, ω. These EAPs were then compared to the prespecified assurance probability at three levels: 50 percent, 80 percent, and 90 percent.

We were able to see a few common trends within both simulations. The ratio between typical and atypical patients, r seemed to be directly proportional to the empirical assurance probability and sample size. The level of prespecified assurance is directly proportional to sample size, where the greater the prespecified assurance, the greater the required sample size. The

difference between θ and θ_0 is inversely proportional to the sample size, where the greater the difference between them, the smaller the required sample size. This is also true of ω where the required sample size is smaller when ω is bigger.

For the case of achieving a prespecified lower bound, we can see that the variance formulas perform quite similarly. The EAPs are generally close to the prespecified assurance probability, though sometimes they may overshoot especially when using the binormal variance estimator. For the case of achieving a certain confidence interval width, the results were unexpected as there were dramatic drops and peaks in the EAP. We suspect that it may be due to the small standard deviation ratio between the study groups of $B = 0.5$ paired with the extreme group size ratio of $r = 0.5$. This may have caused extreme variability amongst the atypical group, thus causing the confidence intervals to be much wider than expected. Additionally, the usage of a non-parametric method on normally distributed data may have compromised the power of the method, thus making the confidence intervals wider than expected.

Overall, the trends in the data match with the results of previous literature and what one would expect. The method based on pilot data performed the best since its empirical assurance probabilities were the closest to the prespecified assurance probabilities out of all of the methods. They were the most consistent and did not contain the drops and peaks seen in the confidence interval width simulation and they were also within a narrower range than those in the lower bound simulation.

Chapter 5 Illustration

After examining the finite sample size performance via simulations in Chapter 4, we now illustrate the application using examples from studies of biomarkers for pancreatic cancer.

Pancreatic cancer is often associated with two biomarkers: CA 125 and CA 19-9. CA 125 is known as cancer antigen 125, a protein that is associated with many types of cancer including ovarian, breast, stomach, and lung cancers. It can also be found in normal cells in the body and is secreted into the bloodstream, thus can be measured (Bast et al., 1983). Carbohydrate antigen 19-9 is a protein found on the surface of cancer cells and is often associated with pancreatic cancer, though can also exist in healthy individuals as well (Del Villano et al., 1983). It is also measured from the bloodstream.

These biomarkers are strongly correlated with having pancreatic cancer but having high levels of these biomarkers do not guarantee that someone will have pancreatic cancer. Wieand et al. (1989) evaluated the sensitivity and specificity of CA 125 (Bast et al., 1983) and CA 19-9 (Del Villano et al., 1983) markers based on a study done at the Mayo Clinic. In 141 patients, 50 had pancreatitis and 91 had pancreatic cancer. Their disease status was recorded as well as the levels of the two biomarkers in units U/mL. Wieand et al. (1989) then used the data from the study to compare the two biomarkers and evaluate how suitable they were for the diagnosis of pancreatic cancer, based on the ROC curves. For this illustration, we used the CA 125 biomarker, which has an AUC of 0.86, to investigate how the position of a lower bound θ_0 would affect the required sample size. Specifically, we want to determine a sample size that would ensure that the lower

bound of 95 percent confidence interval around an AUC of 0.86 is not lower than a preset lower limit of 0.8 or 0.75 or 0.7, with 80 percent assurance.

The first step would be to determine the variances that are needed in the proposed sample size formula (3.6):

$$N = \frac{(r+1)}{r}\left[\frac{Z_{\alpha/2} + Z_{\beta}}{\text{lgt}\,\theta \;-\; \text{lgt}\,\theta_0}\right]^2 \left[\frac{r\,\text{var}(p_j) + \text{var}(q_i)}{\theta^2\,(1-\theta)^2}\right].$$

Recall that the variance formulas for $\text{var}(p_j)$and $\text{var}(q_i)$ are:

$$\text{var}(p_j) = \frac{\sum_{j=1}^{n}\left(p_j - \bar{p}\right)^2}{n-1}$$

$$\text{var}(q_i) = \frac{\sum_{i=1}^{m}(q_i - \bar{q})^2}{m-1}.$$

The rank procedure from Section 3.3.3 is performed in order to determine the proportion of times a biomarker value from the atypical group is greater than biomarker values from the typical group, and vice versa. For the CA 125 data, we get $\text{var}(p_j) = 0.068325$ and $\text{var}(q_i) = 0.009001743$. Assuming an equal typical to atypical group ratio, $r = 1$, the sample size formula gives us

$$N = \frac{(r+1)}{r}\left[\frac{Z_{\alpha/2} + Z_{\beta}}{\text{lgt}\,\theta \;-\; \text{lgt}\,\theta_0}\right]^2 \left[\frac{r\,\text{var}(p_j) + \text{var}(q_i)}{\theta^2\,(1-\theta)^2}\right]$$

$$= \frac{(1+1)}{1}\left[\frac{1.96 + 0.842}{\text{lgt}\,(0.86) - \text{lgt}\,(0.8)}\right]^2 \left[\frac{(1)0.068325 + 0.009001743}{0.86^2\,(1-0.86)^2}\right]$$

$$= 2\left[\frac{1.96 + 0.842}{\ln\left(\frac{0.86}{1-0.86}\right) - \ln\left(\frac{0.8}{1-0.8}\right)}\right]^2 \left[\frac{0.068325 + 0.009001743}{0.86^2\,(1-0.86)^2}\right]$$

$$\approx 439\,.$$

Then a total sample size of 439 is required, with 220 subjects in the atypical group and 219 subjects in the typical group, to achieve a lower bound of 0.8 with 80 percent assurance.

To achieve a lower bound of 0.75 with 80 percent assurance, the sample size required is obtained using the same variances calculated from the pilot data set by the sample size formula:

$$N = \frac{(r+1)}{r}\left[\frac{Z_{\alpha/2} + Z_{\beta}}{\text{lgt}\,\theta\ -\ \text{lgt}\,\theta_0}\right]^2 \left[\frac{r\ \text{var}(p_j) + \text{var}(q_i)}{\theta^2\,(1-\theta)^2}\right]$$

$$= \frac{(1+1)}{1}\left[\frac{1.96 + 0.842}{\ln\left(\frac{0.86}{1-0.86}\right) - \ln\left(\frac{0.75}{1-0.75}\right)}\right]^2 \left[\frac{(1)0.068325 + 0.009001743}{0.86^2\,(1-0.86)^2}\right]$$

$$\approx 161$$

which gives us a total sample size of 161 subjects, with 81 in the atypical group and 80 in the typical group.

Similarly, to achieve a lower bound of 0.7, we have

$$N = \frac{(r+1)}{r}\left[\frac{Z_{\alpha/2} + Z_{\beta}}{\text{lgt}\,\theta\ -\ \text{lgt}\,\theta_0}\right]^2 \left[\frac{r\ \text{var}(p_j) + \text{var}(q_i)}{\theta^2\,(1-\theta)^2}\right]$$

$$= \frac{(1+1)}{1}\left[\frac{1.96 + 0.842}{\ln\left(\frac{0.86}{1-0.86}\right) - \ln\left(\frac{0.7}{1-0.7}\right)}\right]^2 \left[\frac{(1)0.068325 + 0.009001743}{0.86^2\,(1-0.86)^2}\right]$$

$$\approx 89$$

which gives us a total sample size of 89 subjects, with 45 in the atypical group and 44 in the typical group.

We can also calculate the required sample size when the group size ratio is not equal. When $r = 1.5$, assurance is 80 percent, and the lower bound θ_0 is 0.8, we have

$$N = \frac{(r+1)}{r}\left[\frac{Z_{\alpha/2} + Z_{\beta}}{\text{lgt}\,\theta\ -\ \text{lgt}\,\theta_0}\right]^2 \left[\frac{r\ \text{var}(p_j) + \text{var}(q_i)}{\theta^2\,(1-\theta)^2}\right]$$

$$= \frac{(1.5+1)}{1.5}\left[\frac{1.96 + 0.842}{\ln\left(\frac{0.86}{1-0.86}\right) - \ln\left(\frac{0.8}{1-0.8}\right)}\right]^2 \left[\frac{(1)0.068325 + 0.009001743}{0.86^2\,(1-0.86)^2}\right]$$

≈ 527 .

Thus, we would need a total sample size of 527, with 211 subjects in the atypical group and 316 subjects in the typical group.

Chapter 6 Discussion

Existing sample size formulas based on confidence interval width can provide inadequate sample sizes. This is because they do not take into consideration the probability of achieving the desired confidence interval, and usually there is only a 50 percent chance that the observed confidence interval excludes a prespecified value to be excluded. In order to ensure adequate sample sizes for studies assessing accuracy of diagnostic tests, we can incorporate into the sample size formula a prespecified assurance probability of achieving a desired lower limit or confidence interval width.

In this thesis, we derived asymptotic sample size formulas which incorporated a prespecified assurance probability to ensure that sample sizes are not underestimated. These were based on three different formulas for estimating the variance of the AUC. Sample sizes were determined based on the AUC θ, its variance, the ratio between the typical and atypical group sizes r, the ratio between their standard deviations B, and either the lower bound θ_0 or the confidence interval half-width ω. We evaluated the performance of the proposed sample size formulas via simulation studies. Data were then simulated 10000 times based on these calculated sample sizes, and then the means, variances, and AUCs of each data set were determined using the nonparametric method by DeLong et al. (1988). The logit transformation was applied to ensure that confidence intervals were not outside the plausible range of (0,1), then later transformed back into the raw scale in order to be compared. The empirical assurance probabilities were determined based on how many confidence interval lower bounds were greater than θ_0, or how many confidence interval half-widths were smaller than ω. These EAPs were compared to the prespecified assurance probability to evaluate the performance of the

sample size formulas using each of the three variance estimators. We went on to evaluate the robustness of the sample size formula by using these same parametric variance estimators on the generated data for achieving confidence interval width. A new method was developed based on pilot data and DeLong et al.'s (1988) nonparametric method, and a simulation study was also conducted in a similar way to the previous simulation studies. We also applied this new method to a data set in the real world to illustrate its impact on the sample size.

From the results of the simulation study, we concluded that the method based on pilot data performs the best. The empirical assurance probabilities of this method were the closest to the prespecified assurance probabilities, and they were the most consistent. While there is some overshooting, the EAPs do not have dramatic drops and spikes like in the simulation for achieving a prespecified confidence interval width, and they are within a narrower range than the EAPs in the simulation for achieving a prespecified lower bound.

There were certain patterns found in the results of the simulation study. Generally, we found that the sample size is inversely proportional to the ratio between typical and atypical people, where the greater the r, the smaller the required sample size. Sample size also depends on the magnitude of the difference between θ and θ_0 or of the confidence interval width ω, where the greater the difference or width, the smaller the sample size. In general, sample size and EAP are directly proportional, with a larger sample size giving the experiment more assurance. We can also see that EAP tends to increase as B increases, with its highest usually when $B = 1$.

The results do follow previous findings in the existing literature. The binormal based variance formula was known to be more conservative than the exponential based variance, especially when θ is large, and we can see that reflected in the results of this simulation. The sample sizes calculated with the binormal based variance are indeed larger than the other two

variances when θ is large. Additionally, we did comparisons at three different assurance levels—50 percent, 80 percent, and 90 percent. The 50 percent assurance level condition was to act as a control because it is the level of assurance in traditional confidence interval based estimation. As we can see, the sample sizes are indeed larger for the 80 and 90 percent assurance levels than the 50 percent assurance level, indicating that a greater sample size is needed for more precision.

Since the results from the confidence interval width simulation study were not as expected, we suspected that this was due to the standard deviation ratio and group size ratios being too extreme. When a large standard deviation is paired with a small group size for the atypical group, this would create very wide confidence intervals, thus leading to very small EAPs. Another issue would be the usage of a nonparametric method on normally distributed data, as the method may be ineffective. The power of the formula would be reduced, variance would be increased, and this would lead to wide confidence intervals and small EAPs as well. We decided to test the robustness of this method by using the same three parametric variance estimators when calculating the variance of the AUC estimate instead of using the nonparametric method by DeLong et al. (1988). This provided decent EAPs that did not have the extreme drops and peaks as found in the previous setting.

There are several areas for future research. One area would be adjusting for covariates through stratification or regression. We did not consider the possibility that there may be confounding variables or other factors that may influence the results of a diagnostic test, so a diagnostic test may perform differently according to certain characteristics that a group of subjects may have. One could consider stratifying the groups by the confounding characteristic and then conducting the diagnostic test for each group separately. In this case, the proposed sample size formulas would be adjusted for estimating the required sample size within each stratum, and the

AUC may be an indication of the test accuracy within each stratum. More explicitly, the effect of covariates can be adjusted using regression to estimate the AUC and the sample size formula for the AUC can be adjusted as well.

Multiple AUCs can be compared to determine the differences between diagnostic tests—whether one may be more accurate than another or whether one is better suited for diagnostics or screening purposes. Obuchowski (1997) proposed a sample size formula for comparing multiple AUCs, and the formula for the difference between two AUCs can be improved to include a prespecified precision so that the sample size is not underestimated.

Lastly, sample size estimation can also be improved in the case of using repeated measures from the same subject. For repeated measures, the correlation between the measures from the same subject should be taken into account to improve the precision of the AUC estimation. Further study would be required to derive and evaluate a sample size formula with prespecified precision with the incorporation of repeated measures and their correlation

Bibliography

Altman, D.G., & Bland, J.M. (1994). Diagnostic tests 1: sensitivity and specificity. *BMJ, 309*, 1552.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating graph. *Journal of Mathematical Psychology, 12*, 387-415.

Bast, R.C., Klug, T.L., St. John, E., Jenison, E., Nilo , J.M., Lazarus, H., Berkowitz, R.S., Leavitt, T., Griffths, C.T., Parker, L., Zurawski, V.R. & Knapp, R.C. (1983). Radio- immunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer. *New England Journal of Medicine, 309*, 883-887.

Daly, L.E. (1991). Confidence intervals and sample sizes: don't throw out all your old sample size tables. *BMJ. 302*, 333-336.

DeLong, E.R., Delong, D.M., & Clarke-Pearson, D.L. (1988). Comparing the areas un-der two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837-845.

Del Villano, B.C., Brennan, S., Brock, P., Bucher, C., Liu, V., McClure, M., Rake, M., Space, B. & Zurawski, V.R. (1983). Radioimmunometric assay for a monoclonal antibody-defined tumor marker, CA19-9. *Clinical Chemistry, 29*, 549-52.

El Khouli, R.H., Macura, K.J., Barker, P.B., Phil, D., Habba, M.R., Jacobs, M.A., & Bluemke, D.A. (2010). The relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced (DCE) MRI of the breast. *Journal of Magnetic Resonance Imaging: JMRI, 30*(5), 999-10004.

Gardner, M.J., & Altman, D.G. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *Statistics in Medicine, 292*, 746-750.

Gordon, I. (1987). Sample size estimation in occupational mortality studies with use of confidence interval theory. *American Journal of Epidemiology, 125*(1), 158-162.

Green, D.M. & Moses, F.L. (1966). On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin, 66*(3), 228-234.

Green, D.M. & Swets, J. (1966). *Signal Detection Theory and Psychophysics*. New York: John Wiley and Sons, Inc.

Greenland, S. (1988). On sample size and power calculations for studies using confidence intervals. *American Journal of Epidemiology, 128*, 231-237.

Hanley, J.A., & Hajian-Tilaki, K. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update. *Statistics in Radiology, 4*(1), 49-58.

Hanley, J.A., & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Hanley, J.A., & McNeil, B.J. (1983). A method of comparing the area under two ROC curves derived from the same cases. *Radiology, 148*(3), 839-843.

Kupper, L.L., & Hafner, K.B. (1989). How appropriate are popular sample size formulas? *The American Statistician, 43*, 101-105.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. New York: John Wiley.

Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*(1), 50-60

Noether, G.E. (1967). *Elements of nonparametric statistics*. New York: Wiley.

Obuchowski, N.A. (1994). Computing sample size for receiver operating characteristic studies. *Statistics in Radiology, 29*(2), 238-243.

Obuchowski, N.A., & McClish, D.K. (1997). Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. *Statistics in Medicine, 16*, 1529-1542.

Rothman K.J. (1978). A show of confidence. *New England Journal of Medicine, 299*, 1362-1363. DOI: 10.1056/NEJM197812142992410.

Rosner, B., & Glynn, R.J. (2009). Power and sample size estimation for the Wilcoxon rank sum test with application to comparisons of C statistics from alternative prediction models. *Biometrics, 65*, 188-197.

Wieand, S., Gail, M.H., James, B.R., & James, K.L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika, 76*(3), 585-592.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80-83.

Yerushalmy, J. (1947). Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques. *Public Health Reports, 62*(40), 1432-1449.

Zou, G.Y. (2012). Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in Medicine, 31*, 3972-3981.

Zou, G.Y. (2021). Confidence interval estimation for treatment effects in cluster randomization trials based on ranks [Press release]. DOI: 10.1002/SIM.8918.

VITA

EDUCATION

- Master of Science in Biostatistics 2018-2021
  Western University, London, Ontario, Canada

- Bachelor of Science in Psychology, Neuroscience, and Behaviour 2014-2018
  McMaster University, Hamilton, Ontario, Canada

WORK EXPERIENCE

- Research Assistant 2015-2018
  McMaster University, Hamilton, Ontario, Canada