TWO-STAGE NEGATIVE BINOMIAL GROUP TESTING MODEL FOR ESTIMATING PREVALENCE OF A RARE TRAIT

FRANCIS MWANGI KARIUKI

A Thesis Submitted to the Graduate School in Partial Fulfilment of the Requirements for the Master of Science Degree in Statistics of Egerton University

EGERTON UNIVERSITY

SEPTEMBER 2023

DECLARATION AND RECOMMENDATION

Declaration

This thesis is my original work and has not been presented in this university or any other for the award of a degree.

Signature: ………………………………… Date: …5/09/2023…………………………

Kariuki Francis Mwangi

SM123/14597/18

Recommendation

This thesis has been submitted with our approval as university supervisors.

Signature: ………………………………… Date: …5/09/2023…………………………

Dr. Ronald Waliaula Wanyonyi
Department of Mathematics
Egerton University

Signature: ………………………………… Date: …5/09/2023……………………

Prof. Ali Salim Islam
Department of Mathematics
Egerton University

COPYRIGHT

DEDICATION

I dedicate this dissertation work to my family; parents James Kariuki, and Elizabeth Wambui, and brother Wilson Waititu who have been a constant source of support and encouragement during the challenges of graduate school and life. I am truly thankful for having you in my life.

ACKNOWLEDGEMENTS

First, I want to thank my Lord and Saviour Jesus Christ for granting me sufficient grace, good health, wisdom, and favour throughout this work. I express my heartfelt gratitude to my supervisors Dr. Ronald Wanyonyi and Prof. Ali Islam for their selfless efforts, positive criticism, and guidance throughout the period I was preparing the dissertation. I want to acknowledge Dr. Justine Obwoge for being resourceful and for his consistent advice and support in programming and scholarly comments.

I wish to acknowledge my father James Kariuki and my mother Elizabeth Wambui for their prayers, financial support, and understanding during my studies, may God bless you abundantly. My gratitude goes to my colleagues Samwel Karanja, Flavian Awere, Collins Otieno, Samson Tsuma, Wesley Ochoka, and Erick Ogolla for their moral support and encouragement throughout the course. Finally, I am grateful to Charei Munene for his encouragement to enroll in postgraduate studies. Thank you and God bless you all.

ABSTRACT

Group testing is an economical screening strategy that is beneficial in terms of efficiency and cost-cutting. The idea dates back to World War II, and it entails amalgamating individual specimens into pools that are tested for the presence of a trait of interest. Since its inception, group testing literature has branched into two research areas: classification and estimation. Research work in group testing has concentrated on designs without errors and has mainly developed under the binomial model. However, a combination of inverse sampling and group testing has been established to be useful when there is a need to report estimates early in the screening process. The main focus under the negative binomial group testing designs has been to develop more efficient estimators and to determine optimum group sizes under the assumption that the testing process has no misclassification. However, errors associated with labelling, and misclassification are prone to occur in an experimental design. Retesting of pools has been established to improve the efficiency of an estimator and increase the precision of a test. This research has constructed and analyzed a two-stage negative binomial group testing model for estimating the prevalence of a rare trait when imperfect tests with known sensitivity and specificity are used. The study utilized the Maximum Likelihood Estimation (MLE) method to obtain the estimator and the Cramer-Rao bound method to compute the Fischer information of the estimator. The properties of the constructed estimator were examined. The efficiency of the constructed estimator relative to other estimators in pool testing designs was determined by computing the Asymptotic Relative Efficiency (ARE) and the Relative Mean Squared Error (RMSE). The procedure was illustrated, and the model was verified by performing Monte Carlo simulations using R programming language version 3.5.2. The research findings showed that the model was superior to the one-stage negative binomial group testing model with misclassification as low variances were obtained as the proportion p increased. Also, the constructed estimator performed more efficiently for higher values of p. Furthermore, the study can be used for surveillance of pathogens and monitoring the prevalence of infectious diseases such as the Coronavirus disease 2019 (COVID-19) to prevent another pandemic resurgence.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AIDS	Acquired Immunodeficiency Syndrome
ARE	Asymptotic Relative Error
CDC	Centers for Disease Control and Prevention
COVID-19	Coronavirus disease of 2019
FMD	Foot and Mouth Disease
GMO	Genetically Modified Organism
HHS	Department of Health and Human Service
HIV	Human immunodeficiency virus
HPV	Human papillomavirus
IPP	Infertility Prevention Project
MLE	Maximum Likelihood Estimator
MSE	Mean Squared Error
RMSE	Relative Mean Squared Error
STDs	Sexually Transmitted Diseases
WNV	West Nile Virus

LIST OF SYMBOLS

θ	The probability of declaring a pool positive
Ω	Sample space
x	Total number of pools tested
r	Number of pools that test positive on a retest
n	The number of pools to be tested for a trait of interest in the binomial model
k	Pool size
$\cup$	Union
$\cap$	Intersection
$\pi^*(p)$	The probability of declaring a pool positive on a retest
π_0	Specificity of the test
π_1	Sensitivity of the test
$\pi_b(p)$	Probability of classifying a group as positive in the binomial model

CHAPTER ONE

INTRODUCTION

1.1 Background Information

Group testing, also known as pool testing, occurs when individual specimens (e.g., blood, plasma, urine, swabs, etc.) from a population are pooled and then tested for the presence of a trait of interest. The idea of pool testing is credited to Dorfman's (1943) seminar work and was used as a testing strategy to weed out all the syphilitic men who were called for army induction during World War II. In this testing strategy, a portion of the individual samples to be tested are first amalgamated into groups of equal size before they are subjected to testing. If a pool tests negative, further tests are discontinued, and all the members within that group are declared free from the trait of interest. Otherwise, each item is tested individually because a positive test result on a group indicates that at least one member has the trait of interest. When the proportion of a trait of interest is low, group testing unlike the standard method of testing each individual, offers a feasible and economical method that is not prohibitive in terms of cost and time.

The research work in group testing is two-fold, namely classification and estimation. In classification, the aim is to identify the positive members having a trait of interest from their counterparts. The first documented work in classification was applied to reduce the screening cost and to identify all the syphilitic-positive recruits that were called for army induction (Dorfman, 1943). On the other hand, the goal of estimation is to approximate the proportion of a trait of interest without labelling any individual as positive or negative, which is the focus of this study.

Previous research work in group testing has mainly been developed under the binomial model where the number of pools to be tested for a trait of interest is fixed in advance (Nyongesa & Syaywa, 2010; Nyongesa, 2011; Tamba *et al*., 2012). However, other probabilistic models in group testing have also been considered, including the hypergeometric model (Bar-lev *et al.*, 2003), the beta-binomial model (Turechek & Madden, 2003), and the Geometric model (Pritchard, 2008). Furthermore, Bayesian approaches have also been considered in estimation (Pritchard & Tebbs, 2011a; Tebbs *et al.*, 2003). The regression model was proposed when screening for multiple infections (Tebbs *et al*., 2013). The regression models were extended to produce the covariate-adjusted estimates using the individual covariate information for pooled responses (Delaigle & Zhou, 2015; Tebbs *et al*., 2013; Wang *et al.*, 2015).

When the prevalence of a trait of interest is small, and there is a need to report estimates early in the screening process, the negative binomial model has been suggested to be viable in estimation (Pritchard & Tebbs, 2011b). In this model, the number of positive pools having a trait of interest is fixed in advance, and the testing process is deemed complete when the fixed number of positive pools is observed. This model has been applied to emergencies such as disease outbreaks (Solomon *et al.*, 2003), natural disasters (Foppa *et al.*, 2007), and biological attacks (Thavaselvam & Vijayaraghavan, 2010). In recent studies developed under this model, different authors have examined point estimates (Pritchard & Tebbs, 2011b), Confidence intervals (Thong & Shan, 2015; Yu *et al.*, 2016), Optimum pool sizes when imperfect tests are used (Montesinos-López *et al.*, 2013; Xiong, 2016), improved estimators that are almost unbiased by applying a suitable correction factor **(**Hepworth, 2013), and Bayesian approaches (Pritchard & Tebbs, 2011a).

An aspect of concern when using group testing is errors associated with labelling and misclassification of items (Xie *et al*., 2001) and the dilution effect (Mokalled *et al*., 2021). Since the first work in estimation that considered imperfect tests, other scholars have also considered this concept in estimation (Matiri *et al*., 2017; Okoth *et al*., 2017a; Wanyonyi *et al*., 2015a). Their main focus has been to determine more efficient estimators and optimal group sizes when imperfect tests are used. The efficiency of an estimator is of great significance during estimation, and the retesting of pools has been considered to improve the efficiency of an estimator and also improves the precision of a test when the test kits are imperfect (Nyongesa, 2005; Nyongesa, 2011). Until now, no group testing design has been developed under the negative binomial model that has incorporated the sequential retesting of positive pools when imperfect tests are used. Therefore, this study proposes to develop a two-stage group testing procedure for estimating the prevalence of a trait using the Negative Binomial model by incorporating imperfect tests.

1.2 Statement of the Problem

The standard method of screening individuals' samples for the presence of a rare trait in a large population is both uneconomical and time-consuming. A viable way is to use group testing, which offers a cost-effective screening strategy. Research in public health highlights the significance of estimating the prevalence, even if disease identification is the main objective of a study. Most research conducted in group testing has concentrated on designs where pools are not misclassified and have largely presumed a binomial model where a fixed number of pools are tested for a rare trait. However, a combination of inverse sampling and group testing is desirable when there is a need to report estimates early in the screening process. Unlike the binomial model, the negative binomial group testing model is an appealing strategy where samples are continuously screened until a predetermined number of positive groups containing a rare trait are observed. Estimation work under the binomial model has focused on examining efficient estimators, determining the desired sample size, and estimating the proportion of a rare trait when imperfect tests are used. Different authors have considered group testing designs with errors associated with labelling and misclassification. Retesting of pools under the binomial model in group testing was shown to reduce misclassification and improve the efficiency of the estimator. The estimation work in the negative binomial group testing model has developed under the postulation that the testing process is perfect. Alternative estimators that reduced the bias were developed. When imperfect tests are used, the optimal group size in the negative binomial group testing model has been considered. To the author's knowledge, estimation procedures that incorporate imperfect tests and the sequential retesting of a pool that tests positive in the negative binomial group testing model are lacking in the statistical literature. The purpose of this study is to construct and analyse a two-stage negative binomial group testing procedure for estimating the prevalence of a rare trait when imperfect tests with known sensitivity and specificity are used. A pool that tests positive for a rare trait in the initial stage is sequentially given a retest, and the testing process continues until a predefined number of positive pools that test positive on a retest are observed.

1.3 Objectives

1.3.1 General Objective

To construct and analyze a two-stage negative binomial group testing procedure for estimating the prevalence of a rare trait.

1.3.2 Specific Objectives

i. To obtain an estimator for the prevalence of a rare trait using the Two-stage Negative binomial model in group testing.

ii. To determine the properties of the derived estimator such as the bias, and Mean Squared Error.

iii. To compare the proposed model with the one-stage negative binomial group testing model with misclassification.

iv. To apply the proposed model to West Nile Virus data.

1.4 Significance of the Study

When there was a need for an efficient procedure of screening for a rare disease in a population, Dorfman (1943) proposed group testing. In the procedure, tests are carried out on pooled samples and all individuals in a group that tests positive are retested. Since his seminal work, most of the current and recent past research in group testing has been developed under the assumption that the tests used are perfect. When imperfect tests are used, errors in experiments that are associated with labelling and misclassification are susceptible to occur, and thus the need to perform a confirmatory test. This has become an area of interest, and different scholars have adopted this concept when estimating the prevalence of a trait. Unlike the binomial model in group testing which is limited to a fixed number of pools to be tested for a rare trait. A combination of Inverse sampling and group testing is more appealing when there is a need to report estimates early in the screening process. This is because an early and accurate assessment of the prevalence level of a disease can prompt mitigation measures against an outbreak. The main focus under the Inverse binomial pool testing has been to develop more efficient estimators and optimum group sizes under the assumption that assays used for screening are perfect. However, one can envision an experimental situation where inverse sampling has been applied, and experimental errors occur as a result of imperfect tests used. To increase the precision of a test during estimation, there is a need for retesting pools. It has been established that retesting of pools improves the efficiency of an estimator, and it recovers lost sensitivity and specificity. The benefits of this study outweigh the disadvantages of one at a time testing when estimating the proportion of a trait in a low prevalent population. The study also contributes to the existing literature on group testing when negative binomial group testing models with retesting are used in estimation. Furthermore, the study can be used to screen and monitor the prevalence of infectious diseases such as the Coronavirus disease 2019 (COVID-19) to prevent another pandemic resurgence.

1.5 Assumptions

i. The sensitivity and specificity of the test are held constant throughout the testing period.

ii. Individual outcomes are independent and identically distributed.

iii. There were no dilution or shielding effects due to the pooling of samples.

1.6 Definition of terms

i. **Rare trait**

A characteristic or an attribute that is not commonly observed within a specific population

ii. **Pool/ Group**

A set of individuals or specimens combined for testing

iii. **Sensitivity**

This is the probability of a test to correctly classifying a positive pool or individual.

iv. **Specificity**

This is the probability of a test to correctly classify a negative pool or individual.

v. **Group Testing**

A technique where test subjects or samples are combined to form pools, before being subjected to testing as a group rather than testing the subjects or samples individually.

vi. **Re-testing**

This refers to the testing of a group or individual more than once.

vii. **Parameter**

This is a characteristic that defines and describes a population.

viii. **Estimation**

A specified procedure that computes the value of some property of the population

ix. **Prevalence**

This is the proportion of a population who have a rare trait in a given period.

x. **Confidence interval**

This refers to the probability that a population parameter will fall between a set of values for a certain proportion of times.

xi. **Coverage probability**

The proportion of the number of times a confidence interval contains the true value of the parameter.

xii. **Statistic**

A statistic is a function of observable random variables, which is itself an observable random variable, which does not contain any unknown parameters.

CHAPTER TWO

LITERATURE REVIEW

2.1 Foundation of Group Testing

When screening for the presence of a trait of interest (e.g., infection) in a large population, a countless number of individual samples need to be tested. When the prevalence of a trait is low, group testing offers an appealing testing strategy that is not prohibitive in terms of cost and time. Group testing entails amalgamating specimens (e.g., blood, plasma, urine, swabs, etc.) from a population into pools and then screening the pools for a trait of interest. The idea of group testing is credited to Dorfman (1943), who used it as a strategy to weed out all syphilitic men who were called for army induction during World War II. In this testing strategy, individuals' samples were amalgamated into pools and then tested as a unit. If a pool tested negative, all individuals within were declared free from infection. Otherwise, a positive reading on a pool meant that at least one individual was infected, instigating the need for individual testing to decode positive individuals from negative ones. In this testing strategy, Dorfman (1943) recorded a reduction in the expected number of tests and the overall cost of screening when disease prevalence was low.

To further reduce the expected number of tests and the testing cost when the prevalence is small, a modification of the Dorfman testing scheme has been examined and extended to multi-stage (Bilder *et al.*, 2010). This testing scheme involved testing of pooled samples derived from a population of interest. A negative reading on a pool inferred that all members within the pool were free from infection and further tests were discontinued. Otherwise, members within the pool were randomly selected and then retested one by one until the first positive individual was identified. The remaining members would then be combined to form a new pool before they are subjected to another test. If the new pool tests positive, the same procedure would be repeated until all the members were classified as positive or negative.

In multiple stages of group testing, the halving algorithm has been developed, where the individual covariates information was applied to accomplish information retesting (Black *et al*., 2012). In the halving algorithm, pools that test positive are split into two halves before being subjected to a retest. The procedure only stops when a pool tests negative or until individual testing occurs. Other subsequent works have been reviewed because the individual covariate information was used to estimate the probability that an individual tests positive (Saker, 2016). It was noted that an informative approach based on the Dorfman pooling scheme is superior to those not considering individual heterogeneity (McMahan *et al*., 2012). The risk probability of an individual is given preference in that a positive pool that is to be subjected to a retest has to have the highest probability of being positive.

An improved testing scheme based on Dorfman's (1943) testing protocol was later developed when screening for the presence of Human Immuno-deficiency Virus (HIV) and acquired immunodeficiency syndrome (AIDS) in a population (Monzon *et al*., 1992). In their testing design, pools of individual samples were drawn from the population and then tested. Further tests were discontinued for a pool that tested negative, and all members within that pool were declared free from infection. Otherwise, pools that tested positive were retested. In a pool that tested negative after a retest, members within that pool were declared free from infection. Otherwise, the standard method of individual testing was applied.

2.2 Pool Testing Schemes

Based on the number of possible stages in group testing procedures, two forms of pool testing schemes have been proposed, namely hierarchical (adaptive) and non-hierarchical (non-adaptive) testing schemes (Okoth *et al*., 2017a). The hierarchical testing procedure utilizes the information from the previous stage to determine the testing pattern of the subsequent stages. An example of a two-stage adaptive stage is the Dorfman (1943) pool testing strategy. Besides, Sterrett's (1957) pool strategy scheme is also a hierarchical procedure with an unknown number of stages.

The test result of a group is dichotomous, meaning it can either be a positive or a negative result. Non-hierarchical models are constructed by utilizing the test results of a group. The standard individual testing which takes a single step to identify individuals who are either positive or negative, falls under this testing scheme. Array (matrix) testing is also an example of a non-hierarchical model that has considered imperfect tests (Kim *et al.,* 2007). It involves organizing specimens into a square grid before pooling the samples for testing based on rows and columns. Individual testing would be conducted on specimens that fell on the intersection of a positive row and a positive column. Other succeeding works in array testing have also been developed, for instance, array testing in more than one directional (Berger *et al.,* 2000), and three-dimensional array procedure with testing errors (Kim & Hudgens, 2009).

2.3 Estimation in Group Testing

The research work in group testing literature has branched into two distinct areas, namely classification and estimation. The two areas have received substantial attention under the binomial model, which presumes a fixed sampling design where pools are screened for the absence or presence of a trait of interest. Alternatively, other experimental designs in group testing have been suggested where sampling and testing occur sequentially until a predetermined number of positive pools are observed (Haber *et al.,* 2018). This research focuses on the latter by considering the negative binomial model in group testing that incorporates the retesting of pools that test positive in the initial test when imperfect tests are used.

The binomial model was first considered in the insect-vector problem to estimate the proportion of insect vectors capable of transmitting the aster-yellow virus (Thompson, 1962). The method of Maximum Likelihood was used to obtain the estimator and to examine its properties such as bias and best asymptotic normal (BAN). The estimator was noted to be positively biased as the prevalence increased. Other subsequent work in estimation followed, including estimating prevalence based on a pool testing scheme with retesting (Nyongesa, 2011). It was established that pool testing offered a cost-effective method, and retesting of pools improved the estimator's efficiency and recovered the sensitivity of the testing scheme.

Group testing designs have been applied in different situations, and the focus has been to develop more efficient estimators and optimal group sizes. The test assays have a threshold of detecting a trait of interest, and a group testing procedure has been applied in genetically modified organisms (GMOs) to examine the required sample size (Yamamura & Hino, 2007). It was noted that even if a threshold of detection exists, it was possible to estimate the proportion of defective items for any group size. If a group is large and the proportion of infection is small, a group can easily be misclassified as one that is free of infection, and this is termed the dilution effect. The required group size has been examined by considering the dilution effect when detecting the adventitious presence (AP) of transgenic plants (Hernández-Suárez *et al*., 2008).

If individuals in a population are grouped into *n* pools each of size *k*, out of which *x* pools test positive. Then *x* has a binomial distribution with parameters *n* and $1-(1-p)^k$ or simply written as $x\sim Binomial\ (n,\ 1-(1-p)^K\ )$.

The MLE of the prevalence, *p* was obtained by Nyongesa (2011) to be

$$\hat{p}=1-\left(1-\frac{x}{n}\right)^{\frac{1}{k}}. \qquad (1)$$

When group testing is applied in the presence of misclassification, the optimum properties of group testing strategies were considered by Liu *et al.* (2012). The exact range of disease prevalence was computed for which the testing strategy provided more efficient estimators as the group sizes increased. Recently, assays have been developed which can detect multiple diseases simultaneously. Regression models have been proposed to address the challenge when screening for multiple diseases concurrently (Zhang *et al.*, 2013). It was pointed out that this was attributed to a likely correlation between the unobserved individual's disease statuses. When considering hierarchical testing schemes, a multi-stage pooling strategy in estimation was established by Brookmeyer (1999). In this pooling strategy, pools formed were first tested then a retest was performed on pools that tested positive. This was achieved by sequentially subdividing the positive pools before they are retested. A reduction in the variance associated with each additional stage in the multi-stage pooling studies was noted. Besides, the obtained results were extended to estimate the disease incidence rate, and the method was applied to screen for HIV.

When the assay used is not 100% accurate, previous research points out a loss of sensitivity during estimation (Nyongesa, 2011). Retesting of pools has been shown to improve the efficiency of the estimator and reduce misclassifications (Nyongesa, 2018). When the sensitivity and specificity of the tests are held constant throughout the testing scheme, a retesting model in estimation was developed by Nyongesa and Syaywa (2010). They established that retesting of negative pools improved the efficiency of the estimator, and the model suggested a practical use in blood donation. Elsewhere, retesting of the positive pools was established to recover the sensitivity when imperfect tests were used in the Monzon *et al.* (1992) testing scheme. A statistical pool testing model with retesting has been developed based on the Monzon *et al.* (1992) testing scheme, and the work has been extended to a multi-stage pooling strategy (Nyongesa, 2018; Okoth *et al.*, 2017b). The Asymptotic Relative Error (ARE) was computed, and the results confirmed that retesting improved the efficiency of the estimator.

The idea of pool testing in the presence of testing errors was earlier introduced by Nyongesa and Syaywa (2010) and Nyongesa (2011). If the probability that a pool tests positive is $\pi_b(p)$, and that x out of n groups of size k test positive, then x was shown to have a binomial distribution given as

$$x \sim Binomial\ (n, \pi_b(p)),$$

where

$$\pi_b(p) = \pi_1(1-(1-p)^k) + (1-\pi_0)(1-p)^k, \tag{2}$$

and π_1 and π_0 are the sensitivity and specificity of the test respectively.

Using the model, the MLE of p was obtained as

$$\hat{p} = 1 - \left[\frac{1 - \frac{x}{n}}{\pi_0 + \pi_1 - 1}\right]^{\frac{1}{k}}. \tag{3}$$

A comparison was made between individual testing and pool testing in the presence of testing errors to ascertain which provided a better estimator, and the results showed that pool testing improved the efficiency of the estimator Nyongesa (2011). A computational statistical model for pool testing that incorporated retesting was also examined by Tamba *et al.* (2012). The number of misclassifications and the cost incurred in the testing scheme were examined, and it was established that pool testing is economical in a low prevalence population and that retesting of pools reduced misclassification.

Group testing has also found its application in the quality control process and has been applied in the contamination of GMOs when there are inspection errors (Wanyonyi *et al.*, 2015b). It was noted that when the proportion of a trait of interest is relatively high, the batch-testing model was superior to other existing models. In other research, the cut of values in relation to batch testing has been examined (Matiri *et al.*, 2017; Wanyonyi *et al.*, 2015a). The probability of detecting a positive batch was noted to be affected by the batch size and the cut-off values (Wanyonyi *et al.*, 2015a). Moreover, by comparing their model with that of Brookmeyer (1999), they noted that at high prevalence, their model improved the efficiency of the estimator over other existing estimators.

Elsewhere, a statistical model has been constructed to select a combination of two or three experiments when imperfect tests are used in batch testing (Matiri *et al.*, 2017). The MLE and cut-off values were obtained, and the results of Fischer information were compared for the different experimental models. Through comparison, the proposed joint model was spotted to be more efficient than the other two existing models for any range of prevalence when the sensitivity and specificity were held constant. Re-testing of batch testing models based on the quality control process has been examined by Wanyonyi *et al.* (2021). Batches that tested positive were given a retest, and the results indicated that retesting improved the estimator's efficiency over the one-stage batch testing in a quality control process. Moreover, the model was established to be superior to the classical two-stage batch testing in that the estimator recorded smaller variance for relatively high values of the proportion.

The binomial model in group testing has been explored extensively by different authors. The drawback of the model is that it utilizes a fixed number of pools set by the researcher to test for a trait of interest. The model may not be helpful in situations that require quick responses like in an emergency, and disease outbreaks, where sampling and testing are done until a predetermined number of pools having a trait of interest are observed. This calls for a combination of Inverse sampling and group testing to be examined. It has been established that the negative binomial model is more appealing when estimating the prevalence of a rare trait (Hepworth, 2013; Pritchard & Tebbs, 2011b; Xiong, 2016).

2.4 Inverse Binomial Model in Group Testing

When sampling biological samples, inverse binomial sampling is of great importance. If the proportion of individuals possessing a character trait is *p*, and sampling is done until a specified number say, *r* individuals are observed. Then, the number of individuals sampled follows a negative binomial distribution. A combination of inverse sampling and group testing offers an appealing strategy when reporting estimates early in the screening process (Hepworth, 2013).

The model assumes that the number of pools with a trait of interest is fixed in advance and that the testing continues until the desired number of positive pools is observed. Inverse binomial sampling was first used in the estimation of frequencies of an attribute in a population to deduce the unbiased estimator and to investigate the properties of the variance (Haldane, 1945). It was pointed out that most of the experimental errors arise from the sampling process and not from the pooling scheme (Katholi & Unnasch, 2006).

The MLE was used to obtain the point estimator and to discuss the confidence intervals for equal pool sizes by Katholi and Unnasch (2006). The work was extended by examining the point estimators and alternative estimators that reduced the bias for equal and unequal pool sizes (Pritchard & Tebbs, 2011b). The Bayesian approach has also been considered by incorporating the prior knowledge of the incidence rate and different loss functions (Pritchard & Tebbs, 2011a). The results were used to examine point estimators and credible intervals. An improved estimator has been considered by applying a suitable correction factor to obtain an almost unbiased estimator by Hepworth (2013). The score-based method with a correction for skewness and the exact method with a mid-*p* correction factor were recommended for their exceptional coverage properties.

The optimal group sizes under the negative binomial group testing model have been investigated in detecting the adventitious presence (AP) of transgenic plants in a population (Montesinos-L’opez *et al*., 2013). The three proposed methods i.e. two computational and one analytical methods were noted to provide a good approximation and ensured precision in the estimated proportion that guaranteed narrow confidence width. The use of a negative binomial model in group testing was applied during the screening of *Onchocerciasis volvulus,* responsible for causing ocular and skin disease (Rodriguez-Perez *et al*., 2006). The Method of Maximum Likelihood was used by Pritchard and Tebbs (2011b) to show that if $X = x$ is the total number of pools tested until the r^{th} positive pool is observed, then x follows a Negative binomial distribution with waiting parameter r, and success probability $\pi(p) = 1 - (1-p)^k$. The probability density function is given by

$$f(x \mid p, r, k) = \binom{x-1}{r-1}\,(1-(1-p)^k)^r\,(1-p)^{k(x-r)}. \tag{4}$$

The Maximum Likelihood Estimate (MLE) of p was shown to be

$$\hat{p} = 1 - \left[1 - \frac{r}{x}\right]^{\frac{1}{k}}. \tag{5}$$

The variance of the estimate can be obtained from the information function given by

$$I\,(p) = \frac{r\,k^2\,(1-p)^{k-2}}{(1-\,(\,1-p\,)^k)^2}. \tag{6}$$

The MLE obtained in the binomial model was shown to be positively biased, especially for group size $k > 1$ (Muhua, 2010). The MLE of the negative binomial model has the same form as that of the binomial model, and alternative estimators that reduce the bias of the estimator were examined by Pritchard and Tebbs (2011b). Simulation studies indicated that both the shift and combined estimator reduced the bias. This was performed assuming that perfect tests are used in the testing scheme.

The first work in estimation that considered imperfect tests was pioneered by Litvak *et al*. (1994), who applied the procedure to screen for HIV. The results indicated that the testing scheme improved the accuracy of the estimator and lowered the number of false positives. The concept that the tests may not be 100% perfect i.e. the sensitivity and specificity value is less than a unit, was incorporated in the testing scheme. Recent scholarly works have also considered this concept in estimation under the binomial model in group testing (Matiri *et al*., 2017; Okoth *et al*., 2017a; Wanyonyi *et al*., 2015a). Their studies have mainly focused on determining more efficient estimators and optimal group sizes when imperfect tests are used. The optimal group size under the negative binomial model, which has considered an imperfect test to estimate optimal group sizes was examined by Xiong (2016). The MLE and the variance of prevalence p were obtained

$$\hat{p} = 1 - \left\{\frac{{\pi_1}^2 - \frac{r}{x}}{\pi_0 + \pi_1 - 1}\right\}^{\frac{1}{k}} \tag{7}$$

and

$$var(\hat{p}) = \left\{\frac{[\pi_1 - (\pi_0 + \pi_1 - 1) \times (1-p)^k]^2 \times [1 - \pi_1 + (\pi_0 + \pi_1 - 1) \times (1-p)^k]}{rk^2(\pi_0 + \pi_1 - 1)^2 \times (1-p)^{2k-2}}\right\} \tag{8}$$

where k is the group size, n is the waiting parameter denoting the predetermined number of positive pools, and π_1 and π_0 denote the sensitivity and specificity of the tests respectively.

Surveillance of pathogens plays a vital role in public health and risk assessment. The early detection of infectious diseases is paramount to reducing the severity of an outbreak. Although vaccination for infectious diseases like foot and mouth disease (FMD) exists, an outbreak can have a catastrophic effect on the meat and milk industry (Callahan *et al*., 2002). The impacts of technological advancement cannot be undermined, and one of the looming crises is bioterrorism, which is a deliberate act of releasing biological toxins and agents as an act of war (Zilinskas, 1997). The FMD was suggested as a possible biological agent, hence the need for early and accurate detection to prevent an outbreak and limit the spread of infection (Koda, 2002). Pritchard and Tebbs (2011b) suggested that their methodologies can be modified to suit the testing scheme when imperfect tests are used in the negative binomial group testing model. The retesting of pools under the binomial model was established to improve the efficiency of an estimator and to recover the sensitivity of the test during estimation (Nyongesa, 2011; Nyongesa, 2018; Wanyonyi *et al*., 2021). The estimation procedures in group testing design that incorporates the sequential retesting of positive pools in the negative binomial model are still lacking.

Therefore, the problem motivating this study is constructing and analyzing a two-stage negative binomial group testing procedure for estimating the prevalence of a rare trait. In this procedure, sampling and testing occur sequentially, whereby a pool that tests positive is retested. The testing process continues until the desired number of pools that test positive on retesting are observed. The only drawback of the negative binomial model is that depending on the pool size and the prevalence of a trait of interest, countless tests have to be performed to observe a few positive pools. Thus, the number of positive pools with a rare trait was suggested to be small for practicability (Pritchard & Tebbs, 2011b). If the sampling process is done seamlessly, the procedure can save on cost. It can be used for rapid and accurate estimation of infectious diseases such as the Coronavirus disease of 2019 (COVID-19) which is a global pandemic.

2.5 Interval Estimation

One of the basic methodologies in statistics is interval estimation. There are different approaches to constructing confidence intervals that have been reviewed by different authors (Luchen, 2012; Pritchard & Tebbs, 2011a; Thong & Shan, 2015; Yu *et al.,* 2016). The Wald confidence interval is the most common in practice and statistical literature. When the probability distribution of the estimator is known, the confidence interval contains the unknown parameter within its bounds with a high pre-specified probability. The standard Wald confidence interval based on the normal approximation of the MLE of *p* is given by

$$\hat{p} \pm Z_{\alpha/2}\sqrt{var(\hat{p})} \qquad (9)$$

where $Z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile from the $\mathcal{N}(0,1)$ distribution and $var(\hat{p}) = (1-(1-p)^k)/(nk^2(1-p)^{k-2}$ is the asymptotic variance of *p*, recommended by Thompson (1962). The Wald confidence intervals were observed to suffer from problems associated with overshooting and zero confidence width (Orawo, 2021). The Wilson interval of the binomial proportion *p* was pointed to be asymmetric and can be derived from inverting the z-test for *p.* The two-tailed Wilson interval is of the form

$$\frac{\hat{p} + \frac{\left(Z_{\alpha/2}\right)^2}{2n} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{\left(Z_{\alpha/2}\right)^2}{4n^2}}}{\left(1 + \frac{\left(Z_{\alpha/2}\right)^2}{n}\right)} \qquad (10)$$

where *n* is the number of groups tested, and $Z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile from the $\mathcal{N}(0,1)$ distribution. The generalized two-sided Agresti-Coull interval for the binomial distribution with parameters (n, p) has been reviewed by Brown *et al.* (2001) and takes the form

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{\tilde{n}}} \qquad (11)$$

where $\tilde{n} = n + 4$ and $\hat{p} = \left(\frac{x+2}{n+4}\right)$ is a re-centered estimator of the proportion *p.* It was pointed out that the confidence interval has good coverage probabilities, but it is more conservative for the proportion *p* close to 0 (Brown *et al*., 2001). An Exact interval for the negative binomial group testing proportion p has been examined by Pritchard and Tebbs (2011a). When pools of size k are used $x\sim$ negative binomial (r, θ) where $\theta = (1-(1-p)^k)$. An exact interval for θ was derived by Lui (1995) using the relationship between negative binomial distribution and the incomplete beta function. The lower and upper confidence limits are $\theta_L = B_{1-\gamma/_2, r,\ x-r+1}$ and $\theta_U = B\gamma_{/_2,\ r,\ x-r}$ respectively where $x = \sum_{i=1}^{r} G_i$, and $B_{\gamma,\ a,\ b}$ denotes the upper γ quantile of the two-parameter $\text{beta}(a, b)$ distribution and $G_i$follows a geometric distribution. The exact interval was obtained by transforming the endpoints of the θ interval which reduced to $p_L = 1-\left(1-B_{1-\gamma/_2, r,\ x-r+1}\right)^{\frac{1}{k}}$ and $p_U = 1-\left(1-B\gamma_{/_2, r,\ x-r}\right)^{\frac{1}{k}}$.

2.6 Application of Group Testing

Group testing has been applied in a variety of different fields since the inception of Dorfman's (1943) seminar work. In industrial applications, group testing has been applied in making a “leak test” on a large number of electrical devices filled with a gas (say, helium) as outlined by Mundel (1984). Any number of units say *x* can be tested using a single test, and the result of the test is that either all the *x* units are good or at least one of the *x* is defective. Electrical devices such as condensers, resistors, etc. are tested in a similar manner.

Group testing has also been applied in the Infertility Prevention Project (IPP), a national project funded by the Centers for Disease Control and Prevention (CDC) and the Department of Health and Human Services (HHS) (Tebbs *et al.*, 2013). The objective of the project was to identify infected individuals with Chlamydia, or gonorrhoea through screening. The trends in prevalence were monitored and treatment was offered to the infected individuals.

Similarly, group testing was earlier applied to screen for chlamydia and gonorrhoea (Lindan *et al.*, 2005). The two bacterial infections were pointed out to be responsible for causing pelvic inflammatory diseases, ectopic pregnancies, sterility, and infertility. Elsewhere, research showed that the two bacterial infections were also responsible for the transmission of other sexually transmitted diseases (STDs) like HIV and Human papillomavirus (HPV) (Lewis *et al.*, 2012). Group screening has also been applied for a variety of STDs, including HIV (Pilcher *et al.*, 2005), hepatitis B, and hepatitis C (Cardoso *et al.*, 1998).

Red Cross organizations in Japan and Germany used this technique to screen for blood samples (Mine *et al.*, 2003). To curb the spread of HIV infection, group testing was used to screen for the presence of HIV antibodies (Kline *et al.*, 1989; Monzon *et al.*, 1992). It was shown that group testing lowered misclassification when screening for HIV in a low-risk population (Litvak *et al.*, 1994). This testing strategy was suggested to be useful when concealing the identity of the subjects tested due to the stigma associated with the HIV/AIDS virus (Gastwirth & Hammick, 1989).

Pool testing has also been applied in the early stages of drug discovery (Xie *et al.,* 2001). The results demonstrated a reduction in the cost incurred, unlike when the standard testing protocol is applied. Elsewhere, it has been applied in arbovirus literature to screen for WNV (Busch *et al*., 2005; Rutledge *et al*., 2003) and to screen for the H1N1 influenza virus (Van *et al*., 2012). Furthermore, group testing has also been applied in quality control processes (Fang *et al*., 2007; Wanyonyi *et al*., 2015b), and in industrial experimentation (Vine *et al*., 2008).

CHAPTER THREE

MATERIALS AND METHODS

3.1 Probability Theory

Under the Probability Theory, the Indicator function and the Theorem of total probability were used to derive the probabilities of interest.

3.1.1 Indicator Function

A random variable that takes the value 1 when an event happens and 0 when the event does not happen is called an indicator function of an event. If we let Ω be a sample space and $E \subseteq \Omega$ be an event. The indicator function (or indicator random variable) of the event E denoted by I_E is a random variable defined by

$$I_E = \begin{cases} 1, & if\ \omega \in E \\ 0, & if\ \omega \notin E \end{cases} . \tag{12}$$

The following indicator functions were used in the development of the proposed model to simplify the notations:

Define:

$$T_i = \begin{cases} 1, & if\ the\ i^{th}\ group\ is\ tests\ positive. \\ 0, & otherwise. \end{cases}$$

$$T^*{}_i = \begin{cases} 1, & if\ the\ i^{th}\ group\ test\ positive\ on\ the\ retest \\ \\ 0, & otherwise. \end{cases}$$

$$D_i = \begin{cases} 1, & if\ the\ i^{th}\ group\ is\ truly\ positive. \\ 0, & otherwise. \end{cases}$$

$$T^*{}_{ij} = \begin{cases} 1, & if\ the\ j^{th}\ \ individual\ in\ the\ ith\ group\ tests\ positive\ on\ the\ retest \\ \\ 0, & otherwise. \end{cases}$$

3.1.2 Theorem of Total Probability

A partition is a collection of non-empty, non-overlapping subsets of a sample space whose union is the sample space itself (Mood *et al.*, 1974). If we let $A_1, A_2, \ldots, A_n$ be a collection of events that partition the sample space S and B is an arbitrary event within S, then B can be expressed as a union of subsets as follows

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n) \tag{13}$$

where the bracketed events $(B \cap A_i)\ for\ i = 1, 2, \ldots, n$ are mutually exclusive events. Using the additional law of probability for mutually exclusive events

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n). \tag{14}$$

Each partition on the right-hand side may be expressed in terms of conditional probabilities as follows

$$P(B \cap A_i) = P(B / A_i)\, P(A_i). \tag{15}$$

Using the expression in equation (15) into (14) we have

$$\begin{aligned} P(B) &= P(B / A_1)\, P(A_1) + P(B / A_2)\, P(A_2) + \cdots + P(B / A_n)\, P(A_n) \\ &= \sum_{i=1}^{n} P(B / A_i)\, P(A_i). \end{aligned} \tag{16}$$

The theorem of total probability and the indicator functions were of great importance in the development of the proposed model to simplify notations and compute conditional probabilities.

3.2 Parametric Point Estimation

The Method of Maximum Likelihood Estimation and asymptotic variance were considered in parametric point estimation.

3.2.1 Method of Maximum Likelihood

This study utilized the method of maximum likelihood because it is convenient, and it gives estimators with good statistical properties. Suppose that $Y_i = y_i$ denotes the number of groups with size *k* that are tested for a trait of interest until the *r*th positive pool on a retest is observed. Therefore, Y_i has a geometric distribution. If *x* is the total number of pools tested is denoted by $x = \sum_{i=1}^{r} Y_i$, then *x* follows a negative binomial distribution with a with waiting parameters *r* and success probability $\pi^*(p)$. The likelihood function is given by

$$L(x|\, p) = \prod_{i=1}^{r} \pi^*(p)\,(1 - \pi^*(p))^{y_i - 1} \tag{17}$$

The maximum likelihood estimate of *p* was obtained as the solution to the equation,

$$\frac{\partial}{\partial p} \log L(p) = 0. \tag{18}$$

3.2.2 Asymptotic Variance

The asymptotic variance of the estimate *p* was obtained by computing the Fishers information given by;

$$var(\hat{p}) = E\left\{-\left(\frac{\partial^2}{\partial p^2}\log L(p)\right)\right\}^{-1}. \tag{19}$$

3.3 Model Comparison

The properties of the derived estimator such as the bias, and Mean Squared Error (MSE), were computed as follows;

$$Bias(\hat{p}) = \sum_{t=n}^{t^*}(\hat{p} - p)\binom{x-1}{r-1}\pi^*(p)^r(1-\pi^*(p))^{x-r} \tag{20}$$

and

$$MSE(\hat{p}) \approx \sum_{x=r}^{x^*}(\hat{p} - p)^2\binom{x-1}{r-1}\pi^*(p)^r(1-\pi^*(p))^{x-r} \tag{21}$$

respectively, where $\pi^*(p)$ denotes the probability of classifying a group as positive on retesting, *r* is the number of pools that test positive on retesting, and $pr(X \geq x^*) \leq 0.00001$ for small values of p considered. It was pointed out that if 0.00001 is replaced by a smaller number, it makes virtually no difference to the results (Hepworth, 2013). Thus, the constant $v = 0.00001$ was taken throughout the simulation such that $pr(X \leq x^*) \geq 1 - v$ making these approximations very close to the true values of bias and MSE.

3.3.1 Asymptotic Relative Error (ARE)

If the estimator of the one-stage negative binomial group testing model with misclassification as suggested by Xiong (2016) model is denoted by $\hat{p}_T$ and the proposed estimator is denoted by $\hat{p}_R$ then ARE values were calculated as;

$$ARE = \frac{var(\hat{p}_T)}{var(\hat{p}_R)}. \tag{22}$$

The proposed model was found to be efficient for $ARE > 1$.

3.3.2 Relative Mean Squared Error (RMSE)

The RMSE was given by

$$RMSE = \frac{MSE(\hat{p}_T)}{MSE(\hat{p}_R)}. \tag{23}$$

3.4 Comparison of Interval Estimates

The procedure for comparing the confidence interval was based on the coverage probability. In this section, simulation studies were carried out to compare the performances of the Wald, Wilson, Agresti- Coull, and Exact intervals based on their coverage probabilities. The true coverage probability of an estimator can be approximated in the same manner as its bias and the MSE for a given prevalence *p*, waiting parameter *r*, and pool size *k* when the accuracy of the assay used for testing is known. It was pointed out that for any confidence interval method for estimating the proportion *p*, the actual coverage probability can be approximated using the following equation

$$C(p,r,k,\pi_0,\pi_1) = \sum_{r}^{x^*} I(\hat{p})\binom{x-1}{r-1}\pi^*(p)^r(1- \pi^*(p))^{x-r} \qquad (24)$$

where $\pi^*(p) = {\pi_1}^2 + (1-p)^k((1-\pi_0)^2 - {\pi_1}^2)$, $I(\hat{p}) = 1$ if the interval contains p, or $I(\hat{p}) = 0$ otherwise. The plots of the coverage probabilities were computed for the values of p ranging from 0.00005 to 0.10 by increments of 0.00005 for group sizes $k = 20$ and the waiting parameter $r = 2, 5, 20, 50$ at 95% nominal level. Also, a table of coverage probabilities was constructed when the proportion was fixed at $p = 0.001, 0.025, 0.05, 0.10$.

3.5 Simulation

R-Programming language version 4.0.3 (2020-10-10) was used to perform simulation studies. Simulation describes a numerical technique for conducting experiments on a computer that depends on repeated random sampling from probability distributions. A generalised Monte-Carlo algorithm used in this study involved the following:

Step 1: Setting fixed values of *r*, *k*, *p*, sensitivity, and specificity values.

Step 2: Generate *N* independent data sets from negative binomial $\left(r,\ \pi^*(p)\right)$ that is

$T_i{\sim}nbinom\left(r,\ \pi^*(p)\right)$ for $i = 1, \dots, N$

Step 3: Compute the numerical value of the test statistics *T* for each data set $T_1, T_{2,}\ \dots, T_N$

Step 4: If *N* is large enough, summary statistics $T_1, T_{2,}\ \dots, T_N$ should be a good approximation to the true sampling properties of the test condition under the conditions of interest.

The study simulated different data sets of pool sizes, $k = 5, 10, 30,$ and 50 when the waiting parameter r was fixed at $r = 1, 3, 5, 10,$ and 15. The sensitivity and specificity value of the tests were assumed to remain constant throughout the entire study and were examined at $99\%, 98\%, 95\%,$ and 90% respectively.

3.6 Real Data

Data derived from a public health study that described the first documented field transmission of the West Nile Virus (WNV) conducted by Rutledge *et al*. (2003) has been utilized. The study used the method of reverse transcription-polymerase chain reaction to screen the pools of mosquitos collected through trapping and surveillance programs as a way to assess transmission of WNV. Collectively, out of 11,948 mosquitos that were collected, 14 pools tested positive for WNV. The study was conducted for four days in August of 2001 following the outbreak of WNV in Jefferson County, Florida. Both equal and unequal pool sizes were considered. Inverse sampling was not used by Rutledge *et al*. (2003), but we use this field study as a foundation to illustrate our testing procedures as suggested by Pritchard and Tebbs (2011b).

CHAPTER FOUR

RESULTS AND DISCUSSIONS

4.1 Overview of Chapter Four

This chapter presents a negative binomial model in group testing that incorporates imperfect tests and the retesting of a positive pool. The entire chapter is arranged as follows: Section 4.2 describes the proposed model, and Section 4.3 describes the formulation of the model using the indicator function. Section 4.4 presents the behaviour of $\pi^*(p)$ against p. Section 4.5 presents the method of Maximum Likelihood that has been used to obtain the point estimators while section 4.6 examines the characteristics of the estimator. The results of the properties of the estimator are presented in section 4.7. The confidence interval and coverage probabilities are outlined in Sections 4.8 and 4.9 respectively. Section 4.10 presents the comparison of the proposed model with the one-stage negative binomial group testing model when imperfect tests are used. The results of this research are applied to WNV prevalence estimation data in Section 4.11. Lastly, the discussion of the results is presented in section 4.12.

4.2 The Proposed Model

Suppose that $G_i = g_i$ denotes the number of pools that are sequentially tested until the first positive pool is detected on a retest and $G_1, G_2, \dots, G_r$ are observed to contain the r^{th} pool that tests positive on a retest. Therefore G_i has a geometric distribution and the overall number of pools that are sequentially tested to obtain r positive pools on a retest is denoted as $x = \sum_{i=1}^{r} G_i$. It is assumed that the pools are of equal size. If the size of the pool is denoted as k and the prevalence of a disease is denoted as p, then the sufficient statistic $x = \sum_{i=1}^{r} G_i$ follows a negative binomial distribution with waiting parameters r and success probability $\pi^*(p)$. The model is diagrammatically described below:

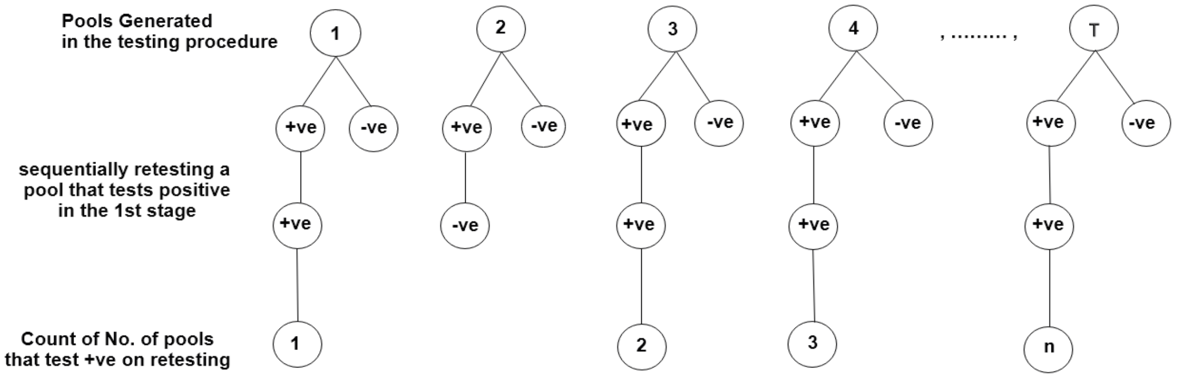


Figure 4. 1: Diagrammatic description of the Retesting model

4.3 Indicator Function

Define the indicator function as follows;

$$T_i = \begin{cases} 1, & if\ the\ i^{th}\ group\ is\ tests\ positive. \\ 0, & otherwise. \end{cases}$$

$$T^*{}_i = \begin{cases} 1, & if\ the\ i^{th}\ group\ test\ positive\ on\ the\ retest \\ \\ 0, & otherwise. \end{cases} \qquad (25)$$

$$D_i = \begin{cases} 1, & if\ the\ i^{th}\ group\ is\ truly\ positive. \\ 0, & otherwise. \end{cases}$$

If the sensitivity and the specificity are defined as $\pi_1 = \Pr\,(X_i = 1|\,D_i = 1)$ and $\pi_0 = \Pr(X_i = 1\,|D_i = 0)$ respectively. The probability that a pool tests positive on retest $i.e\ \pi^*(p)$ is

$$\begin{aligned} \pi^*(p) &= Pr(X^*{}_i = 1\ X_i = 1) \\ &= Pr(X^*{}_i = 1\ X_i = 1\ D_i = 0\ or\ D_i = 1) \\ &= \Pr\,(X^*{}_i = 1\ X_i = 1\ D_{in} = 0\,) + \ Pr(X^*{}_i = 1\ X_i = 1\ D_{in} = 1) \\ &= \Pr(X^*{}_i = 1|D_i = 0) \times \Pr\,(X_i = 1|D_i = 0) \times \Pr(D_i = 0) \\ &\qquad + \Pr(X^*{}_i = 1|D_i = 1) \times \Pr(X_i = 1|D_i = 1) \times \Pr(D_i = 1) \\ &= (1-\pi_0)^2(1-p)^k + \ \pi_1{}^2(1-(1-p)^k) \end{aligned}$$

$$\pi^*(p) = \pi_1{}^2 + (1-p)^k((1-\pi_0)^2 - \pi_1{}^2). \qquad (26)$$

4.4 Characteristics of $\boldsymbol{\pi^*(p)}$ against $\boldsymbol{p}$

In this section, we investigated the relationship between $\pi^*(p)$ and p. This was accomplished by plotting $\pi^*(p)$ against p at different k and when the specificity and sensitivity of a test are known. The specificity and sensitivity values of the tests were established and considered equal throughout the testing procedure.

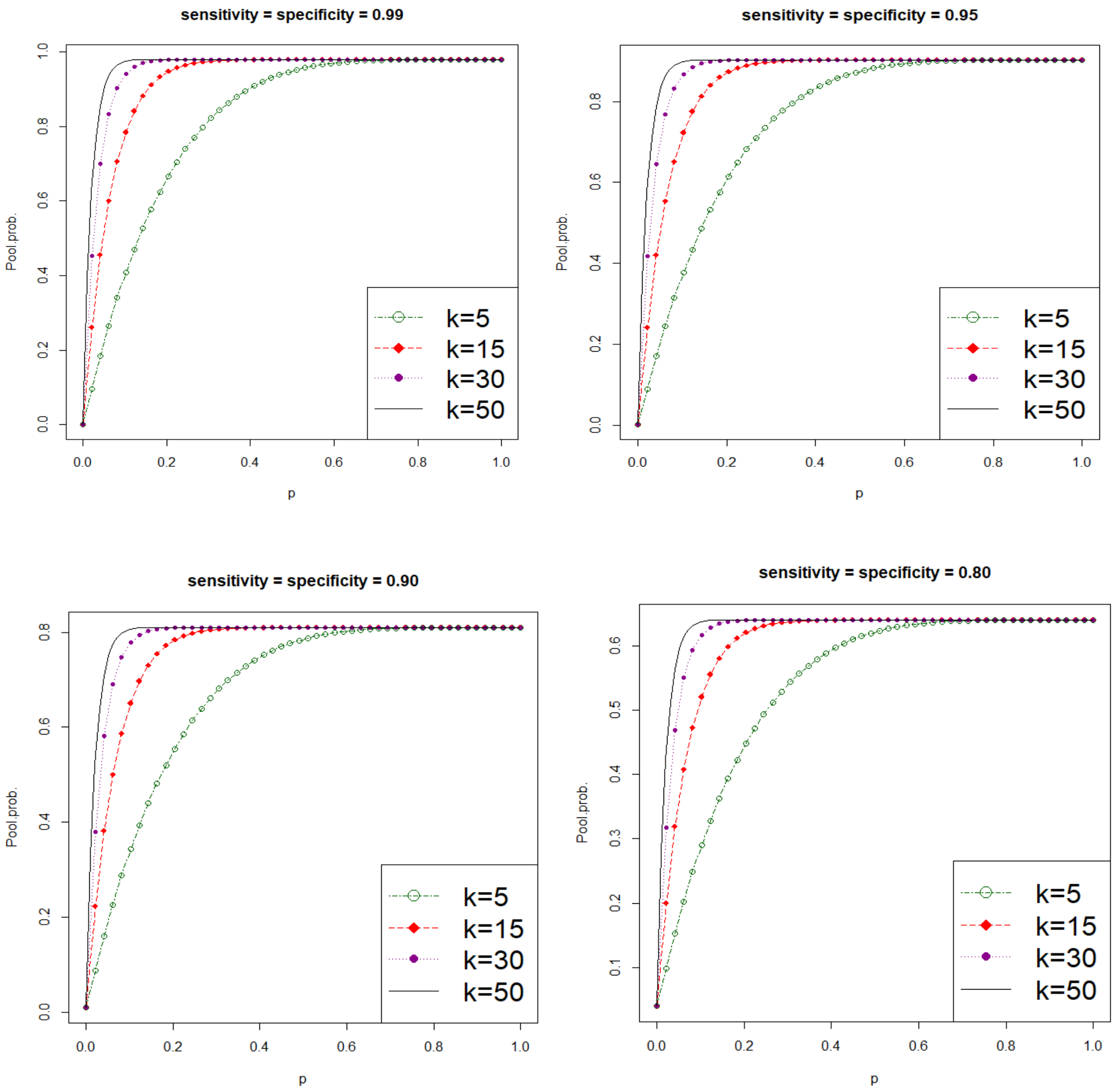


Figure 4. 2: Relationship between $\pi^*(p)$ and p for k = 5, 15, 30, and 50 with Sensitivity= Specificity = 0.99, 0.95, 0.90 and 0.80

Figure (4.2) shows the relationship between $\pi^*(p)$ and p for various group sizes when the accuracy of the assay used is known. The relationship was observed to be monotonic and that $\pi^*(p)$ ranged between 0 and 1. A steep slope was observed at low p values and large group sizes. It was also noted that the probability increased rapidly for low values of p converging towards the sensitivity and specificity value of the tests, especially for high sensitivity and specificity values. Thus, when the accuracy of the tests is known, the probability of detecting a positive pool is sensitive to both the prevalence level and the size of a pool. The relationship was also considered when the accuracy of the test was 80%. The value was selected solely to study the relationship, and this is because the assay used for testing has a higher sensitivity and specificity.

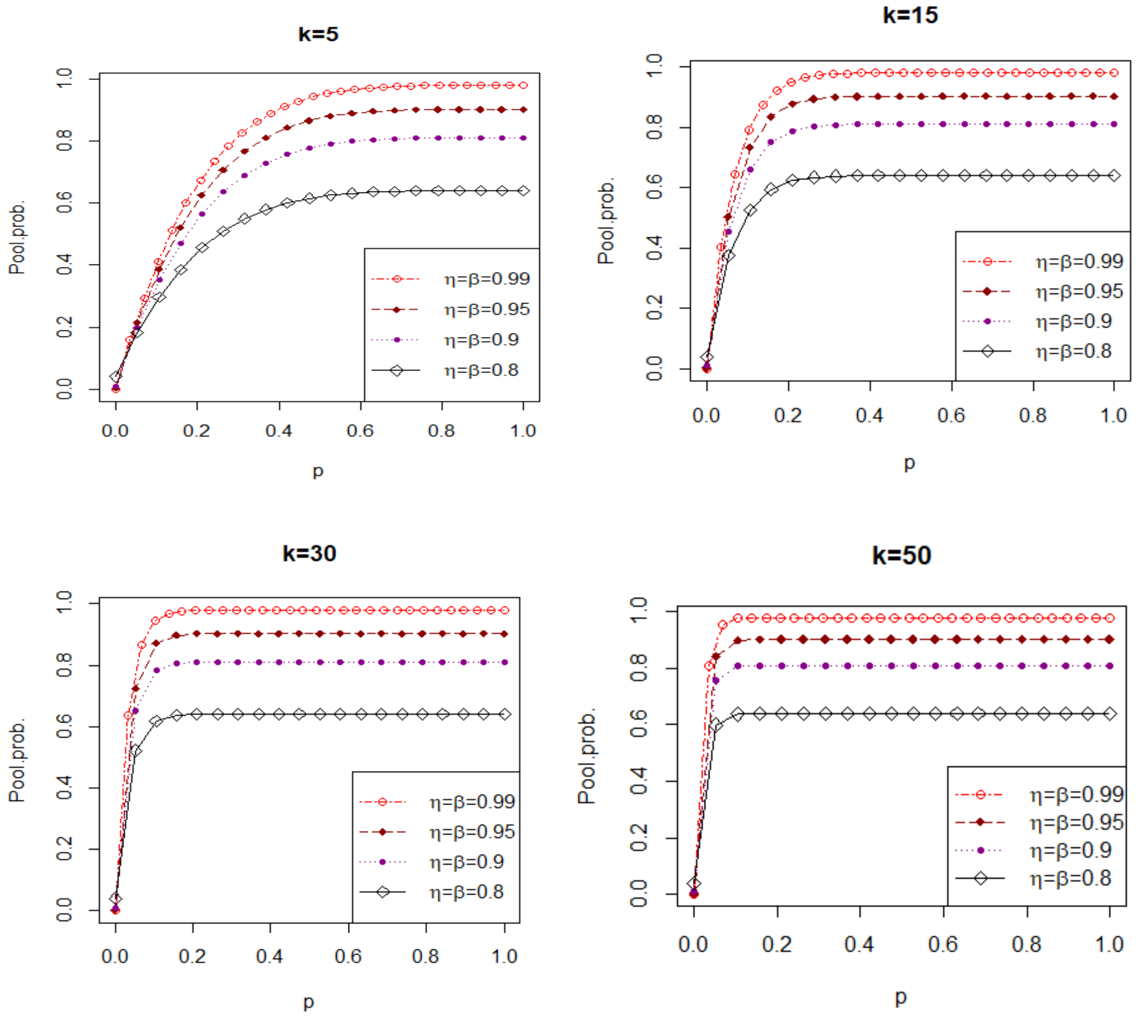


Figure 4. 3: Relationship between $\pi^*(p)$ and p for k = 5, 15, 30, and 50 with Sensitivity(η) = Specificity(β) = 0.99, 0.95, 0.90 and 0.80

Figure (4.3) shows the relationship between the probability of classifying a pool positive on retesting and prevalence, *p* at fixed group sizes. A striking characteristic observed was that the relationship was monotonic, and the gradient of the plots became steeper at low values of *p* as *k* increased. It was sufficient to note that in group testing, both the sensitivity and specificity of the test kits and the size of the group affected the probability of accurately detecting a rare trait. Notably, the probability increased rapidly at low prevalence values to a maximum as *k* increased, after which there was no rate of change. When *k* was fixed, and the prevalence was low, the probability of detecting a trait of interest was highest at high values of the sensitivity and specificity of the test.

4.5 The Maximum Likelihood Estimator

Suppose *x* is the total number of pools tested to obtain *r* pools that test positive on retesting. If *p* is the prevalence of a rare trait, and *k* denotes the size of the group then x follows a negative binomial distribution with parameters $\left(r, \pi^*(p)\right)$.

The model can simply be written as;

$$f(x/p) = \binom{x-1}{r-1}[\pi^*(p)]^r[1-\ \pi^*(p)]^{x-r} \tag{27}$$

In terms of the log-likelihood function, equation (27) can be expressed as;

$$L(p\ /\ x, r) = log\binom{x-1}{r-1} + r\log\pi^*(p) + (x-r)\log[1-\pi^*(p)] \tag{28}$$

The first-order derivative of the log maximum likelihood is

$$\frac{\partial}{\partial p}L(p) = \left(\frac{r}{\pi^*(p)} - \frac{(x-r)}{1-\pi^*(p)}\right)\left(\frac{\partial}{\partial p}\pi^*(p)\right)$$

To obtain the MLE of *p*, we solve the equation $\frac{\partial}{\partial p}L(p) = 0$ which reduces to

$$\frac{r}{\pi^*(p)} = \frac{x-r}{1-\pi^*(p)}$$

$$\pi^*(p) = \frac{r}{x} \tag{29}$$

Substituting the value of $\pi^*(p) = {\pi_1}^2 + (1-p)^k((1-\pi_0)^2 - {\pi_1}^2)$ into equation (29) and solving for $\hat{p}$ we obtain

$${\pi_1}^2 + (1-p)^k((1-\pi_0)^2 - {\pi_1}^2) = \frac{r}{x} \tag{30}$$

$$(1-p)^k((1-\pi_0)^2 - {\pi_1}^2) = \frac{r}{x} - {\pi_1}^2$$

$$(1-p)^k = \left\{\frac{\frac{r}{x} - {\pi_1}^2}{(1-\pi_0)^2 - {\pi_1}^2}\right\} \tag{31}$$

$$(1-p) = \left\{\frac{\frac{r}{x} - {\pi_1}^2}{(1-\pi_0)^2 - {\pi_1}^2}\right\}^{\frac{1}{k}}$$

$$\hat{p} = 1 - \left\{ \frac{{\pi_1}^2 - \frac{r}{x}}{{\pi_1}^2 - (1 - \pi_0)^2} \right\}^{\frac{1}{k}} . \tag{32}$$

To compute the MLE, Monte Carlo simulation was used, and the R code implementing the simulation study is annexed in Appendix C.

Table 4. 1: **MLE of p for $k = 5, 10, 30, 50$ and $r = 1, 3, 5, 10, 15$ with Sensitivity = Specificity = 99%**

Sensitivity=Specificity=0.99					
			r		
p	1	3	5	10	15
			$k = 5$		
0.005	0.037083	0.007335	0.006203	0.005540	0.005367
0.01	0.072883	0.014681	0.012411	0.011042	0.010696
0.05	0.277838	0.077258	0.062284	0.055110	0.053197
0.10	0.454758	0.175649	0.128055	0.110799	0.106557
0.20	0.697056	0.419199	0.306039	0.230217	0.215374
0.30	0.837360	0.643550	0.526881	0.392980	0.347869
			$k = 10$		
0.005	0.060468	0.007355	0.006216	0.005534	0.005351
0.01	0.108036	0.015609	0.012384	0.011074	0.010694
0.05	0.421494	0.116445	0.069269	0.055804	0.053448
0.10	0.658052	0.330505	0.199350	0.120100	0.108827
0.20	0.885850	0.709192	0.587668	0.403237	0.309704
0.30	0.957030	0.879560	0.815683	0.696210	0.613936
			$k = 30$		
0.005	0.143000	0.010042	0.006235	0.005524	0.005344
0.01	0.267490	0.027645	0.013566	0.011109	0.010681
0.05	0.774480	0.475560	0.302159	0.120312	0.070893
0.10	0.943910	0.833082	0.740952	0.569289	0.442061
0.20	0.978910	0.940975	0.903185	0.824583	0.754329
0.30	0.980270	0.944985	0.908833	0.834533	0.766407
			$k = 50$		
0.005	0.228030	0.015558	0.006729	0.005549	0.005339
0.01	0.394580	0.067095	0.019821	0.011208	0.010726
0.05	0.911220	0.746833	0.620492	0.394786	0.257628
0.10	0.975650	0.928954	0.883985	0.787707	0.704396
0.20	0.980080	0.944023	0.906544	0.828924	0.756956
0.30	0.980080	0.944120	0.906641	0.829114	0.756958

It was observed from Table (4.1) that as prevalence increased, so did the MLE for any fixed value of *r* and *k*. When the group size and the waiting parameter *r* were large, low values of *p* resulted in a small MLE. Further, high values of *k* and *p* were sufficient to overestimate the prevalence even at large values of *r*. Thus, when imperfect tests are used with a known accuracy, one can approximate the proportion p using a combination of r and k values. For example, when the tests are 99% accurate, a close approximation of $p = 0.005$ was observed when $r = 15$ and $k = 50$.

Table 4. 2: MLE of p for $k = 5, 10, 30, 50$ and $r = 1, 3, 5, 10, 15$ with Sensitivity = Specificity = 95%

Sensitivity=Specificity=0.95					
	r				
p	1	3	5	10	15
$k = 5$					
0.005	0.038622	0.007610	0.006360	0.005603	0.005395
0.01	0.071528	0.015042	0.012561	0.011117	0.010732
0.05	0.265143	0.077676	0.063117	0.055301	0.053403
0.10	0.431582	0.171177	0.129080	0.111386	0.106820
0.20	0.654666	0.380445	0.289774	0.242944	0.220649
0.30	0.783361	0.563529	0.462084	0.435045	0.364367
$k = 10$					
0.005	0.059278	0.007505	0.006288	0.005569	0.005385
0.01	0.104496	0.015570	0.012492	0.011102	0.010694
0.05	0.396336	0.109971	0.068458	0.056371	0.053830
0.10	0.613479	0.283287	0.178538	0.132960	0.112020
0.20	0.819533	0.590752	0.454833	0.457813	0.327689
0.30	0.888036	0.720806	0.613523	0.682071	0.550181
$k = 30$					
0.005	0.131180	0.009665	0.006296	0.005564	0.005354
0.01	0.249600	0.025059	0.013159	0.011157	0.010702
0.05	0.715270	0.382271	0.216061	0.168315	0.082148
0.10	0.869140	0.663005	0.520282	0.581617	0.399407
0.20	0.908870	0.746513	0.625290	0.725650	0.573274
0.30	0.909650	0.748814	0.628502	0.729691	0.577710
$k = 50$					
0.005	0.211400	0.014051	0.006738	0.005566	0.005356
0.01	0.366630	0.056549	0.018093	0.011475	0.010757
0.05	0.835370	0.592239	0.423074	0.454208	0.265026
0.10	0.903890	0.732667	0.602974	0.702818	0.539048
0.20	0.908720	0.743264	0.617231	0.721335	0.560229
0.30	0.908720	0.743360	0.617327	0.721430	0.560235

Scrutiny of Table (4.2) showed the performance of the MLE when the assay used for testing was 95% accurate. When k and r are sufficiently large, the MLE is small for low values of p. Conversely, when large group sizes and a sufficiently large waiting parameter r are used, the MLE overestimated the prevalence at high values of the proportion p. Furthermore, it was observed that when $r = 1$ and $k = 50$, the MLE increased to a maximum value as the p increased, and afterward, there was no rate of change observed. Finally, it was noted that at any fixed values of k and p, the MLE decreased as the waiting parameter r increased.

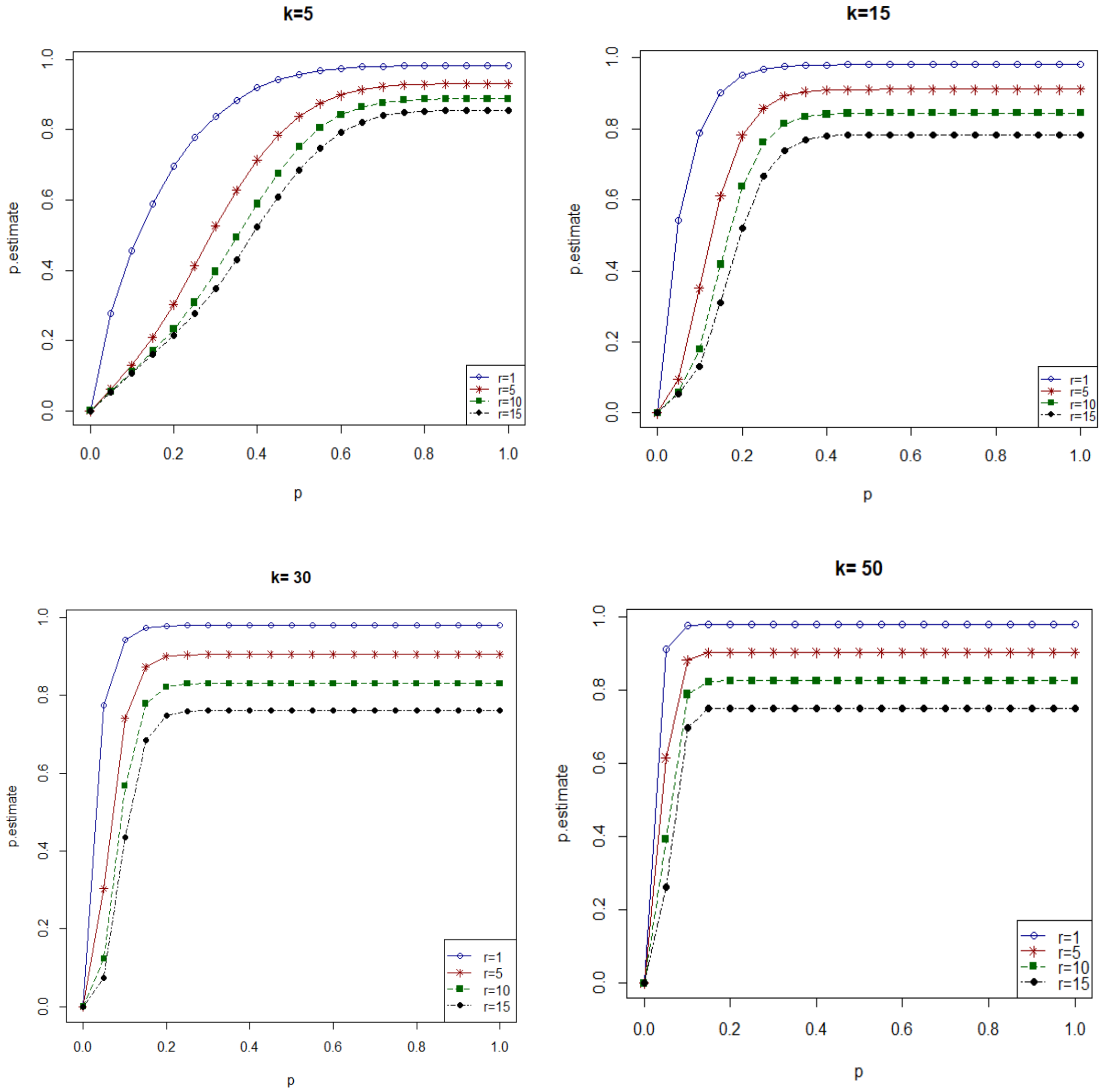
Table 4. 3: MLE of p for $k = 5, 10, 30, 50$ and $r = 1, 3, 5, 10, 15$ with Sensitivity = Specificity = 90%

Sensitivity= Specificity= 0.90					
			r		
p	1	3	5	10	15
			$k = 5$		
0.005	0.049307	0.008527	0.006815	0.005810	0.005531
0.01	0.076926	0.016140	0.013089	0.011332	0.010878
0.05	0.250348	0.079561	0.064613	0.055680	0.053480
0.10	0.409826	0.170627	0.137896	0.113133	0.107650
0.20	0.605089	0.354179	0.332336	0.252703	0.228741
0.30	0.720972	0.505047	0.530204	0.436243	0.390851
			$k = 10$		
0.005	0.063313	0.008169	0.006531	0.005687	0.005447
0.01	0.104177	0.016312	0.012717	0.011231	0.010779
0.05	0.373008	0.102652	0.076881	0.057527	0.054077
0.10	0.561554	0.247224	0.230322	0.139804	0.118663
0.20	0.745903	0.481790	0.549537	0.440285	0.379229
0.30	0.804308	0.581238	0.682959	0.615026	0.574926
			$k = 30$		
0.005	0.125410	0.009260	0.006489	0.005617	0.005383
0.01	0.234050	0.023474	0.015182	0.011231	0.010771
0.05	0.646420	0.297767	0.330595	0.172557	0.113522
0.10	0.782640	0.500725	0.620685	0.512636	0.452614
0.20	0.814220	0.562920	0.694645	0.623274	0.582967
0.30	0.815090	0.564349	0.697481	0.626114	0.587064
			$k = 50$		
0.005	0.193200	0.012956	0.008029	0.005616	0.005377
0.01	0.339180	0.046552	0.026416	0.011909	0.010857
0.05	0.752080	0.439925	0.546659	0.412779	0.334123
0.10	0.809120	0.544807	0.679821	0.600676	0.555674
0.20	0.812950	0.552098	0.690189	0.614254	0.571822
0.30	0.812950	0.552098	0.690189	0.614254	0.571822

Scrutiny of Table (4.3) shows that when the accuracies of the assays are 90% accurate, the MLE was observed to be a monotone increasing function of *r*, *k*, and *p*. Large group sizes produced close approximations of *p* when the waiting parameter *r* was large. Generally, when the waiting parameter *r* was small, the MLE tended to overestimate the prevalence level, even at large group sizes. Therefore, to obtain a close approximation of *p*, both the waiting parameter *r* and group size *k* have to be relatively large.

4.6 Characteristic of $\widehat{p}$

In this section, the behaviour of the MLE was investigated by plotting $\hat{p}$ against *p* at different group sizes by varying the waiting parameter *r* when the sensitivity and specificity of the tests were set at 0.99 as illustrated below.



\

Figure 4. 4: Relationship between $\widehat{p}$ and *p* for *k* = 5, 15, 30, 50 and *r* = 2, 5, 10 and 15 when Sensitivity = Specificity = 0.99

Figure (4.4) illustrates the relationship between the MLE and the prevalence level at different waiting parameters *r* when the group sizes are fixed. It was observed that when the accuracy of the assay was 99%, the relationship was observed to be monotonic. The relationship increased to a maximum and afterward, there was no observable rate of change. Secondly, as the group size increased, a steep gradient was observed at low values of the prevalence when the waiting parameter *r* was small. The plots depict that the MLE overestimated the prevalence and it confirmed previous studies that the MLE is positively biased, especially at high prevalence.

The relationship between the MLE and the prevalence level was also considered at different group sizes, and different waiting parameters *r* when the sensitivity and specificity of the tests are set at 99% as presented in Figures (4.5) below

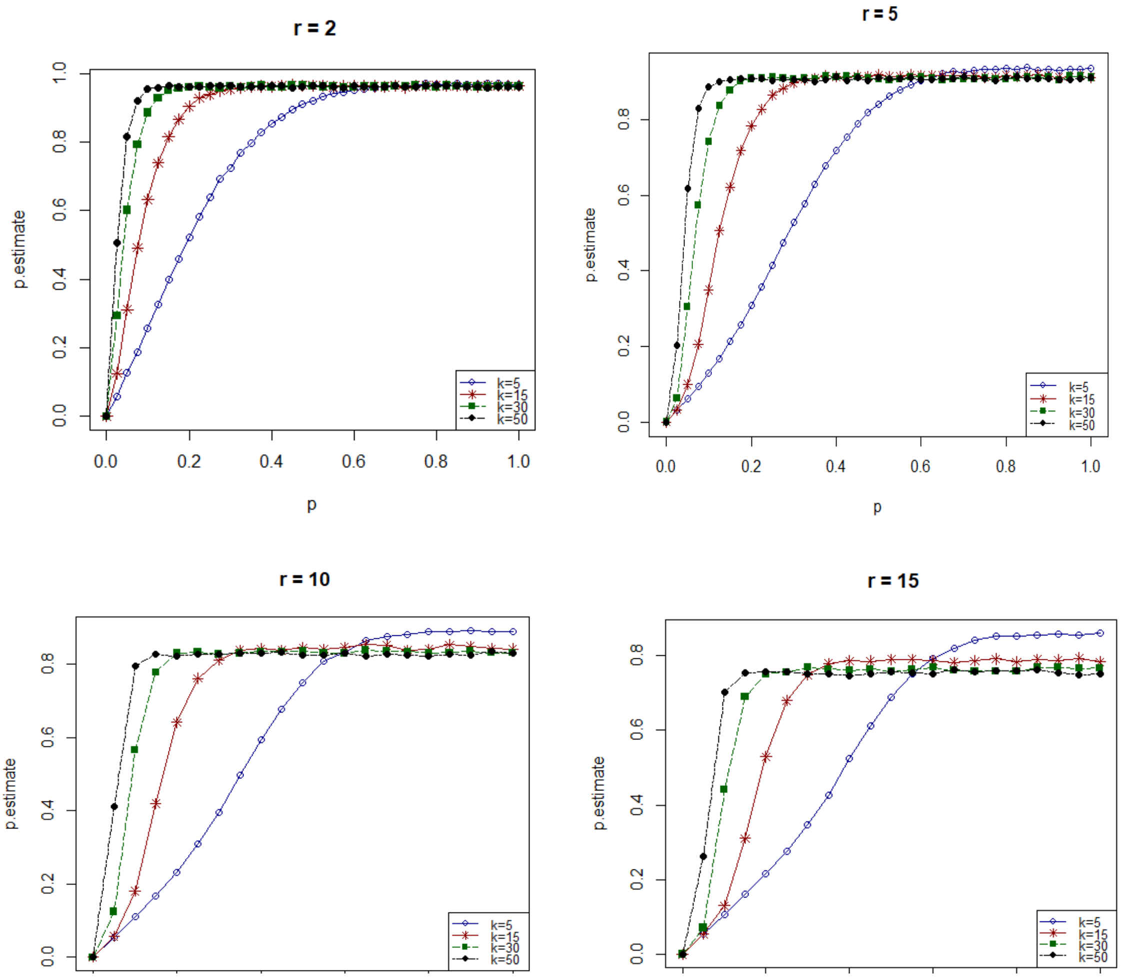


Figure 4. 5: Relationship between $\hat{p}$ and p for $r = 2, 5, 10, 15$ and $k = 5, 15, 30, 50$ with Sensitivity = Specificity = 0.99

When the assays used are 99% accurate, it was observed that the MLE is a non-decreasing function of the proportion *p*, group size *k*, and the waiting parameter *r*. The MLE increased to a maximum as the prevalence increased, and afterward, no rate of change was observed. Lastly, when the group sizes were large, a steep gradient was observed at a low prevalence level for any predetermined waiting parameter *r*.

The relationship between the MLE and the prevalence level was also investigated for various group sizes and at different sensitivity(η) and specificity(β) values when the waiting parameter $r = 1$ as illustrated in Figures (4.6) below

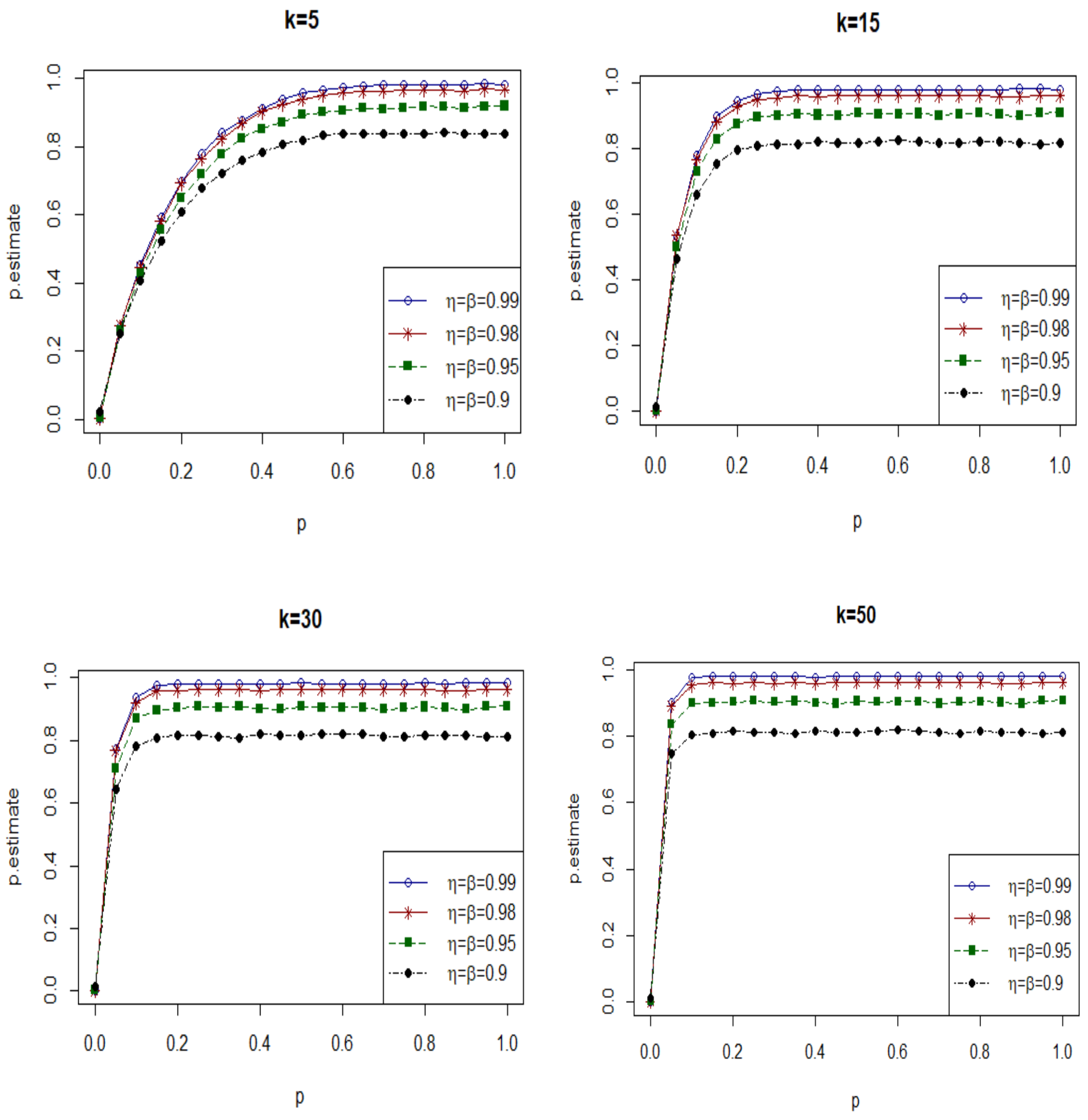


Figure 4. 6: Relationship between $\hat{p}$ and p for $k = 5, 15, 30, 50$ and $r = 1$ with sensitivity = specificity = 0.99, 0.98, 0.95, 0.90

Figure (4.6) showed that at a certain waiting parameter *r*, a steep gradient appeared at low prevalence values as the group sizes increased. It was also observed that the relationship between MLE and the proportion *p* increased monotonically regardless of the test accuracy. The MLE was observed to increase to a maximum and afterward, no rate of change was observed. Lastly, it was observed that the MLE is affected by group size, and prevalence level, assuming the accuracy of the testing assay is known. Lastly, as p increased, high sensitivity and specificity values were associated with high MLE values.

Lastly, the relationship between the MLE and the prevalence level was investigated at various waiting parameters r and known sensitivity(η) and specificity(β) values of the tests when the group size $k = 5$ as illustrated in the figures below.

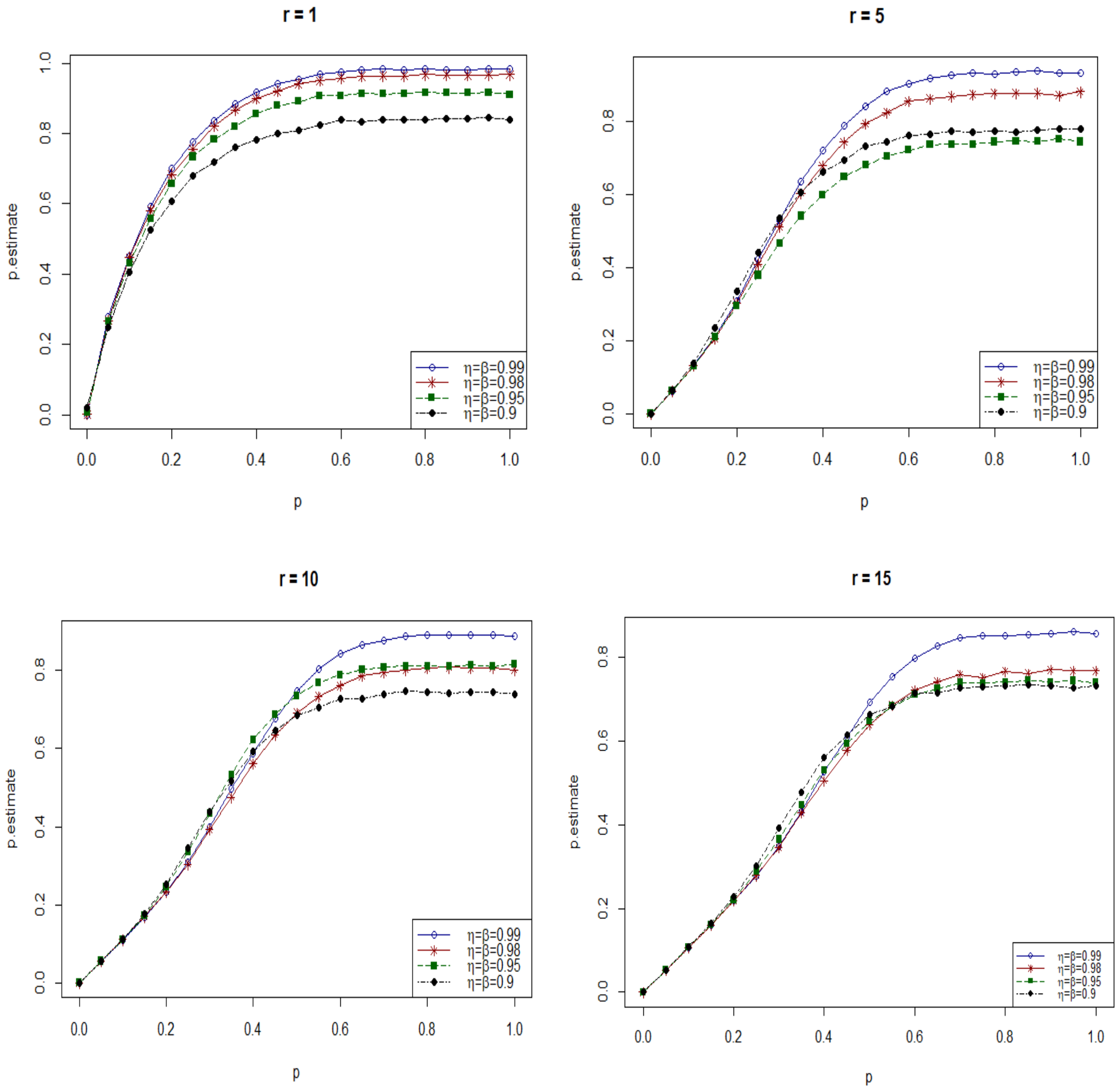


Figure 4. 7: Plots of $\widehat{p}$ against p for $r = 1, 5, 10, 15$ and sensitivity = specificity = 0.99, 0.98, 0.95 and 0.90 with k = 5

Figure (4.7) exhibits the relationship between the MLE and the proportion p at fixed group sizes. It was observed that as the value of r increased, the prevalence level in the interval $0 < p < 0.6$ showed an almost linear relationship. As r increased, it was observed that the behaviour of the gradient became more gradual. High sensitivity and specificity resulted in higher MLE values as prevalence increased. In general, the MLE increases as the group size, prevalence level, and waiting parameter r increase, assuming that the accuracy of the assay used for testing is known.

4.7 Properties of the Estimator

In this section, the properties of the estimator such as biasedness, asymptotic variance, and mean squared error of the estimator were considered.

4.7.1 Biasedness of the estimator

The MLE of the binomial model was shown to be unbiased when $k = 1$, but for $k \geq 1$, Jensen's inequality was used to show that the estimator overestimated the prevalence (Muhua, 2010). Jensen's inequality states that if a function of $f(t)$ is a convex function, then $E[f(t)] \geq f[E(t)]$, and if $f(t)$ is a concave function, then $E[f(t)] \leq f[E(t)]$, provided that the expectations exist and are finite (Billingsley, 1995).

A similar approach using Jensen's inequality was used to prove that the MLE overestimated the prevalence level for $k > 1$.

Preposition 1

The value of the estimator $\hat{p}$ given in equation (32) overestimates the prevalence of a rare trait when k is greater than one $(k > 1)$.

Proof

To prove this proposition, Jensen's inequality was used.

Approach 1:

From equation (32), if it is assumed that

$$\hat{p} = 1 - \left\{\frac{{\pi_1}^2 - \frac{r}{x}}{{\pi_1}^2 - (1-\pi_0)^2}\right\}^{\frac{1}{k}}$$

is continuous. Then

$$\frac{\partial \hat{p}}{\partial x} = -\frac{r}{kx^2}\left\{\frac{{\pi_1}^2 - \frac{r}{x}}{{\pi_1}^2 - (1-\pi_0)^2}\right\}^{\frac{1-k}{k}} \times \left(\frac{1}{{\pi_1}^2 - (1-\pi_0)^2}\right) \tag{33}$$

and

$$\frac{\partial^2\hat{p}}{\partial x^2} = \frac{1}{kx^3}\times\left(\frac{2r}{{\pi_1}^2-(1-\pi_0)^2}\right)\times\left\{\frac{{\pi_1}^2-\frac{r}{x}}{{\pi_1}^2-(1-\pi_0)^2}\right\}^{\frac{1-k}{k}}$$
$$+\frac{r^2}{kx^4}\left(\frac{k-1}{k}\right)\times\left(\frac{1}{{\pi_1}^2-(1-\pi_0)^2}\right)^2 \tag{34}$$
$$\times\left\{\frac{{\pi_1}^2-\frac{r}{x}}{{\pi_1}^2-(1-\pi_0)^2}\right\}^{\frac{1-2k}{k}} \geq 0.$$

Hence, $\hat{p}$ is convex as long as $k > 1$ and ${\pi_1}^2 \geq \frac{r}{x}$.

Since $\hat{p}$ is a function of x, say $\phi(x)$, then by Jensen's inequality,

$$E[\phi(x)] \geq \phi[E(x)].$$

Therefore, under this condition

$$E(\hat{p}) = E\left\{1-\left[\frac{{\pi_1}^2-\frac{r}{x}}{{\pi_1}^2-(1-\pi_0)^2}\right]^{\frac{1}{k}}\right\}$$

$$E(\hat{p}) \geq 1-E\left[\frac{{\pi_1}^2-\frac{r}{x}}{{\pi_1}^2-(1-\pi_0)^2}\right]^{\frac{1}{k}}. \tag{35}$$

Since $E\left(\frac{1}{x}\right) \geq \left(\frac{1}{E[x]}\right)$ by Jensen inequality, then

$$E(\hat{p}) \geq 1-\left[\frac{{\pi_1}^2-\frac{r}{E[x]}}{{\pi_1}^2-(1-\pi_0)^2}\right]^{\frac{1}{k}}. \tag{36}$$

Substituting $E[x] = \frac{r}{\pi^*(p)}$ where $\pi^*(p) = {\pi_1}^2 + (1-p)^k((1-\pi_0)^2 - {\pi_1}^2)$, into equation (36) it reduces to

$$E(\hat{p}) \geq 1-\left[\frac{({\pi_1}^2-(1-\pi_0)^2)(1-p)^k}{{\pi_1}^2-(1-\pi_0)^2}\right]^{\frac{1}{k}}$$
$$E(\hat{p}) \geq p. \tag{37}$$

Hence for $k > 1$ the estimator $\hat{p}$ overestimates p.

Approach 2:

If we let,

$$y(x) = \left\{ \frac{{\pi_1}^2 - \frac{r}{x}}{{\pi_1}^2 - (1 - \pi_0)^2} \right\}^{\frac{1}{k}} \tag{38}$$

$$\frac{\partial y(x)}{\partial x} = \frac{r}{kx^2} \left(\frac{{\pi_1}^2 - \frac{r}{x}}{{\pi_1}^2 - (1 - \pi_0)^2} \right)^{\frac{1-k}{k}} \times \left(\frac{1}{{\pi_1}^2 - (1 - \pi_0)^2} \right). \tag{39}$$

$$\frac{\partial^2 y(x)}{\partial x^2} = \left\{ \frac{r^2}{kx^4} \times \left(\frac{1-k}{k} \right) \times \left(\frac{{\pi_1}^2 - \frac{r}{x}}{{\pi_1}^2 - (1 - \pi_0)^2} \right)^{\frac{1-2k}{k}} \times \left(\frac{1}{({\pi_1}^2 - (1 - \pi_0)^2)^2} \right)^2 \right\} \tag{40}$$

$$- \frac{2r}{kx^3} \times \left(\frac{{\pi_1}^2 - \frac{r}{x}}{{\pi_1}^2 - (1 - \pi_0)^2} \right)^{\frac{1-k}{k}} \times \left(\frac{1}{{\pi_1}^2 - (1 - \pi_0)^2} \right) \leq 0.$$

Therefore, $\mathrm{y}(x)$ is concave for $k > 1$ and ${\pi_1}^2 > \frac{r}{x}$.

Thus, by Jensen inequality $E[y(x)] \leq y[E(x)]$,

$$E \left\{ \frac{{\pi_1}^2 - \frac{r}{x}}{{\pi_1}^2 - (1 - \pi_0)^2} \right\}^{\frac{1}{k}} \leq \left\{ \frac{{\pi_1}^2 - \frac{r}{E(x)}}{{\pi_1}^2 - (1 - \pi_0)^2} \right\}^{\frac{1}{k}} \tag{42}$$

$$1 - E \left\{ \frac{{\pi_1}^2 - \frac{r}{x}}{{\pi_1}^2 - (1 - \pi_0)^2} \right\}^{\frac{1}{k}} \geq 1 - \left\{ \frac{{\pi_1}^2 - \frac{r}{E(x)}}{{\pi_1}^2 - (1 - \pi_0)^2} \right\}^{\frac{1}{k}}. \tag{43}$$

Since $E\left(\frac{1}{x}\right) \geq \left(\frac{1}{E[x]}\right)$ by Jensen Inequality, and $E(x) = \frac{r}{\pi^*(p)}$, where $\pi^*(p) = {\pi_1}^2 + (1-p)^k((1-\pi_0)^2 - {\pi_1}^2)$ substituting in equation (38) simplifies to

$$E(\hat{p}) \geq \left\{ \frac{({\pi_1}^2 - (1 - \pi_0)^2)(1-p)^k}{{\pi_1}^2 - (1 - \pi_0)^2} \right\}^{\frac{1}{k}} \tag{44}$$

$$\geq 1 - (1 - p)$$

$$E(\hat{p}) \geq p. \tag{45}$$

Hence for $k > 1$ the estimator $\hat{p}$ overestimates p.

Approach 3:

The expectation of equation (32) can be expressed as

$$E(\hat{p}) = 1 - E\left[\left(\frac{1}{{\pi_1}^2 - (1-\pi_0)^2}\right)\left({\pi_1}^2 - \frac{r}{x}\right)\right]^{\frac{1}{k}}. \tag{46}$$

Let $E(\hat{p}) = 1 - E[y(x)]$

where $y(x) = \left[\left(\frac{1}{{\pi_1}^2 - (1-\pi_0)^2}\right)\left({\pi_1}^2 - \frac{r}{x}\right)\right]^{\frac{1}{k}}.$

Then, $$\frac{\partial y(x)}{\partial x} = \frac{r}{kx^2}\left[\frac{{\pi_1}^2 - \frac{r}{x}}{{\pi_1}^2 - (1-\pi_0)^2}\right]^{\frac{1-k}{k}} \times \left(\frac{1}{{\pi_1}^2 - (1-\pi_0)^2}\right). \tag{47}$$

$$\frac{\partial^2 y(x)}{\partial x^2} = \left\{\frac{r^2}{kx^4} \times \left(\frac{1-k}{k}\right) \times \left(\frac{{\pi_1}^2 - \frac{r}{x}}{{\pi_1}^2 - (1-\pi_0)^2}\right)^{\frac{1-2k}{k}} \times \left(\frac{1}{{\pi_1}^2 - (1-\pi_0)^2}\right)^2\right\} \tag{48}$$

$$- \frac{2r}{kx^3} \times \left(\frac{{\pi_1}^2 - \frac{r}{x}}{{\pi_1}^2 - (1-\pi_0)^2}\right)^{\frac{1-k}{k}} \times \left(\frac{1}{{\pi_1}^2 - (1-\pi_0)^2}\right) \leq 0.$$

$\text{Therefore, y(t) is concave for } k > 1 \text{ and } {\pi_1}^2 \geq \frac{r}{x}.$

Therefore, by Jensen's inequality $E[y(x)] \leq y[E(x)]$,

$$E\left[\left(\frac{1}{{\pi_1}^2 - (1-\pi_0)^2}\right)\left({\pi_1}^2 - \frac{r}{x}\right)\right]^{\frac{1}{k}} \leq \left[\left(\frac{1}{{\pi_1}^2 - (1-\pi_0)^2}\right)\left({\pi_1}^2 - \frac{r}{E(x)}\right)\right]^{\frac{1}{k}} \tag{49}$$

Substituting $E[T] = \frac{r}{\pi^*(p)}$ where $\pi^*(p) = {\pi_1}^2 + (1-p)^k((1-\pi_0)^2 - {\pi_1}^2)$ equation (44) reduces to

$$E\left[\left(\frac{1}{{\pi_1}^2 - (1-\pi_0)^2}\right)\left({\pi_1}^2 - \frac{r}{x}\right)\right]^{\frac{1}{k}} \leq \left\{\frac{(1-p)^k({\pi_1}^2 - (1-\pi_0)^2)}{{\pi_1}^2 - (1-\pi_0)^2}\right\}^{\frac{1}{k}}.$$

Which implies that

$$-E\left[\left(\frac{1}{{\pi_1}^2 - (1-\pi_0)^2}\right)\left({\pi_1}^2 - \frac{r}{x}\right)\right]^{\frac{1}{k}} \geq -(1-p). \tag{50}$$

$$1 - E\left[\left(\frac{1}{{\pi_1}^2 - (1-\pi_0)^2}\right)\left({\pi_1}^2 - \frac{r}{x}\right)\right]^{\frac{1}{k}} \geq 1 - (1-p). \tag{51}$$

Hence,

$$E(\hat{p}) \geq p\,. \tag{52}$$

Thus for $k > 1$ and ${\pi_1}^2 \geq \frac{r}{x},\ \hat{p}$ overestimates p.

The bias of an estimator measures the average error incurred when using the estimate of a parameter. It was pointed out that bias was particularly useful in evaluating point estimates (Thompson, 1962). The exact bias of an estimator is given by:

$$bias(\hat{p}) = E(\hat{p}) - p. \tag{53}$$

Because x~Negative Binomial $(r, \pi^*(p))$ where $\pi^*(p) = {\pi_1}^2 + (1-p)^k((1-\pi_0)^2 - {\pi_1}^2)$. The exact bias and the MSE can be expressed as an infinite sum. It was noted that the sums do not reduce to anything tractable. Thus, the bias and the MSE of an estimator $\hat{p}$ were suggested by Pritchard and Tebbs (2011b) to be approximated as shown in equation (21) given as follows

$$Bias(\hat{p}) = \sum_{x=r}^{x^*} (\hat{p} - p)\binom{x-1}{r-1}\pi^*(p)^r(1-\pi^*(p))^{x-r}$$

Since the sums do not reduce to anything tractable Monte Carlo simulations were invoked and the R code implementing the simulation study that computes the Bias Tables is annexed in Appendix C.

Table 4. 4: Bias of $\hat{p}$ for various values of p with $r = 1, 3, 5, 10, 15$ and $k = 5, 10, 30, 50$ at sensitivity = specificity = 99%

Bias $\times 10^{-4}$					
			r		
p	1	3	5	10	15
			$k = 5$		
0.005	362.80610	37.39135	23.41135	14.38200	11.23476
0.01	653.08290	74.66184	46.54191	28.62074	22.36416
0.05	2371.60600	418.93271	231.33740	140.27249	109.71514
0.10	3840.02890	1037.38995	511.51869	279.00437	217.11517
0.20	5573.34120	2682.07576	1513.74906	650.96485	456.72240
0.30	6035.75530	4163.31983	2864.88066	1455.92420	925.87968
0.40	5685.07450	4860.81024	4057.69919	2519.13300	1829.42647
			$k = 10$		
0.005	567.82770	38.23104	23.37473	14.36144	11.21754
0.01	1048.95920	80.94800	46.73266	28.64085	22.37064
0.05	3856.78190	821.10962	309.72758	145.44342	112.11377
0.10	5901.92240	2564.13220	1247.39388	391.76074	247.93661
0.20	7167.71460	5611.21106	4330.80178	2387.50558	1421.94666
0.30	6777.53620	6278.09096	5776.61834	4625.86607	3656.79622
0.40	5930.78080	5748.91242	5547.54441	5027.30965	4520.02533
			$k = 30$		
0.005	1407.68200	61.49968	24.03181	14.39814	11.23803
0.01	2572.09800	228.50082	57.97831	29.02309	22.58504
0.05	7379.39100	4393.14376	2659.68468	818.70792	315.13048
0.10	8494.38800	7535.94242	6669.32861	4886.44017	3558.92440
0.20	7868.38400	7599.10928	7331.94058	6691.10470	6096.45891
0.30	6915.64500	6738.41412	6559.49977	6122.70470	5709.80285
0.40	5955.53000	5855.73383	5751.20042	5487.76257	5231.61935
			$k = 50$		
0.005	2191.00100	135.81610	28.52969	14.49044	11.28822
0.01	3858.27000	625.75440	131.99412	30.48499	23.01542
0.05	8629.29400	7095.11760	5821.19533	3537.86433	2194.31129
0.10	8796.89600	8395.50360	8007.35776	7103.89562	6294.55529
0.20	7877.70300	7632.13780	7390.91512	6816.34526	6285.50365
0.30	6917.58000	6749.42800	6582.56056	6181.28160	5807.13767
0.40	5957.37300	5866.46570	5773.79640	5545.46300	5327.73594

Table (4.4) showed that for any fixed values of the group size k and the waiting parameter r, the absolute bias of the estimator increased to a maximum as the optimal value of the prevalence was attained, and afterward decreased as the prevalence increased. Conversely, for a given group size and prevalence level, the absolute bias of the estimator decreased as the waiting parameter r increased. The results showed that it was possible to get a set of r and k values that yield a minimum absolute bias at a particular value of the prevalence. For example, when $p = 0.005$, the minimum bias was observed at $r = 15$ and $k = 10$.

Table 4. 5: Bias of $\hat{p}$ for various values of p with $r = 1, 3, 5, 10, 15$ and $k = 5, 10, 30, 50$ at sensitivity = specificity = 95%

Bias $\times 10^{-4}$					
	r				
p	1	3	5	10	15
$k = 5$					
0.005	380.76220	41.52092	25.95154	15.93273	12.44444
0.01	653.68000	79.14075	49.22275	30.22827	23.61278
0.05	2255.26870	419.97179	238.73971	144.13236	112.44601
0.10	3602.94990	978.32673	521.57424	293.12960	225.82892
0.20	5175.96070	2294.87368	1354.88762	808.08288	519.35439
0.30	5677.46480	3303.48420	2321.27912	1906.67545	1137.52123
0.40	5434.48780	4046.78395	2823.63207	2950.41699	1981.60932
$k = 10$					
0.005	560.99250	40.47671	24.74973	15.17049	11.84578
0.01	1008.15300	83.12783	48.40546	29.57685	23.08846
0.05	3593.23780	735.62461	302.91655	153.64969	116.63904
0.10	5458.47350	2124.49065	1032.87077	524.91204	284.51085
0.20	6698.41200	4483.44720	3047.88096	2993.50067	1674.04049
0.30	6418.43550	5219.55184	4137.03897	4614.50812	3137.79043
0.40	5718.11860	5026.75652	4321.83066	4731.85814	3693.31595
$k = 30$					
0.005	1325.23900	58.81119	24.68150	14.79048	11.52974
0.01	2396.62500	201.03076	57.00882	29.91896	23.19216
0.05	6817.86500	3479.66278	1848.61660	1315.15555	442.30362
0.10	7881.07700	5984.48217	4502.89514	5074.04357	3235.39221
0.20	7383.34200	6252.38117	5276.53975	6002.40539	4670.04657
0.30	6585.20200	5799.59064	5103.11115	5617.62121	4654.29024
0.40	5780.27900	5329.62519	4905.94283	5201.99972	4597.72182
$k = 50$					
0.005	2042.00100	118.19540	28.08787	14.90331	11.57584
0.01	3573.01200	513.19890	108.85200	33.54519	23.92129
0.05	7974.66300	5579.64670	3875.45159	4143.91552	2256.64394
0.10	8167.33000	6701.98660	5484.81003	6367.58059	473092535
0.20	7400.01500	6325.07340	5414.47400	6117.29080	4889.21737
0.30	6595.08500	5855.0844	5217.26849	5701.61852	4832.57489
0.40	5790.07600	5384.89830	5019.79203	5285.59631	4775.47022

Scrutiny of Table (4.5) showed that when the test kits used were 95% accurate, the absolute bias of the estimator increased to a maximum as the optimal prevalence value was attained and afterward decreased for fixed group sizes and waiting parameter r. However, when the group sizes and the prevalence levels were fixed, the absolute bias decreased to a minimum value as the optimal value of the waiting parameter r was attained, and afterward increased. The results showed that one can estimate the prevalence level at a predetermined optimal value of the waiting parameter r that would register the least bias when the group sizes are known. For example, when $r = 15$ and $k = 30$ the least bias was observed at $p = 0.005$.

Table 4. 6: Bias of $\hat{p}$ for various values of p with $r = 1, 3, 5, 10, 15$ and $k = 5, 10, 30, 50$ with sensitivity and specificity = 90%

Bias $\times 10^{-4}$					
r					
p	1	3	5	10	15
$k = 5$					
0.005	467.90050	56.36712	35.08247	21.51645	16.80186
0.01	714.20770	94.79655	58.57850	35.89432	28.02674
0.05	2163.61020	444.46113	260.33995	152.67300	118.80335
0.10	3349.73070	981.77386	621.57221	315.40921	239.84322
0.20	4694.99310	2105.61612	1829.82767	917.39839	617.72843
0.30	5232.37050	2894.22145	3058.52693	1988.99850	1477.08288
0.40	5112.76240	3044.61997	3735.69578	2865.89693	2393.39554
$k = 10$					
0.005	605.22390	48.60525	29.48837	18.02489	14.06701
0.01	1007.63300	92.46712	53.79461	32.62030	25.43748
0.05	3309.68030	689.20618	403.12087	165.89926	123.97599
0.10	4933.90670	1803.39974	1585.10099	613.01676	354.75481
0.20	6129.29920	3414.35979	4061.27359	2860.10880	2204.29724
0.30	5977.05470	4019.89376	4898.19411	4077.44902	3561.66834
0.40	5450.54620	4153.55514	4796.22757	4277.17188	3927.98009
$k = 30$					
0.005	1268.80300	60.38560	27.75081	15.92456	12.39351
0.01	2223.72000	183.55030	76.01481	31.74067	24.46841
0.05	6153.52500	2636.61550	2929.40045	1398.82986	761.90235
0.10	7146.73100	4396.39780	5549.06096	4428.67280	3783.53477
0.20	6800.14500	4860.69110	5878.30405	5234.381140	4843.02142
0.30	6186.25000	4810.24840	5534.04576	5069.59304	4782.58953
0.40	5566.31800	4747.52150	5174.39042	4881.96047	4693.82582
$k = 50$					
0.005	1902.02900	106.72100	37.78836	15.81898	12.23804
0.01	3262.31600	416.76720	209.42977	37.59929	25.50609
0.05	7194.70200	4056.87530	5118.21118	3748.93108	2947.16756
0.10	7414.01000	4992.51820	6242.69631	5441.75945	4962.77907
0.20	6827.53100	4996.38140	5966.69413	5376.87823	5025.23179
0.30	6207.59300	4933.63780	5607.01724	5189.20952	4936.42050
0.40	5587.58400	4870.75320	5247.16406	5001.28083	4847.28771

Scrutiny of Table (4.6) showed that the absolute bias increases to a maximum value when the tests are 90% accurate, irrespective of the waiting parameter *r* and group size *k*, then decreases as the optimal prevalence value is reached. before decreasing. Also noted was for any fixed value of proportion *p* and the group size *k*, the absolute bias decreased as the optimal value of the waiting parameter *r* increased. Generally, an optimized absolute bias of the estimator can be obtained from a combination of *r* and *k* values at a particular prevalence level.

4.7.2 The Asymptotic Variance of the Estimator

The asymptotic variance of the estimator is obtained by computing the Fisher's information given by

$$var(\hat{p}) = \left\{-E\left(\frac{\partial^2 \ell(p)}{\partial p^2}\right)\right\}^{-1}. \tag{54}$$

It can be shown that

$$var(\hat{p}) = \left[\frac{1-({\pi_1}^2+(1-p)^k \times ((1-\pi_0)^2-{\pi_1}^2))}{rk^2(1-p)^{2k-2} \times ((1-\pi_0)^2-{\pi_1}^2)^2}\right] \times ({\pi_1}^2+(1-p)^k \times ((1-\pi_0)^2-{\pi_1}^2))^2. \tag{55}$$

The Cramer-Rao bound method was used to derive the variance of the estimator and to show that when the waiting parameter *r* was large, the variance of the estimator was best asymptotically normally distributed.

Method Based on Cramer-Rao Bound

$$\lim_{r\to\infty} var(\hat{p}) = \frac{1}{-E\left(\frac{\partial^2 \ell(p)}{\partial p^2}\right)}. \tag{56}$$

The Likelihood function is given by

$$L(p) = \binom{x-1}{r-1}({\pi_1}^2+(1-p)^k((1-\pi_0)^2-{\pi_1}^2))^r \times (1-({\pi_1}^2+(1-p)^k \times ((1-\pi_0)^2-{\pi_1}^2)))^{x-r}. \tag{57}$$

The log-likelihood function $\ell(p)$ becomes

$$\ell(p) = r\log({\pi_1}^2+(1-p)^k((1-\pi_0)^2-{\pi_1}^2)) + (x-r)\log\Big(1-({\pi_1}^2+(1-p)^k((1-\pi_0)^2-{\pi_1}^2))\Big). \tag{58}$$

$$\frac{\partial \ell(p)}{\partial p} = \frac{-rk(1-p)^{k-1} \times ((1-\pi_0)^2-{\pi_1}^2)}{{\pi_1}^2+(1-p)^k \times ((1-\pi_0)^2-{\pi_1}^2)} + (x-r)\frac{k(1-p)^{k-1}((1-\pi_0)^2-{\pi_1}^2)}{1-({\pi_1}^2+(1-p)^k \times ((1-\pi_0)^2-{\pi_1}^2))} \tag{59}$$

which reduces to

$$\begin{aligned}\frac{\partial \ell(p)}{\partial p} = \frac{k}{(1-p)}\Bigg[& r\left(\frac{{\pi_1}^2}{{\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)} - 1\right) \\ & - (x-r)\left(1 - \frac{(1-{\pi_1}^2)}{1-({\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2))}\right)\Bigg]\end{aligned} \tag{60}$$

$$\begin{aligned}\frac{\partial^2 \ell(p)}{\partial p^2} = \frac{k}{(1-p)^2}\Bigg[& r\left(\frac{{\pi_1}^2}{{\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)} - 1\right) \\ & - (x-r)\left(1 - \frac{(1-{\pi_1}^2)}{1-({\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2))}\right)\Bigg] \\ & + \frac{k}{(1-p)}\Bigg[r\left(\frac{{\pi_1}^2 k(1-p)^{k-1} \times ((1-\pi_0)^2 - {\pi_1}^2))}{[{\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)]^2}\right) \\ & - k(x-r)\left(\frac{(1-{\pi_1}^2) \times (1-p)^{k-1} \times ((1-\pi_0)^2 - {\pi_1}^2))}{[\,1-({\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)]^2}\right)\Bigg].\end{aligned} \tag{61}$$

Equation (61) can be simplified to

$$\begin{aligned}\frac{\partial^2 \ell(p)}{\partial p^2} = \frac{k}{(1-p)^2}\Bigg[& -r\left(\frac{(1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)}{{\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)}\right) \\ & + (x-r)\left(\frac{(1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)}{1-({\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2))}\right)\Bigg] \\ & + \frac{k^2}{(1-p)^2}\Bigg[r\left(\frac{{\pi_1}^2 \times (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)}{[{\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)]^2}\right) \\ & - (x-r)\left(\frac{(1-{\pi_1}^2) \times (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2))}{[\,1-({\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)]^2}\right)\Bigg].\end{aligned} \tag{62}$$

Since x follows the negative binomial distribution with parameters$(r, \pi^*(p))$, where $\pi^*(p) = {\pi_1}^2 + (1-p)^k((1-\pi_0)^2 - {\pi_1}^2)$ then

$$E(t) = \frac{r}{\pi^*(p)}$$

$$E(x-r) = r\left(\frac{1-\pi^*(p)}{\pi^*(p)}\right). \tag{63}$$

Substituting the value of $\pi^*(p) = {\pi_1}^2 + (1-p)^k((1-\pi_0)^2 - {\pi_1}^2)$ into equation (63) it reduces to

$$E(x-r) = r\left(\frac{1-{\pi_1}^2 - (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)}{{\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)}\right). \tag{64}$$

Substituting equation (64) into equation (62) reduces to

$$\begin{aligned}
-E\left(\frac{\partial^2 \ell(p)}{\partial p^2}\right) &= \frac{rk^2}{(1-p)^2}\left[\left(\frac{(1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)}{{\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)}\right)\right. \\
&\quad - \left(\frac{1 - {\pi_1}^2 - (1-p)^k \times ({\pi_1}^2 - (1-\pi_0)^2)}{{\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)}\right) \\
&\quad \left.\times \left(\frac{(1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)}{[\,1 - ({\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2))]^2}\right)\right] \\
&\quad + \frac{rk^2}{(1-p)^2}\left[\begin{array}{l} -\left(\dfrac{{\pi_1}^2 \times (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)}{[\,{\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)]^2}\right) + \\ \left(\dfrac{1 - ({\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2))}{{\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)}\right) \times \\ \left(\dfrac{(1 - {\pi_1}^2)}{[\,1 - ({\pi_1}^2 + (1-p)^k((1-\pi_0)^2 - {\pi_1}^2))]^2}\right) \end{array}\right]
\end{aligned} \tag{65}$$

which simplify to

$$\begin{aligned}
&-E\left(\frac{\partial^2 \ell(p)}{\partial p^2}\right) \\
&= \frac{rk^2(1-p)^{2k}}{(1-p)^2}\left[\left(\frac{((1-\pi_0)^2 - {\pi_1}^2)^2}{1 - ({\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2))}\right)\right] \\
&\times \left(\frac{1}{({\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)^2}\right).
\end{aligned} \tag{66}$$

$$\operatorname*{Lim}_{n\to\infty} var(\hat{p}) = \frac{1}{-E\left(\frac{\partial^2 \ell(p)}{\partial p^2}\right)}$$

which simplifies to

$$\begin{aligned}
&\operatorname*{Lim}_{r\to\infty} var(\hat{p}) \\
&= \left(\frac{(1-p)^2}{rk^2(1-p)^{2k}}\right)\left[\frac{1 - ({\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2))}{((1-\pi_0)^2 - {\pi_1}^2)^2}\right] \times ({\pi_1}^2 \\
&+ (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2))^2
\end{aligned} \tag{67}$$

$$\begin{aligned}
&= \left(\frac{(1-p)^{2-2k}}{rk^2}\right)\left[\frac{1 - ({\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)))}{({\pi_1}^2 - (1-\pi_0)^2)^2}\right] \times ({\pi_1}^2 \\
&\qquad + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2)))^2.
\end{aligned} \tag{68}$$

$$\begin{aligned}
\operatorname*{Lim}_{r\to\infty} var(\hat{p}) &= \left[\frac{1 - ({\pi_1}^2 + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2))}{rk^2(1-p)^{2k-2} \times ({\pi_1}^2 - (1-\pi_0)^2)^2}\right]({\pi_1}^2 \\
&\qquad + (1-p)^k \times ((1-\pi_0)^2 - {\pi_1}^2))^2.
\end{aligned} \tag{69}$$

Hence, the variance of $\hat{p}$ becomes

$$var(\hat{p}) = \left[\frac{1-({\pi_1}^2+(1-p)^k \times ((1-\pi_0)^2-{\pi_1}^2))}{rk^2(1-p)^{2k-2} \times ((1-\pi_0)^2-{\pi_1}^2)^2}\right]$$
$$\times ({\pi_1}^2+(1-p)^k \times ((1-\pi_0)^2-{\pi_1}^2))^2.$$

Thus, The Cramer Rao Lower Bound gives the same asymptotic variance of p and shows that the estimator is the Best Asymptotic Normal for large waiting parameter *r*. That is for fixed *k*, sensitivity and specificity of a test, and $r \to \infty$,

$$I(\hat{p}\ )^{-\frac{1}{2}}(\ \hat{p}-p) \to N\big(0, var(\hat{p})\big). \tag{70}$$

Table 4. 7: variance of $(\hat{p})$ against p for various values of p when k = 15, $r = 1, 3, 5, 10$ and with sensitivity and specificity = 99%, 98%, 95%, and 90%

		k = 15		
		r		
p	1	3	5	10
		Sensitivity = Specificity = 0.99		
0.005	40011.0040	120033.0120	200055.0200	400110.0401
0.010	10035.8073	30107.4219	50179.0365	100358.0730
0.050	391.8239	1175.4718	1959.1197	3918.2394
0.100	84.2147	252.6440	421.0733	842.1466
0.200	8.5956	25.7869	42.9781	85.9562
		Sensitivity = Specificity = 0.98		
0.005	39617.1518	118851.4553	198085.7589	396171.5178
0.010	9962.0121	29886.0362	49810.0604	99620.1207
0.050	382.9342	1148.8025	1914.6709	3829.3418
0.100	78.6214	235.8642	393.1070	786.2141
0.200	6.3662	19.0985	31.8309	63.6617
		Sensitivity = Specificity = 0.95		
0.005	37079.1739	111237.5216	185395.8693	370791.7385
0.010	9565.5124	28696.5373	47827.5622	95655.1244
0.050	357.3788	1072.1365	1786.8942	3573.7884
0.100	65.5688	196.7063	327.8440	655.6879
0.200	3.5959	10.7967	17.9944	35.9889
		Sensitivity = Specificity = 0.90		
0.005	29084.0152	87252.0456	145420.0759	290840.1519
0.010	8322.2950	24966.8849	41611.4749	83222.9497
0.050	316.4389	949.3168	1582.1946	3164.3892
0.100	51.0235	153.0703	255.1172	510.2345
0.200	2.0887	6.2662	10.4437	20.8873

Scrutiny of Table (4.7) showed that the variance of the estimator is a monotone decreasing function of *r* and *k* when the sensitivity and specificity of the test kits are known. It was noted that for any fixed value of the predetermined waiting parameter *r*, the variance of the estimator decreased as the proportion *p* increased. Furthermore, it was also observed that at any fixed value of the proportion *p*, the variance of the estimator increased as the predetermined parameter *r* increased. A striking feature noted was that even when the accuracy of the diagnostic test was relatively low, small variances were observed as the proportion *p* increased. Also noted was that small variances are observed in small predetermined waiting parameters *r*. Thus, in practice, the waiting parameter *r* may not be large enough since countless number of pools may need to be tested before one can observe a predetermined number of pools that tests positive on a retest.

4.7.3 Mean Squared Error of the Estimator (MSE)

The Mean squared error of an estimator is the average squared deviation derived from the true value of the parameter, which incorporates measures of both the accuracy (bias) and the precision (variance) of the estimator. It is used as a measure for the goodness of an estimator that is given by:

$$MSE(\hat{p}) = \left(bias(\hat{p})\right)^2 + var(\hat{p}).$$

The MSE can be expressed as an infinite sum, however, the sums do not reduce to anything tractable. Thus, the MSE of an estimator $\hat{p}$ was suggested by Pritchard and Tebbs (2011b) to be approximated as shown in equation (22) given as follows

$$MSE(\hat{p}) \approx \sum_{x=r}^{x^*} (\hat{p} - p)^2 \binom{x-1}{r-1} \pi^*(p)^r (1 - \pi^*(p))^{x-r}$$

where $pr(X \leq x^*) \geq 1 - v$ for v small. The value of $v = 0.00001$ was taken throughout, making these approximations very close to the true values of bias and MSE.

Since the sums do not reduce to anything tractable, Monte Carlo simulations were invoked and the R code implementing the simulation study that computes the bias tables is annexed in Appendix section.

Table 4. 8: MSE of $\hat{p}$ for various values of p with $r = 1, 3, 5, 10, 15$ and $k = 5, 10, 30, 50$ and with sensitivity and specificity = 99%

MSE $\times 10^{-4}$					
	r				
p	1	3	5	10	15
$k = 5$					
0.005	248.0600	0.704640	0.144780	0.041657	0.023382
0.01	482.9300	3.130540	0.571370	0.164670	0.092545
0.05	2016.7300	123.779720	17.538790	3.936600	2.215701
0.10	3261.6100	564.307690	121.615100	16.517000	8.731067
0.20	4247.3500	1857.421800	839.387760	150.740000	49.387109
0.30	4054.6700	2695.858660	1797.260460	685.060000	289.682941
0.40	3325.5500	2732.301220	2228.374240	1346.600000	832.097851
$k = 10$					
0.005	478.310000	1.629700	0.147543	0.041564	0.023317
0.01	923.560000	9.917900	0.647720	0.165220	0.092695
0.05	3552.640000	561.137100	95.807460	5.135700	2.374300
0.10	5178.770000	2115.678900	873.769169	107.580000	20.79600
0.20	5622.070000	4306.743200	3295.318680	1694.70000	881.810000
0.30	4692.580000	4272.603200	3879.890790	3040.100000	2381.000000
0.40	3535.480000	3382.286300	3224.232340	2844.800000	2501.200000
$k = 30$					
0.005	1356.100000	25.828000	0.625270	0.042023	0.023457
0.01	2501.600000	163.923000	11.184170	0.184380	0.095335
0.05	6949.000000	4117.492000	2440.360710	661.960000	181.220000
0.10	7606.000000	6702.058000	5904.136630	4299.700000	3131.500000
0.20	6271.500000	6017.749000	5772.132500	5197.400000	4677.500000
0.30	4817.600000	4650.590000	4486.036800	4094.100000	3732.900000
0.40	3556.600000	3462.454000	3366.198140	3129.500000	2905.000000
$k = 50$					
0.005	2152.033000	101.968400	4.951560	0.045301	0.023790
0.01	3795.008000	569.470800	85.782540	0.930301	0.107290
0.05	8166.129000	6684.168800	5470.766540	3315.612502	2009.934470
0.10	7899.744000	7512.479800	7143.493100	6297.265286	5550.580170
0.20	6279.433000	6042.506400	5813.196300	5275.105616	4785.490050
0.30	4818.755000	4656.493700	4497.833960	4121.942498	3776.736100
0.40	3557.990000	3470.234900	3382.076640	3168.059613	2967.003440

Scrutiny of Table (4.8) showed that when the group size and the waiting parameter r are fixed, the MSE of $\hat{p}$ increased to a maximum value as the optimal value of the prevalence was attained and subsequently decreased. It was also noted that for any fixed values of the prevalence and group size, the MSE of the estimator decreased as the waiting parameter, r increased. It was therefore possible to obtain a combination of r and k which provided the minimum MSE value when the tests were 99% accurate. For instance, when $p = 0.005$ the minimum value of MSE was obtained at $r = 15$ and $k = 10$.

Table 4. 9: MSE of $\widehat{p}$ for various values of p with $r = 1, 3, 5, 10, 15$ and k = 5, 10, 30, 50 and with sensitivity and specificity = 95%

MSE $\times 10^{-4}$					
	r				
p	1	3	5	10	15
k=5					
0.005	253.980000	0.857360	0.178530	0.051170	0.028702
0.01	470.500000	3.373740	0.643590	0.184030	0.103284
0.05	1879.900000	110.271640	17.963370	4.213000	2.344348
0.10	3018.780000	471.545300	107.371930	20.338320	9.666764
0.20	3919.450000	1480.895760	620.064980	269.110180	73.930991
0.30	3753.850000	2122.574940	1238.619640	1022.769360	407.850012
0.40	3113.380000	2171.630380	1509.867600	1589.588120	885.358940
k=10					
0.005	464.030000	1.624700	0.165970	0.046476	0.026033
0.01	873.100000	8.942400	0.686450	0.176990	0.098977
0.05	3285.510000	451.303000	72.453480	7.534300	2.642158
0.10	4776.330000	1667.629100	603.683580	224.190000	35.406000
0.20	5190.500000	3373.496200	2196.958960	2189.500000	1018.600000
0.30	4361.120000	3384.466500	2605.707320	2989.800000	1987.200000
0.40	3339.040000	2774.564400	2265.512850	2591.900000	1910.800000
k=30					
0.005	1269.400000	21.336000	0.508233	0.044754	0.024797
0.01	2321.200000	131.450000	7.997270	0.233470	0.101530
0.05	6403.900000	3222.576000	1624.382680	1139.400000	284.890000
0.10	7008.900000	5239.424000	3914.380480	4447.800000	2765.300000
0.20	5799.000000	4745.247000	3875.095560	4549.100000	3396.300000
0.30	4495.700000	3763.511000	3141.892010	3619.400000	2785.400000
0.40	3385.900000	2965.426000	2586.401920	2860.900000	2337.000000
k=50					
0.005	2000.300000	82.018000	3.507400	0.056069	0.025197
0.01	3508.900000	450.479000	58.327400	2.725400	0.132810
0.05	7522.200000	5223.102000	3626.523400	3885.200000	2049.800000
0.10	7280.000000	5875.751000	4739.874600	5584.600000	4084.200000
0.20	5809.200000	4779.079000	3929.325800	4601.500000	3476.000000
0.30	4501.300000	3792.030000	3196.577500	3659.800000	2863.100000
0.40	3393.300000	3004.790000	2663.567700	2917.700000	2449.600000

When the assay used was 95% accurate, scrutiny of Table (4.9) showed that for any fixed values of the group size and the waiting parameter *r*, the MSE of the estimator increased to a maximum as the optimal value of the prevalence was attained before it decreased. Secondly, when both the group size and the prevalence rate were fixed, the MSE of the estimator decreased to a minimum as the optimal value of the waiting parameter, *r* was attained and afterward increased. Finally, the results showed that one can obtain a combination of *r* and *k* values that would yield a minimum approximation of the MSE at a given prevalence. For instance, the minimum value of the MSE was obtained at $p = 0.005$ when $r = 15$ and $k = 30$.

Table 4. 10: MSE of $\hat{p}$ for various values of p with $r = 1, 3, 5, 10, 15$ and $k = 5, 10, 30, 50$ and with sensitivity and specificity = 90%

MSE $\times 10^{-4}$					
r					
p	1	3	5	10	15
k=5					
0.005	308.540000	1.611800	0.329340	0.093451	0.052364
0.01	501.510000	4.807400	0.933170	0.260210	0.145740
0.05	1752.250000	108.656500	27.553940	4.824100	2.640800
0.10	2754.610000	413.895500	198.541390	25.665000	11.342000
0.20	3538.330000	1169.715800	1083.567340	329.850000	130.450000
0.30	3393.920000	1597.604400	1854.672870	1037.800000	652.190000
0.40	2850.870000	1612.619500	1992.584970	1450.200000	1146.700000
k=10					
0.005	491.410000	2.160000	0.251430	0.065823	0.036780
0.01	854.830000	9.419000	1.051160	0.216630	0.120570
0.05	2993.080000	359.660000	169.242300	10.426000	3.194290
0.10	4309.840000	1254.709000	1162.433250	279.540000	81.656590
0.20	4676.460000	2461.274000	3054.481370	2037.500000	1471.695850
0.30	3960.790000	2480.735000	3138.887370	2578.800000	2244.231320
0.40	3096.200000	2113.019000	2596.268890	2244.000000	2031.518110
k=30					
0.005	1205.700000	18.584000	1.620100	0.052651	0.028810
0.01	2139.900000	104.117000	24.556500	0.317140	0.114980
0.05	5763.000000	2351.566000	2687.423700	1200.400000	587.524940
0.10	6298.100000	3795.734000	4842.426700	3856.700000	3261.408630
0.20	5234.100000	3473.853000	4401.661600	3843.900000	3519.085690
0.30	4109.300000	2859.746000	3519.539300	3116.500000	2877.620150
0.40	3178.700000	2433.942000	2822.791200	2567.700000	2409.781690
k=50					
0.005	1855.200000	65.734000	11.762000	0.076824	0.028519
0.01	3195.300000	341.118000	158.306000	4.603300	0.264050
0.05	6757.600000	3785.824000	4787.254000	3483.900000	2713.000000
0.10	6541.000000	4256.810000	5439.538000	4704.800000	4276.400000
0.20	5247.600000	3520.340000	4438.477000	3898.000000	3584.200000
0.30	4121.100000	2919.170000	3556.268000	3173.000000	2947.000000
0.40	3194.700000	2517.862000	2873.886000	2647.700000	2509.500000

The results of Table (4.10) showed that when the assays used for testing were 90% accurate, the MSE of $\hat{p}$ depends on the prevalence rate, the group size, and the waiting parameter r. It was generally noted that when the group size and the waiting parameter r are fixed, the MSE of the estimator increased to a maximum as the optimal value of the prevalence was attained before decreasing. Conversely, when the group size and the prevalence rate were fixed, it was observed that the MSE of the estimator decreased to a minimum as the optimal value of the waiting parameter *r* was reached and afterward increased. Typically, group testing can be used to minimize the MSE of an estimator in low-prevalence populations. When the assay used for testing has a 90% accuracy of detecting a rare trait, the choice of group size that minimizes the MSE of the estimator has to be large.

The behaviour of the MSE$(\hat{p})$ was investigated by plotting MSE$(\hat{p})$ against p for different group sizes k and waiting parameter r when the tests are 99% accurate as shown in the figures below:

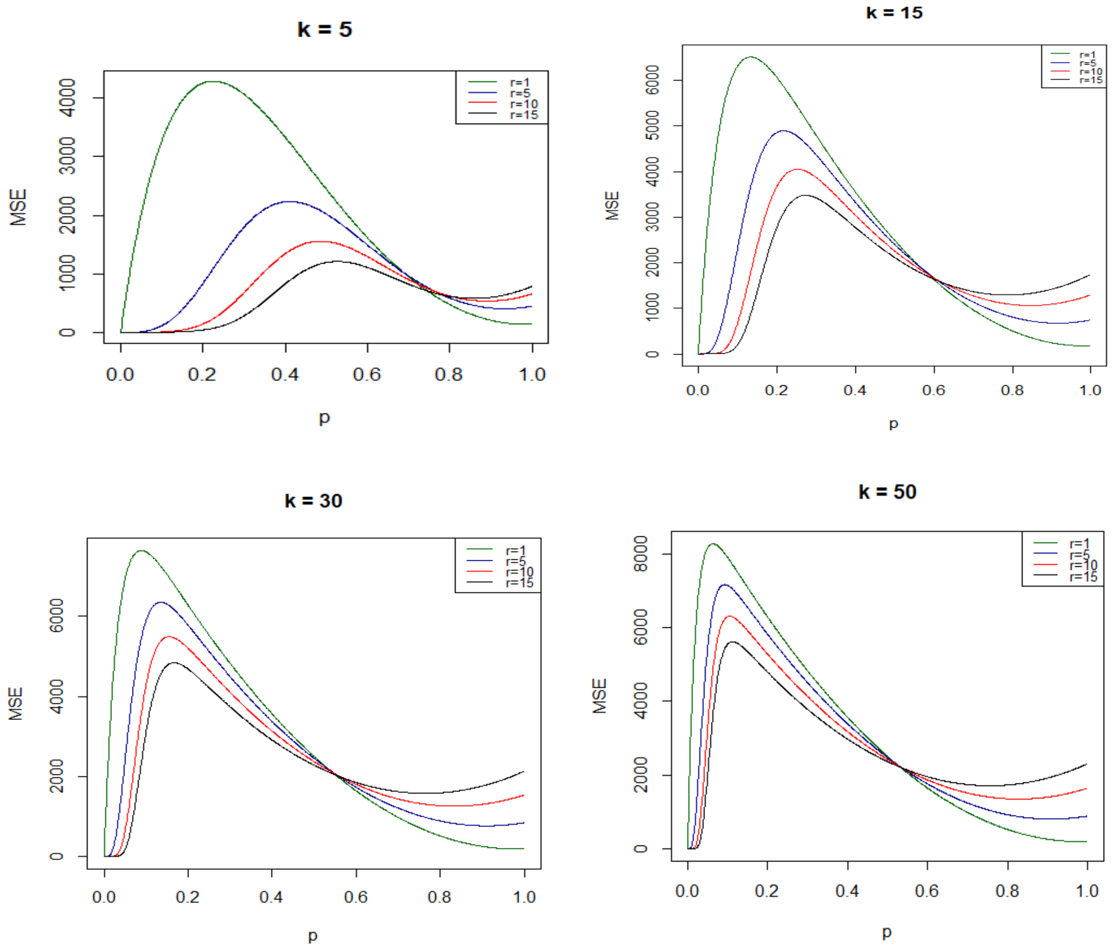


Figure 4. 8: Plot of $MSE(\widehat{p})$ as a function of p for different k = 5, 10, 30, 50 and $r = 1, 5, 10, 15$ with sensitivity and specificity fixed at 0.99

Figure (4.8) shows the MSE of the estimator plotted against the prevalence for different values of the waiting parameter *r* and group size obtained by simulation. It was noted that the optimal value of the MSE of the estimator was observed at a low prevalence level. Besides, the maximum value of the MSE was observed when the group sizes were sufficiently large and when the predetermined waiting parameter r was small. When the waiting parameter r was increased, the optimal value of the MSE also decreased as the prevalence rate increased.

Next, the relationship between the MSE($\hat{p}$) and the prevalence was examined for the various group sizes and waiting parameter r when the tests used were 95% accurate as presented below.

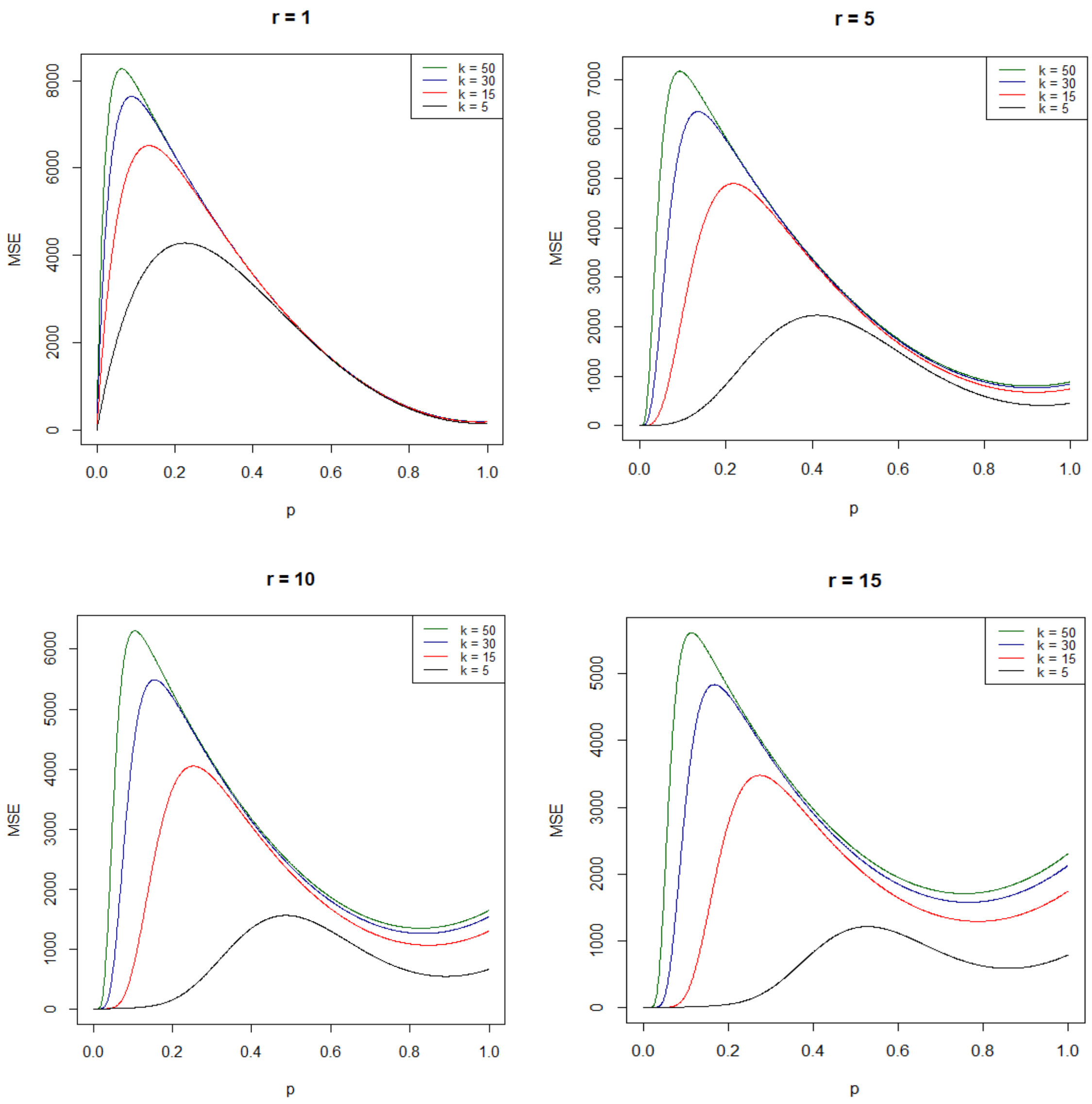


Figure 4. 9: Plot of $MSE(\widehat{p})$ as a function of p for different k = 5, 15, 30, 50 and $r = 1, 5, 10, 15$ with sensitivity and specificity fixed at 0.95

Figure (4.9) shows the relationship between the MSE of the estimator plotted against the prevalence for various group sizes when the waiting parameter, r was predetermined. It was noted that when the waiting parameter r was fixed, the maximum value of the MSE was observed when large group sizes were used. Secondly, the maximum value of the MSE at the optimal value of both the group size and the prevalence decreased as the waiting parameter r

increased. Moreover, it was observed that when $k = 5$ the level of prevalence at which the optimal MSE value was observed increased with an increase in the waiting parameter r.

4.8 Confidence Intervals of Prevalence

When r is large, the MLE is approximately normal with mean p and variance equal to $[I(p)]^{-1}$. A Wald confidence interval for p can be constructed using the estimator for the prevalence and its variance multiplied by the appropriate quantile of the standard normal distribution as shown below

$$\hat{p} \pm Z_{\alpha/2}\sqrt{var(\hat{p})}$$

where $Z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile from the $\mathcal{N}(0,1)$ distribution and

$$var(\hat{p}) = \left[\frac{1-({\pi_1}^2+(1-p)^k \times ((1-\pi_0)^2-{\pi_1}^2))}{rk^2(1-p)^{2k-2} \times ({\pi_1}^2-(1-\pi_0)^2)^2}\right] \times ({\pi_1}^2+(1-p)^k \times ((1-\pi_0)^2-{\pi_1}^2))^2.$$

Monte Carlo simulations were used to generate the confidence intervals and the R code implementing the simulation study is annexed in Appendix C.

Table 4. 11: The 95% CI for different values of p with $r = 2, 5, 10$ and $k = 5$ at sensitivity = specificity = 0.99

p	Lower limit	Upper limit	Width
	$r = 2$		
0.0025	-0.001818	0.011282	0.013100
0.005	-0.003549	0.022454	0.026003
0.01	-0.006866	0.046180	0.053046
0.05	-0.023997	0.209450	0.233447
0.07	-0.028817	0.281430	0.310247
0.10	-0.032835	0.385140	0.417975
	$r = 5$		
0.0025	0.000370	0.005838	0.005468
0.005	0.000768	0.011639	0.010871
0.01	0.001591	0.023232	0.014239
0.05	0.008993	0.114970	0.105977
0.07	0.012822	0.160040	0.147218
0.10	0.018197	0.229310	0.211113
	$r = 10$		
0.0025	0.001046	0.004511	0.003465
0.005	0.002103	0.008977	0.006874
0.01	0.004223	0.017861	0.013638
0.05	0.021732	0.088488	0.066756
0.07	0.030775	0.123730	0.092955
0.10	0.044542	0.177060	0.132518

The results of Table (4.11) showed that when the tests are 99% accurate, the width of the confidence interval increased as the prevalence increased for fixed group sizes and waiting parameter r. It was observed that for small values of the waiting parameter r, the Wald confidence interval performed poorly. Negative values were reported at the lower bound of the confidence interval, which produce unreliable interval estimates. The width of the confidence interval indicated how informative the confidence interval was relative to the parameter, and a narrower length of the confidence interval was observed at $r = 10$ when p=0.0025. In general, it was observed that when the prevalence rate was fixed, the width of the confidence interval decreased as the waiting parameter r increased.

Table 4. 12: The 95% CI for different values of p with $r = 2, 5, 10$ and $k = 15$ at sensitivity = specificity = 0.99

p	Lower limit	Upper limit	Width
		$r = 2$	
0.0025	-0.001842	0.012351	0.014193
0.005	-0.003550	0.026084	0.029634
0.01	-0.006356	0.055199	0.061555
0.05	-0.014694	0.363470	0.378164
0.07	-0.013719	0.503160	0.516879
0.10	-0.010703	0.665730	0.676433
		$r = 5$	
0.0025	0.000381	0.005832	0.005451
0.005	0.000769	0.011624	0.010855
0.01	0.001545	0.023374	0.021829
0.05	0.004683	0.146430	0.141747
0.07	0.003976	0.241400	0.237424
0.10	0.002348	0.413840	0.411492
		$r = 10$	
0.0025	0.001050	0.004485	0.003435
0.005	0.002111	0.008985	0.006874
0.01	0.004222	0.017920	0.013698
0.05	0.020128	0.093936	0.073808
0.07	0.025358	0.141640	0.116282
0.10	0.027239	0.247780	0.220541

The results of Table (4.12) showed that the confidence interval was affected by the group size, the prevalence, and the predetermined waiting parameter r when the sensitivity and the specificity values of the test are held constant throughout the testing process. The width of the confidence interval widened as the proportion p increased at any fixed values of the waiting parameter r. Negative lower bounds of the confidence interval were reported, especially at small values of the waiting parameter r. This produced unreliable interval estimates confirming that the waiting parameter r has to be sufficiently large for the asymptotic properties to take hold. Lastly, at any fixed value of the proportion p, the width of the Wald confidence interval decreased as the waiting parameter r increased.

Next, the behaviour of the Wald confidence interval was examined when the group size was increased to $k = 30$.

Table 4. 13: The 95% CI for different values of p with $r = 2, 5, 10$ and $k = 30$ at sensitivity = specificity = 0.99

p	Lower limit	Upper limit	Width
		$r = 2$	
0.0025	-0.001809	0.015334	0.017143
0.005	-0.003280	0.036323	0.039603
0.01	-0.005603	0.094930	0.100533
0.05	-0.005862	0.626170	0.632032
0.07	-0.003818	0.778580	0.782398
0.10	-0.001948	0.893090	0.895038
		$r = 5$	
0.0025	0.000379	0.005826	0.005447
0.005	0.000749	0.011720	0.010971
0.01	0.001389	0.024643	0.023254
0.05	0.000361	0.336260	0.335899
0.07	-0.000263	0.552140	0.552403
0.10	-0.000404	0.757310	0.757714
		$r = 10$	
0.0025	0.001050	0.004484	0.003434
0.005	0.002096	0.008952	0.006856
0.01	0.004185	0.018034	0.013849
0.05	0.012837	0.157790	0.144953
0.07	0.010171	0.337310	0.327139
0.10	0.005767	0.598910	0.593143

Table (4.13) indicates the confidence interval and the associated width when *k=30*. The findings indicate that the width of the confidence interval increased with an increase in the proportion *p* for a predetermined waiting parameter *r*. It was observed that the Wald confidence interval was associated with wider confidence width and negative lower bound as the proportion *p* increased especially for small *r*. Also, a narrow confidence width was observed at large values of the waiting parameter *r* when the proportion *p* is small which allowed the parameter *p* to be estimated more precisely.

4.9 Coverage Probabilities

The performance of the Wald interval was compared with that of the Wilson interval, Agresti- Coull interval, and the Exact interval by plotting their coverage probabilities against the prevalence as shown in the figures below

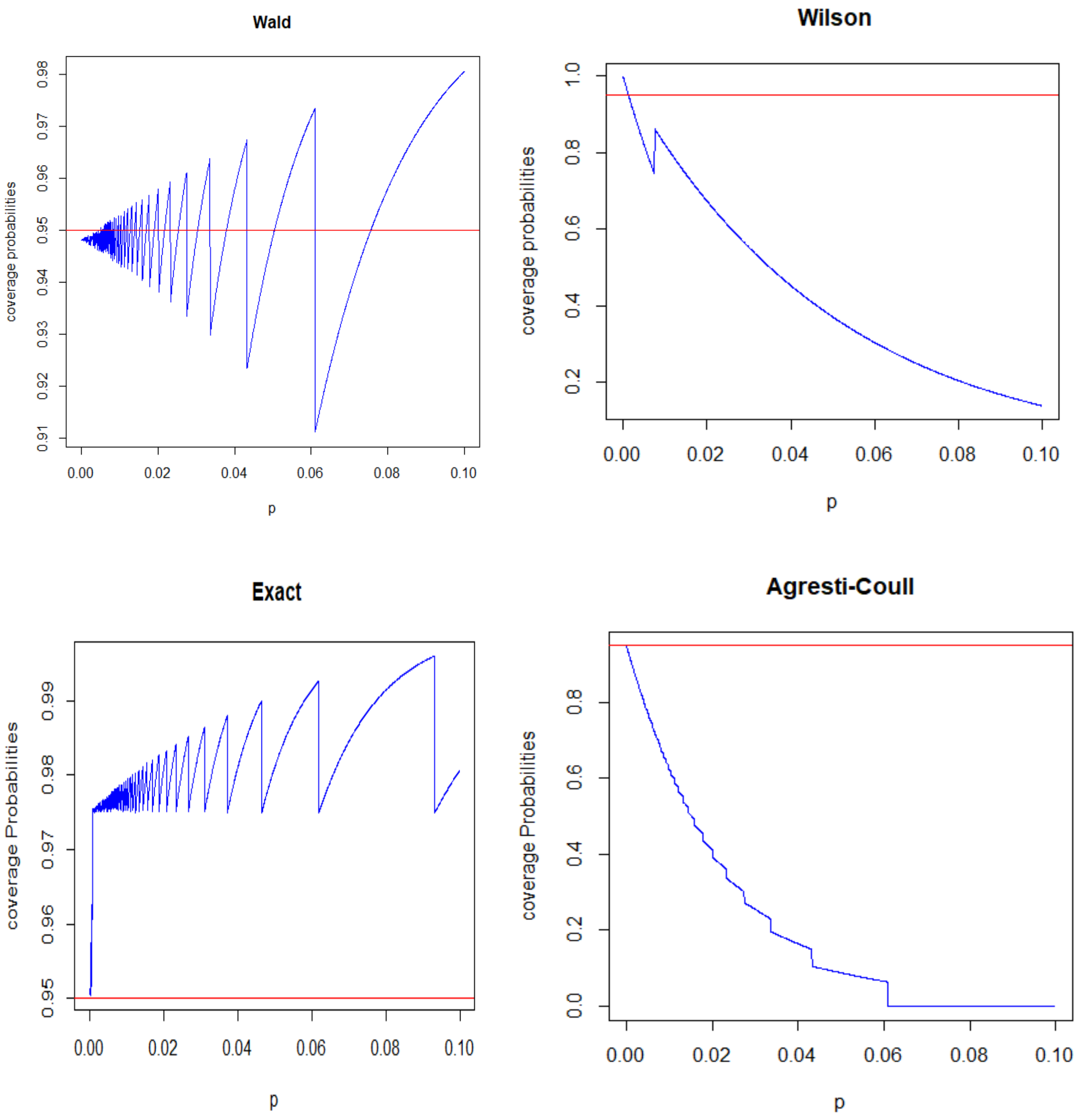


Figure 4. 10: Plot of coverage probability for 95% Wald, Exact Wilson and Agresti- Coull interval when $r = 2$, $k = 20$, sensitivity = specificity = 0.99

Figure (4.10) illustrates the 95% coverage probabilities for the four confidence intervals when $k = 20$, and $r = 2$, when the assays used are 99% accurate. The values of the prevalence ranged from 0.005 to 0.10 by increments of 0.00005. A distinct jagged pattern was observed due to the discrete nature of the negative binomial distribution. The results showed that when $r = 2$ the Wald interval was centered around the 95% nominal level, which was unusual from the available literature. The exact interval was observed to be conservative in that the confidence interval was wider as the prevalence level increased. Both the Wilson and Agresti-

Coull interval showed that 95% coverage was obtained at low prevalence. Lastly, the Wilson interval converged to zero much faster as the prevalence level increased.

Next, the relationship between the coverage probabilities and the prevalence was examined when the waiting parameter r was increased to $r = 5$.

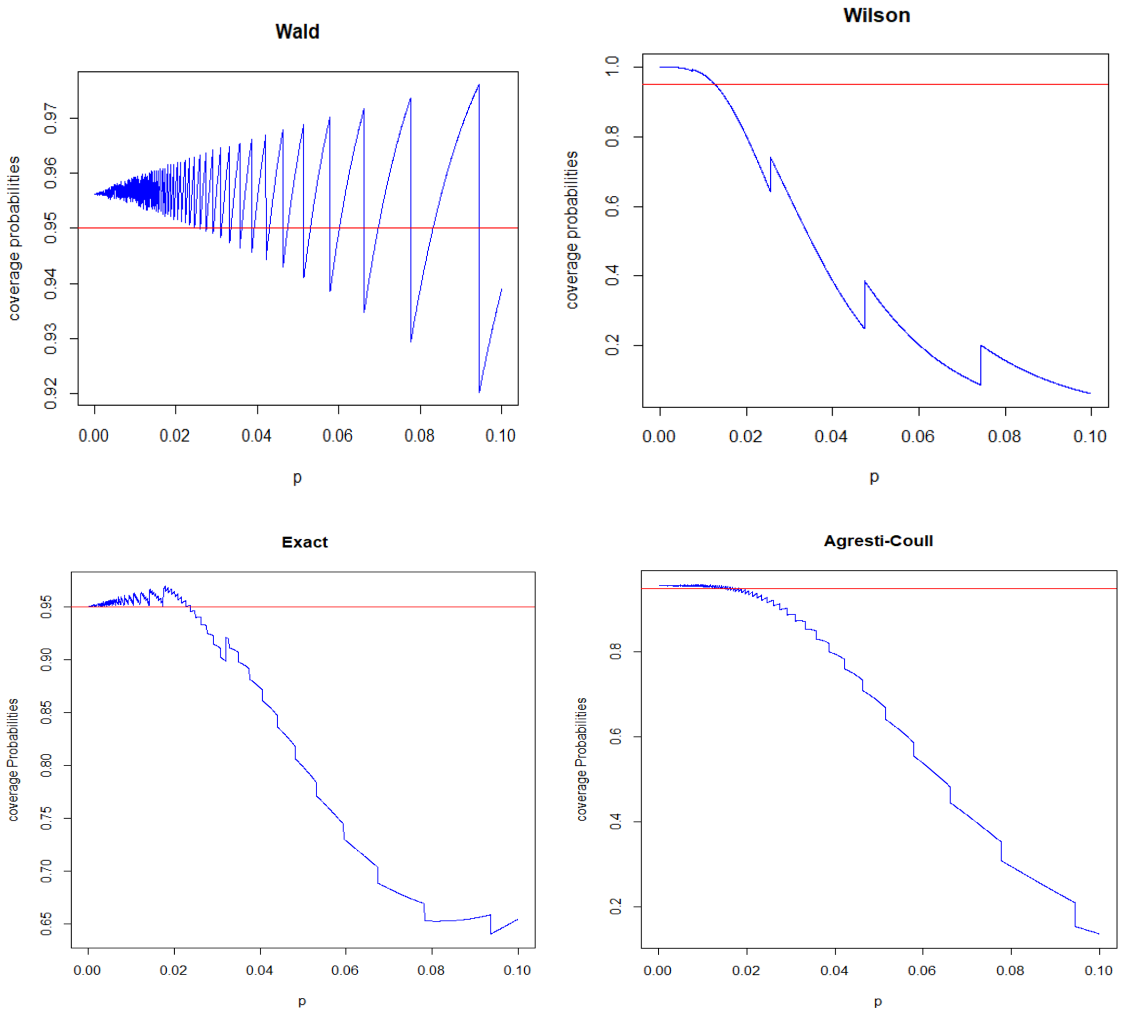


Figure 4. 11: Plot of coverage probability for 95% Wald, Exact Wilson and Agresti- Coull interval when $r = 5$, $k = 20$, sensitivity = specificity = 0.99

Figure (4.11) displays the coverage probabilities for the four different intervals when r was increased to 20. The 95% coverage of wald intervals to some degree was observed to be centered around the nominal level, although their variability increased as p increased. The coverage probabilities for the Agresti- Coull interval were close to the nominal level for low values of p. However as the prevalence increased, a drop in the coverage probability was observed. A similar pattern was observed for the Exact interval. For $p < 0.02$, the Wilson interval was observed to be conservative, whereas, for $p > 0.02$, a drop in the coverage probabilities was observed below the nominal level.

Lastly, the waiting parameter r was increased to 50, and the results are shown in figures (4. 12)

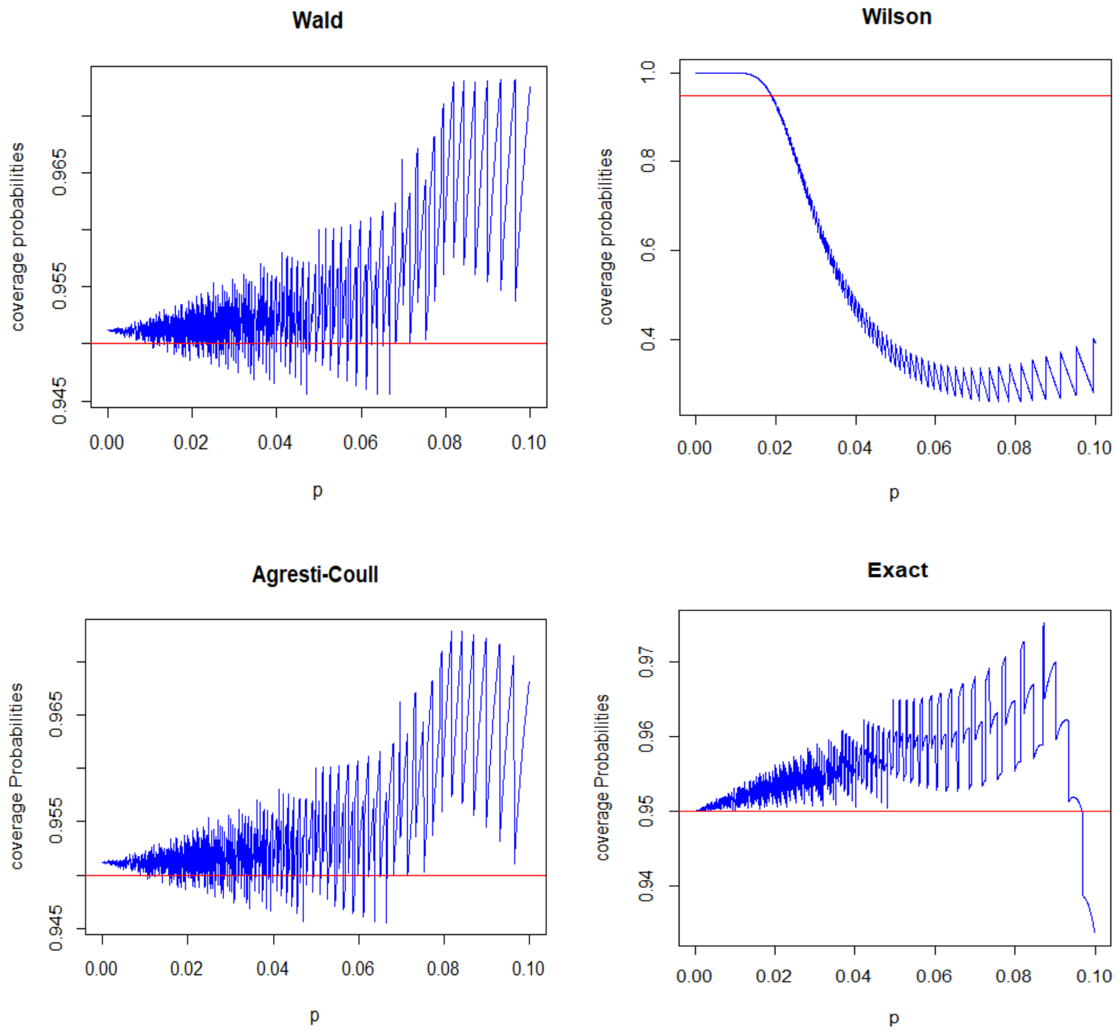


Figure 4. 12: Plot of coverage probability for 95% Wald, Exact Wilson and Agresti- Coull interval when $r = 50$, $k = 20$, sensitivity = specificity = 0.99

Figure (4.12) displays the coverage probabilities of the four intervals when the waiting parameter $r = 50$. The coverage probabilities of the Wald, Agresti- Coull, and Exact intervals were observed to be chaotic at low values of p and served as a shred of visual evidence that the asymptotic properties were taking hold. The coverage probabilities of the Wilson interval were observed to be conservative at $p < 0.02$, but as p increased, there was a drop in the coverage probabilities below the nominal level. Both the Wald and Agresti-Coull intervals exhibited that their coverage probabilities were close to the nominal level, and as the proportion p increased, their variability also increased. Lastly, the Exact interval was observed to be conservative as the proportion p increased and at $p > 0.10$, a drop in the coverage probabilities was observed below the 95% nominal level.

Table 4. 14: The estimated coverage probability of the confidence interval by fixing $k = 20$ given that the sensitivity and specificity of the tests are set at 0.99

		Sensitivity=Specificity=0.99			
		$p = 0.001$	$p = 0.025$	$p = 0.05$	$p = 0.10$
$r = 2$	**Exact**	0.953260	0.791570	0.693980	0.784240
	Wald	0.951440	0.946900	0.933250	0.947380
	Wilson	0.997773	0.663099	0.311096	0.052624
	Agresti-Coull	0.997773	0.663099	0.311096	0.052624
$r = 5$	**Exact**	0.950544	0.939589	0.798446	0.654737
	Wald	0.956110	0.954590	0.963330	0.939100
	Wilson	0.999998	0.659274	0.340133	0.060898
	Agresti-Coull	0.956112	0.918279	0.682613	0.137184
$r = 20$	**Exact**	0.950338	0.956098	0.966471	0.646439
	Wald	0.952710	0.951270	0.966080	0.973460
	Wilson	1.000000	0.806100	0.322080	0.23434
	Agresti-Coull	0.952706	0.951275	0.965294	0.784134

Table (4.14) shows the coverage probabilities for the four methods of confidence estimations at various waiting parameters r when $k = 20$. The Wilson's interval was observed to be very conservative at $p = 0.001$ but as p increased, the coverage probabilities were smaller than the nominal level. On the other hand, the Wald interval was observed to have coverage probabilities that were close to the nominal level or all the values of r. The Agresti- Coull interval performed poorly at $p = 0.001$ when at $r = 2$ but as r increased, the coverage probabilities were centered around the nominal level. Finally, the Exact interval for large values of *r* showed that their coverage probabilities were very close to the nominal level for $p \leq 0.05$.

Table 4. 15: The estimated coverage probability of the confidence interval by fixing $k = 20$ given that the sensitivity and specificity of the tests is 0.98

		Sensitivity=Specificity=0.98			
		$p = 0.001$	$p = 0.025$	$p = 0.05$	$p = 0.10$
$r = 2$	**Exact**	0.956389	0.742640	0.741829	0.843686
	Wald	0.951270	0.941510	0.966040	0.934340
	Wilson	0.997773	0.663099	0.311096	0.052624
	Agresti-Coull	0.909776	0.326340	0.090750	0.000000
$r = 5$	**Exact**	0.951122	0.946001	0.805348	0.641284
	Wald	0.955980	0.961160	0.955190	0.918240
	Wilson	0.999998	0.659274	0.340133	0.060898
	Agresti-Coull	0.955983	0.927962	0.695779	0.156672
$r = 20$	**Exact**	0.950267	0.958160	0.957398	0.505123
	Wald	0.952790	0.952800	0.965000	0.950810
	Wilson	1.000000	0.806100	0.322080	0.234340
	Agresti-Coull	0.952790	0.952790	0.965958	0.813037

Table (4.15) shows the estimated coverage probabilities for the four methods of confidence estimation at various waiting parameters r when the assays used for testing are 98%. Scrutiny of the table showed that when $p = 0.001$, both the Wald interval and the Exact interval have their coverage probabilities that are close to the nominal level for all the values of r. Also, the two confidence intervals outperformed the other methods of confidence interval at large values of r as p increased. It was noted that the Wilson's interval was very conservative when $p = 0.001$ but a drop in their coverage probabilities was observed as p increased. Lastly, it was observed that the coverage probabilities for the Agresti-Coull interval were close to the nominal level at large values of r which indicated that the asymptotic properties were taking hold.

Table 4. 16: The estimated coverage probability of the confidence interval by fixing *k=20* given that the sensitivity and specificity of the tests are 0.95

		Sensitivity=Specificity=0.95			
		$p = 0.001$	$p = 0.025$	$p = 0.05$	$p = 0.10$
$r = 2$	**Exact**	0.955288	0.751583	0.743288	0.827037
	Wald	0.951290	0.947020	0.948960	0.889270
	Wilson	0.997590	0.704400	0.381220	0.110730
	Agresti-Coull	0.907815	0.365555	0.102355	0.000000
$r = 5$	**Exact**	0.950191	0.918453	0.674412	0.451302
	Wald	0.956380	0.954930	0.955340	0.937150
	Wilson	1.000000	0.728970	0.310000	0.160140
	Agresti-Coull	0.956377	0.929504	0.752070	0.298790
$r = 20$	**Exact**	0.950307	0.958069	0.939747	0.275743
	Wald	0.952920	0.952050	0.966080	0.930740
	Wilson	0.999999	0.999974	0.997031	0.863075
	Agresti-Coull	1.000000	0.846940	0.389650	0.282620

Table (4.16) shows the estimated coverage probabilities for the four methods of confidence estimation at various waiting parameters r when the assays used for testing are 95%. The coverage probabilities for both Wilson and Agresti-Coull intervals decreased as the proportion p increased. Secondly, for small values of r, the coverage probabilities of the Agresti-Coull interval were below the 95% nominal levels and converged to zero much faster as the proportion p was increased. In general, when $r = 20$, Wilson's interval was observed to be conservative for $p < 0.10$. Thirdly, when r was large, the Wald interval generally outperformed the other confidence interval since most of its coverage probabilities were close to the nominal level. The exact interval for large values of r was centered around the nominal level, but as p increased, their coverage probabilities decreased.

4.10 Model Comparison

In this section, the computed results were compared with the negative binomial group testing model when imperfect tests are used without retesting. This was accomplished by computing ARE and the RMSE.

4.10.1 Asymptotic Relative Error (ARE)

The sample size based on the one-stage negative binomial group testing scheme that has considered misclassification was examined by Xiong (2016). The variance of the estimator was computed as follows:

$$var(\hat{p}) = \left\{\frac{[\pi_1 - (\pi_0 + \pi_1 - 1)(1-p)^k]^2[1 - \pi_1 + (\pi_0 + \pi_1 - 1)(1-p)^k]}{rk^2(\pi_0 + \pi_1 - 1)^2(1-p)^{2k-2}}\right\}.$$

If the estimator of the One-Stage negative binomial group testing scheme with misclassification is denoted by $\hat{p}_d$, and the computed estimator is denoted p_i then

$$ARE = \frac{var(\hat{p}_d)}{var(p_i)}$$

Therefore, $ARE > 1$ implies that the proposed model is more efficient than Xiong (2016) model with inspection errors.

Table 4. 17: ARE of the proposed model relative to Xiong's (2016) model with errors in inspection for $k = 5, 10, 15, 30, 50$ with sensitivity and specificity = $99\%, 98\%, 95\%,$ and 90%

			r = 2		
			k		
p	5	10	15	30	50
			Sensitivity=Specificity=0.99		
0.005	0.510710	0.694550	0.778750	0.879780	0.926490
0.01	0.695130	0.827230	0.880060	0.939190	0.966370
0.05	0.928180	0.967290	0.983950	1.018560	1.091890
0.10	0.968460	0.996990	1.020620	1.167320	1.624620
0.20	0.999640	1.061750	1.197370	1.845370	1.949040
0.30	1.031470	1.236870	1.642150	1.948260	1.950400
0.40	1.090820	1.589170	1.907640	1.950380	1.950400
			Sensitivity=Specificity=0.98		
0.005	0.311080	0.510600	0.623230	0.779350	0.860140
0.01	0.511330	0.695100	0.779810	0.883120	0.933710
0.05	0.863170	0.935440	0.967240	1.032360	1.157080
0.10	0.937670	0.992120	1.036120	1.268290	1.714580
0.20	0.997140	1.108010	1.308530	1.848710	1.900950
0.30	1.055660	1.358330	1.726590	1.900570	1.901590
0.40	1.155380	1.689460	1.880780	1.901580	1.901590
			Sensitivity=Specificity=0.95		
0.005	0.123530	0.257540	0.363850	0.562960	0.697000
0.01	0.258150	0.446940	0.563670	0.739770	0.841720
0.05	0.702500	0.845380	0.914120	1.050000	1.251560
0.10	0.850110	0.968040	1.057190	1.382690	1.686220
0.20	0.978710	1.180650	1.422250	1.740970	1.759650
0.30	1.093370	1.466630	1.691570	1.759520	1.759870
0.40	1.249260	1.674700	1.752600	1.759870	1.759870
			Sensitivity=Specificity=0.90		
0.005	0.068206	0.137670	0.204830	0.368190	0.513130
0.01	0.138016	0.266410	0.368870	0.566510	0.708750
0.05	0.519800	0.714240	0.820610	1.022540	1.240740
0.10	0.721365	0.904830	1.032030	1.342620	1.508590
0.20	0.921050	1.174130	1.369000	1.531480	1.538910
0.30	1.077610	1.396600	1.510900	1.538860	1.539000
0.40	1.238728	1.503560	1.536130	1.539000	1.539000

Table (4.17) shows ARE values at different group sizes, prevalence levels, and waiting parameter r when the accuracies of tests used are known. Scrutiny of the table showed that retesting a pool that tested positive in the initial test made the proposed model more efficient, as the proportion p increased. This was indicated by $ARE > 1$. It was observed that for a given prevalence of a rare trait, the efficiency of the proposed model increased as the size of the groups was increased. Similarly, for a fixed group size, the efficiency of the model increased as p increased. It was observed that even when the sensitivity and specificity of the test were relatively low, retesting improved the efficiency of the model as the proportion p increased. From the table, it was possible to get the combination of p and k where the model was more efficient. Thus, in cases where group testing is used to screen for a rare trait, retesting of pools is more desirable as it yields more accurate results, and the retesting model outperforms the one-stage negative binomial group testing design as p increases.

4.10.2 Relative Mean Squared Error (RMSE)

To compare the efficiency of a model, a convenient way was to compare the MSE of estimates of the same p with other existing models obtained using different experimental procedures. The MSE of the proposed model was compared with the MSE of the one-stage negative binomial group testing model with misclassification denoted by $\hat{p}_d$. The RMSE was calculated as follows:

$$RMSE = \frac{MSE(\hat{p}_d)}{MSE(p_i)}$$

Table 4. 18: Relative Mean Squared Error of the Estimator (RMSE) with sensitivity and specificity value set at 99%

	r		
p	1	3	5
$k = 5$			
0.005	38.565142	12733.370308	58400.557061
0.01	19.133470	2643.906573	13086.300135
0.05	3.443933	34.238002	154.682671
0.10	1.463359	3.042883	5.576623
0.20	0.525123	0.194483	0.192145
0.30	0.337283	0.241976	0.365225
$k = 15$			
0.005	12.899748	1896.419895	39821.816914
0.01	6.205910	226.377645	4210.861195
0.05	0.882280	0.687266	0.538078
0.10	0.279830	0.036577	0.030166
0.20	0.105515	0.070938	0.079776
0.30	0.196267	0.195440	0.205301
$k = 30$			
0.005	6.228922	238.248501	7177.123081
0.01	2.879419	23.761346	188.761043
0.05	0.287960	0.027176	0.008577
0.10	0.066378	0.014407	0.016307
0.20	0.073766	0.066298	0.069187
0.30	0.193811	0.193251	0.200459
$k = 50$			
0.005	3.555019	44.820292	551.766813
0.01	1.557081	3.773527	9.121756
0.05	0.096818	0.004225	0.004213
0.10	0.027673	0.013245	0.013947
0.20	0.072861	0.066115	0.068759
0.30	0.194272	0.193116	0.199999

Scrutiny of Table (4.18) showed that for small group size, the proposed model performed well when the prevalence was low. This was indicated by the value of $RMSE > 1.$ For instance, when $r = 1$, and $k = 5$, the model was observed to be 38.6 times more efficient than the one-stage negative binomial group testing model with misclassification at $p = 0.05$. In general, the proposed model was examined to be more efficient, especially in a low prevalent population when the group sizes are fairly small. Lastly, large values of the waiting parameter r, when the prevalence was low, yielded a more efficient model.

Table 4. 19: Relative Mean Squared Error of the Estimator (RMSE) at 95% sensitivity and specificity value

	r		
p	1	3	5
$k = 5$			
0.005	36.204290	9359.937964	39805.251986
0.01	18.897137	2201.840151	9836.147127
0.05	3.589811	35.559549	135.836994
0.10	1.561931	3.537761	6.190444
0.20	0.592202	0.265764	0.272134
0.30	0.395585	0.308493	0.524353
$k = 15$			
0.005	13.034037	1879.883663	30942.258974
0.01	6.403645	242.737094	3812.958260
0.05	0.961991	0.887832	0.810215
0.10	0.336188	0.055905	0.046155
0.20	0.153077	0.089761	0.118998
0.30	0.242057	0.241927	0.299192
$k = 30$			
0.005	6.429417	260.242107	7446.285644
0.01	3.022253	27.380735	231.694665
0.05	0.344612	0.044954	0.014607
0.10	0.113660	0.019012	0.024423
0.20	0.119842	0.083489	0.102518
0.30	0.240808	0.237495	0.285215
$k = 50$			
0.005	3.714542	51.049440	673.427405
0.01	1.660597	4.574145	12.515071
0.05	0.145590	0.007191	0.006349
0.10	0.073270	0.016973	0.020894
0.20	0.119303	0.083294	0.101403
0.30	0.242020	0.236409	0.280821

Table (4.19) shows that when the assays used are 95% accurate, the proposed model that incorporated the sequential testing of positive pools performed well compared to the one-stage negative binomial group testing model that has considered misclassifications, especially at low prevalence. As r increased, the proposed model was observed to be more efficient, especially at low prevalence. This indicated that the asymptotic properties of the model were taking hold. Thus, it was possible to get a combination of k and r values where the model would be more efficient for a given prevalence. For instance, when $k = 30$ and $r = 5$, the model was reported to be approximately 7446 more efficient for $p = 0.005$. Lastly, the prevalence at which the proposed model was more efficient reduced as the size of the group increased.

4.11 Application of the Model to Real Data

Surveillance is an important aspect of public health for the prevention and early detection of infectious diseases that would otherwise cause an outbreak. A public health study that involved the surveillance of West Nile Virus was conducted in Jefferson County, Florida following the 2001 outbreak of WNV transmitted by the North American mosquito, *Culex nigripalpus* (Rutledge *et al*., 2003). In North America, many mosquito species were infected with WNV. However, in this study, the primary focus was on the *Culex* species as the primary West Nile vector. To access transmission of WNV, a total of 11,948 mosquitos were captured and tested in various pool sizes using reverse-transcription polymerase chain reaction assays. A total of 14 mosquito pools contained WNV after testing. The authors documented the first field study on the mosquito transmission rate of WNV. By the end of the outbreak, 12 human cases were reported to have West Nile *Meningoencephalitis,* and 483 cases among the horses were documented.

Rutledge *et al*. (2003) did not consider inverse sampling design when imperfect tests are used. However, this investigation was used as a basis to illustrate the proposed group testing design by considering imperfect assays and sequentially retesting a pool that tests positive in the initial stage. The authors were only interested in the pools that tested positive. Based on the field study, the estimated prevalence was reported to be approximately 0.005. Hence it was presumed that 0.005 was the true value of the prevalence. R programming language was used to simulate 10,000 Monte Carlo data sets for the waiting parameters r= 1, 5, 10, and 15 at equal group sizes k= 5, 15, 30, and 50 similar to the public study conducted by Rutledge *et al*. (2003). The values of the sensitivity and specificity of the tests were held constant throughout the testing procedure, and the simulated data contained assays used for testing which were 99%, 98%, 95%, and 90% accurate.

Based on 10,000 Monte Carlo data sets, simulated results of the MLE of the proposed estimator for the waiting parameters $r = 1, 5, 10,$ and 15 have been presented. For a low prevalent population, it was observed that the MLE of the proposed model was consistent with Rutledge *et al*. (2003). When the waiting parameter was small say $r = 1$, the MLE was observed to be exorbitantly positively biased as the prevalence level increased. The results further reaffirm our discovery that group testing is useful in a low prevalent population and that retesting increased the precision of the estimates.

The confidence interval of the prevalence was generated from 10,000 Monte Carlo data and the results are presented in Tables (4.11), (4.12), and (4.13) respectively. It was observed that the Wald confidence interval was associated with wider confidence width and negative lower bound as the proportion *p* increased especially for small *r*. As the waiting parameter *r* increased, the Wald interval and Exact interval provided good coverage probabilities that were close to the nominal level. Both Wilson and Agresti-Coull intervals were observed to be conservative in that the confidence intervals were likely to be wider as *r* increased. The results are presented in section (4.9). Notably, the coverage probabilities for the Wald interval were observed to be centered around the 95% nominal level when $r = 2$. This was somewhat unusual from the available literature. However, its variability increased with an increase in prevalence as illustrated in Figures (4.10), Figures (4.11), and Figures (4.12).

4.12 Discussions

A two-stage negative binomial group testing procedure for estimating the prevalence of a rare trait has been constructed and analysed. The testing procedure differs from the formal pool testing procedure suggested by Dorfman (1943) in that a pool that tested positive was sequentially retested as illustrated in Figure (4.1). The probability of declaring a pool as positive on a retest was obtained, and the results suggest that $\pi^*(p)$ increases with an increase in prevalence. Further, it was observed that $\pi^*(p)$ remained invariant at high prevalence as observed in Figure (4.2) and Figure (4.3) respectively. In special cases $\pi^*(p)$ converged to the sensitivity and specificity of the test kits which were held constant throughout the testing process. The results coincide with observations made when batch testing has been applied in a quality control process (Wanyonyi *et al.*, 2015a), and in multiple test pool testing strategies when estimating HIV/ AIDS prevalence (Nyongesa, 2018).

The MLE of the prevalence has also been derived and its properties investigated. It sufficed to note that the MLE was observed to be positively biased as seen in Tables (4.1), (4.2), and (4.3) respectively. The MLE was observed to be a monotone-increasing function of *r, k,* and *p.* When both r and k are large, a close approximation of the prevalence was observed at $p = 0.005$. This result agrees with the estimated prevalence level reported in a public health study that involved the surveillance of WNV (Rutledge *et al.*, 2003). Alternative estimators have been developed by different scholars to reduce the bias when the waiting parameter *r* is small (Hepworth, 2013; Pritchard & Tebbs, 2011b). Gart's bias correction to the MLE has been recommended as an effective estimator in reducing the bias (Hepworth & Watson, 2009).

The Wald confidence interval was examined, and it was noted that the interval performed poorly for small values of the waiting parameter *r*. Subsequently, a negative lower limit of the confidence interval was observed, and the width of the confidence interval increased as the prevalence increased. It was established that the Wald interval had complications when all pools tested positive that is $\hat{p} = 1$ under the negative binomial group testing model (Pritchard & Tebbs, 2011b). If this happens, the Wald interval was set to be equivocal (0, 1), since $I(\hat{p})$ is undefined, and the upper limit of the interval cannot be computed.

The confidence intervals for the binomial proportion have been explored extensively by Orawo (2021). The standard Wald confidence interval has been compared with other alternative confidence intervals namely, Wilson confidence intervals, Clopper-Pearson intervals, and Likelihood intervals in terms of coverage probabilities and expected mean length. In this study, the coverage probabilities of the Wald interval relative to that of the Wilson interval, Agresti- Coull interval, and the exact interval were examined based on the coverage probabilities through simulation. At small values of the waiting parameter r, the coverage probabilities of the Wald interval showed an unusual pattern that was somewhat different from the literature. This is because the coverage probabilities were centered around the 95% nominal level.

The behaviour of the coverage probabilities for the Wald Interval and Agresti-Coull interval is quite Chaotic for large values of r as observed in Figure (4.12). This confirmed that the asymptotic properties were taking hold as most of the coverage probabilities were close to the nominal level. Both the Wilson and Exact intervals were observed to be conservative as the prevalence increased. It can be deduced that the Wilson interval showed a substantial departure from the nominal level implying it has a longer expected length.

The performance of the estimator was compared with other estimators to determine its efficiency by computing the ARE values. The results showed that the proposed estimator was superior to the one-stage negative binomial group testing model with misclassification as suggested by Xiong (2016). This can be observed in Table (4.17) where $ARE > 1$ indicates that the estimator has a smaller variance than the Xiong (2016) estimator as the proportion p increases. Moreover, the proposed estimator in this study was established to be more efficient in situations where the sensitivity and specificity of the test kits are low. Thus, retesting of positive pools improves the efficiency of an estimator which confirms what other authors have established (Nyongesa, 2011; 2018; Wanyonyi *et al*., 2021).

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary

The results of the study show that group testing is generally feasible in a low-prevalent population. When estimating the prevalence of a rare trait, quick responses and immediate intervention could curb the spread of an outbreak. A special factor to consider when group testing is applied is the size of the groups and the sensitivity and specificity of the assays used for testing. This thesis focused on a combination of negative binomial model and group testing to estimate the prevalence of a rare trait by incorporating the retesting of positive pools. The testing procedure involved the sequential testing and retesting of a pool that tested positive until a predetermined number of pools that tested positive on retesting was observed. This was performed to increase the precision of a test during estimation. Throughout the testing procedure, the sensitivity and specificity of the tests were held constant.

The general objective of this research which was to construct and analyse a two-stage negative binomial group testing procedure for estimating the prevalence of a rare trait has been achieved. The indicator function was utilized to formulate the probability that a pool tests positive for a rare trait on retesting using the negative binomial group testing model. The finding showed that since the probability is a monotone-increasing function of the prevalence, it ranged between 0 and 1.

The study also utilized the Method of Maximum Likelihood Estimation to obtain the MLE of *p* and the performance of the constructed estimator was assessed. The results showed that the MLE of the proportion *p* under the negative binomial group testing model is positively biased. Furthermore, the MSE and the variance of the constructed estimator were examined to measure the goodness and the accuracy of the estimator respectively. The results showed that the MSE increased with an increase in the prevalence.

A comparison of the constructed estimator relative to the one-stage negative binomial group testing model with misclassification was performed by computing the ARE and RMSE. The results showed that the constructed estimator had a smaller variance, and the proposed model was observed to be more efficient in situations where the sensitivity and specificity of the test kits were low. This result agrees with what other authors have established that retesting of pools improves the efficiency as highlighted in the discussion. Lastly, the proposed model was applied to West Nile Virus.

5.1 Conclusions

This research has constructed and analysed a two-stage negative binomial group testing procedure for estimating the prevalence of a rare trait. The study was able to obtain the point estimator for the prevalence using the negative binomial group testing model with the retesting of positive pools. The interval estimates and coverage probabilities of the constructed estimator were also examined. The following conclusions are made:

i. The estimator obtained using the Maximum Likelihood Estimation method was observed to be positively biased under the negative binomial group testing model.

ii. The properties of the constructed estimator such as the bias and the Mean Squared Error both increased as the prevalence of a rare trait increased. Alternative estimators that reduce the bias and the MSE, when the predetermined number of positive pools to be observed, is small can be considered.

iii. The proposed model was observed to be more efficient than the one-stage negative binomial group testing model with misclassification. The constructed estimator had a smaller variance, and retesting of pools suggested that the model performed well even in situations where the efficient of the test kits was low.

iv. Data from a public health study that involved the surveillance of the West Nile virus was used as a foundation to illustrate the proposed negative binomial group testing model with retesting. The results show the applicability of the proposed model in the surveillance and monitoring of infectious diseases.

5.3 Recommendations

The study has presented a new way of estimating the prevalence of a rare trait using the negative binomial group testing model with retesting. The proposed design is appealing in situations that require action to be taken early in the screening process when reporting estimates of rare traits in a population. Early interventions and accurate assessment of the prevalence level of a rare trait (e.g., infectious diseases) can lower disease mortality and the severity of an outbreak. Researchers in public health may use the findings of the study for the continuous testing and surveillance of pathogens and other infectious diseases. Future research can extend the negative binomial group testing model to a multi-stage pool testing scheme. The Bayesian approach in estimation remains an open area to be explored when imperfect tests are used under the negative binomial group testing model with retesting.

REFERENCES

Bar-Lev, S. K., Stadje, W., & Van der Duyn Schouten, F. A. (2003). Hypergeometric group testing with incomplete information. *Probability in the Engineering and Informational Sciences, 17*(3), 335-350. https://doi.org/10.1017/S0269964803173032

Berger, T., Mandell, J. W., & Subrahmanya, P. (2000). Maximally efficient two-stage screening. *Biometrics*, *56*(3), 833-840. https://doi.org/10.1111/j.0006-341X.2000.00833.x

Bilder, C. R., Tebbs, J. M., & Chen, P. (2010). Informative retesting. *Journal of the American Statistical Association*, *105*(491), 942-955. https://doi.org/10.1198/jasa.2010.ap09231

Billingsley, P. (1995). *Probability and measure* (Vol. 3). John Wiley and Sons, pp.276.

Black, M. S., Bilder, C. R., & Tebbs, J. M. (2012). Group testing in heterogeneous populations by using halving algorithms. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *61*(2), 277-290. https://doi.org/10.1111/j.1467-9876.2011.01008.x

Brookmeyer, R. (1999). Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence. *Biometrics*, *55*(2), 608-612. https://doi.org/10.1111/j.0006-341X.1999.00608.x

Busch, M.P., Caglioti, S., Robertson, E.F., McAuley, J.D., Tobler, L.H., Kamel, H., Linnen, J.M., Shyamala, V., Tomasulo, P., & Kleinman, S.H. (2005). Screening the blood supply for West Nile virus RNA by nucleic acid amplification testing. *The New England Journal of Medicine*, *353*(5), 460-467. https://doi.org/10.1056/NEJMoa044029

Callahan, J. D., Brown, F., Osorio, F. A., Sur, J. H., Kramer, E., Long, G. W., & Rock, D. L. (2002). Use of a portable real-time reverse transcriptase-polymerase chain reaction assay for rapid detection of foot-and-mouth disease virus. *Journal of the American Veterinary Medical Association*, *220*(11), 1636-1642. https://doi.org/10.2460/javma.2002.220.1636

Cardoso, M. S., Koerner, K., & Kubanek, B. (1998). Mini-pool screening by nucleic acid testing for hepatitis B virus, hepatitis C virus, and HIV: preliminary results. *Transfusion, 38*(10), 905-907. https://doi.org/10.1046/j.1537-2995.1998.381098440853.x

Delaigle, A., & Zhou, W. X. (2015). Nonparametric and parametric estimators of prevalence from group testing data with aggregated covariates. *Journal of the American Statistical Association*, *110*(512), 1785-1796. https://doi.org/10.1080/01621459.2015.1054491

Dorfman, R., (1943). The detection of defective members of large population. *Annals of Mathematical Statistics*, *14*(4), 436-440. https://doi.org/10.1214/aoms/1177731363

Fang, X., Stroup, W. W., & Zhang, S. (2007). Improved empirical bayes estimation in group testing procedure for small proportions. *Communications in Statistics-Theory and Methods*, *36*(16), 2937-2944. https://doi.org/10.1080/03610920701386935

Foppa, I. M., Evans, C. L., Wozniak, A., & Wills, W. (2007). Mosquito fauna and arbovirus surveillance in a coastal Mississippi community after Hurricane Katrina. *Journal of the American Mosquito Control Association*, *23*(2), 229-232. https://doi.org/10.2987/8756-971X(2007)23[229:MFAASI]2.0.CO;2

Gastwirth, J. L., & Hammick, P. A. (1989). Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: Application to estimating the prevalence of AIDS antibodies in blood donors. *Journal of Statistical Planning and Inference*, *22*(1), 15-27. https://doi.org/10.1016/0378-3758(89)90061-X

Haber, G., Malinovsky, Y., & Albert, P. S. (2018). Sequential estimation in the group testing problem. *Sequential Analysis*, *37*(1), 1-17. https://doi.org/10.1080/07474946.2017.1394716

Haldane, J. B. S. (1945). On a method of estimating frequencies. *Biometrika*, *33*(3), 222-225. https://doi.org/10.1093/biomet/33.3.222

Hepworth, G. (2013). Improved estimation of proportions using inverse binomial group testing. *Journal of Agricultural, Biological, and Environmental Statistics*, *18*(1), 102-119. https://doi.org/10.1007/s13253-012-0126-6

Hepworth, G., & Watson, R. (2009). Debiased estimation of proportions in group testing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *58*(1), 105-121. https://doi.org/10.1111/j.1467-9876.2008.00639.x

Hernández-Suárez, C. M., Montesinos-López, O. A., McLaren, G., & Crossa, J. (2008). Probability models for detecting transgenic plants. *Seed Science Research*, *18*(2), 77-89. https://doi.org/10.1017/S0960258508975565

Katholi, C. R., & Unnasch, T. R. (2006). Important experimental parameters for determining infection rates in arthropod vectors using pool screening approaches. *The American Journal of Tropical Medicine and Hygiene*, *74*(5), 779-785. https://doi.org/10.4269/ajtmh.2006.74.779

Kim, H. Y., & Hudgens, M. G. (2009). Three-dimensional array-based group testing algorithms. *Biometrics*, *65*(3), 903-910. https://doi.org/10.1111/j.1541-0420.2008.01158.x

Kim, H. Y., Hudgens, M. G., Dreyfuss, J. M., Westreich, D. J., & Pilcher, C. D. (2007). Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics*, *63*(4), 1152-1163. https://doi.org/10.1111/j.1541-0420.2007.00817.x

Kline, R. L., Brothers, T. A., Brookmeyer, R., Zeger, S., & Quinn, T. C. (1989). Evaluation of human immunodeficiency virus seroprevalence in population surveys using pooled sera. *Journal of Clinical Microbiology*, *27*(7), 1449-1452. https://doi.org/10.1128/jcm.27.7.1449-1452.1989

Koda, E. K. (2002). Could foot and mouth disease be a biological warfare incident? *Military Medicine*, *167*(2), 91-92. https://doi.org/10.1093/milmed/167.2.91

Lewis, J. L., Lockary, V. M., & Kobic, S. (2012). Cost savings and increased efficiency using a stratified specimen pooling strategy for *Chlamydia trachomatis* and *Neisseria gonorrhoeae*. *Sexually Transmitted Diseases*, *39*(1), 46-48. https://doi.org/10.1097/OLQ.0b013e318231cd4a

Lindan, C., Mathur, M., Kumta, S., Jerajani, H., Gogate, A., Schachter, J., & Moncada, J. (2005). Utility of pooled urine specimens for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in men attending public sexually transmitted infection clinics in Mumbai, India, by PCR. *Journal of Clinical Microbiology*, *43*(4), 1674-1677. https://doi.org/10.1128/JCM.43.4.1674-1677.2005

Litvak, Eugene, Xin M. Tu, & Marcello Pagano. (1994). Screening for the presence of a disease by pooling sera samples. *Journal of the American Statistical Association*, *89*(426), 424-434. https://doi.org/10.1080/01621459.1994.10476764

Liu, A., Liu, C., Zhang, Z., & Albert, P. S. (2012). Optimality of group testing in the presence of misclassification. *Biometrika*, *99*(1), 245-251. https://doi.org/10.1093/biomet/asr064

Luchen, L. (2012). *Interval Estimation for Binomial Proportion, Poisson Mean, and Negative-binomial Mean* [Master dissertation, Uppsala University].

Matiri, G., Nyongesa, K., & Islam, A. (2017). Sequentially Selecting Between Two Experiment for Optimal Estimation of a Trait with Misclassification. *American Journal of Theoretical and Applied Statistics*, *6*(2), 79-89. https://doi.org/10.11648/j.ajtas.20170602.12

McMahan, C. S., Tebbs, J. M., & Bilder, C. R. (2012). Informative dorfman screening. *Biometrics*, *68*(1), 287-296. https://doi.org/10.1111/j.1541-0420.2011.01644.x

Mine, H., Emura, H., Miyamoto, M., Tomono, T., Minegishi, K., Murokawa, H., & Japanese Red Cross NAT Research Group. (2003). High throughput screening of 16 million serologically negative blood donors for hepatitis B virus, hepatitis C virus, and human immunodeficiency virus type-1 by nucleic acid amplification testing with specific and sensitive multiplex reagent in Japan. *Journal of Virological Methods*, *112*(1-2), 145-151. https://doi.org/10.1016/S0166-0934(03)00215-5

Mokalled, S. C., McMahan, C. S., Tebbs, J. M., Andrew Brown, D., & Bilder, C. R. (2021). Incorporating the dilution effect in group testing regression. *Statistics in Medicine*, *40*(11), 2540-2555. https://doi.org/10.1002/sim.8916

Montesinos-López, O. A., Montesinos-Lopez, A., Crossa, J., & Eskridge, K. (2013). Sample size for detecting transgenic plants using inverse binomial group testing with dilution effect. *Seed Science Research*, *23*(4), 279-288. https://doi.org/10.1017/S0960258513000238

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., & Eskridge, K. (2012). Sample size under inverse negative binomial group testing for accuracy in parameter estimation. *PloS ONE*, *7*(3), 1-11. https://doi.org/10.1371/journal.pone.0032250

Monzon, O. T., Paladin, F. J., Dimaandal, E., Balis, A. M., Samson, C., & Mitchell, S. (1992). Relevance of antibody content and test format in HIV testing of pooled sera. *AIDS*, *6*(1), 43-48. https://doi.org/10.1097/00002030-199201000-00005

Mood, A., Graybill, F., & Boes, D. (1974). *Introduction to the Theory of Statistics* (Vol. 3). McGraw-Hill, pp. 35.

Muhua, G. O. (2010). *Estimation problem in group screening designs* [Doctoral dissertation, University of Nairobi, Kenya].

Mundel, L., (1984). Group-testing. *Journal of Quality Technology*, *16*(4), 181-188. https://doi.org/10.1080/00224065.1984.11978916

Nyongesa, L. K. (2011). Dual Estimation of Prevalence and Disease Incidence in Pool Testing Strategy. *Communication in Statistics Theory and Method*, *40*(18), 3218-3229. https://doi.org/10.1080/03610926.2010.493257

Nyongesa, L. K. (2018). Multiple-Test Pool-Testing Strategy for Estimating HIV/AIDS-Prevalence and Its Extension to Multi-Stage. *Annals of Reviews and Research*, *2*(3), 58-68. https://doi.org/10.19080/ARR.2018.02.555589

Nyongesa, L. K., (2005). Hierarchical Screening with Retesting in a low Prevalence Population. *The Indian Journal of Statistics*, *66*(4), 779-790.

Nyongesa, L.K., & Syaywa, J. P. (2010). Group Testing with Test Errors Made Easier. *Journal of Computational Statistics*, *1*(1), 1-9.

Okoth A. W., Nyongesa, L. K., & Kwatch, B.O. (2017a). A comparative study between a Multi-stage adaptive pool-testing model with test errors and the non- adaptive model. *International Journal of Applied Mathematical Research*, *6*(3), 93-97. https://doi.org/10.14419/ijamr.v6i3.7802

Okoth A. W., Nyongesa, L. K., & Kwatch, B.O. (2017b). A Multi-stage adaptive pool-testing model with Test errors; Improved efficiently. *Journal of Mathematics*, *13*(1), 43-55. https://doi.org/10.9790/5728-1301024355

Orawo, L. A. O. (2021). Confidence Intervals for the Binomial Proportion: A Comparison of Four Methods. *Open Journal of Statistics*, *11*(5), 806-816. https://doi.org/10.4236/ojs.2021.115047

Pilcher, C. D., Fiscus, S. A., Nguyen, T. Q., Foust, E., Wolf, L., Williams, D., & Hightow, L. (2005). Detection of acute infections during HIV testing in North Carolina. *New England Journal of Medicine*, *352*(18), 1873-1883. https://doi.org/10.1056/NEJMoa042291

Pritchard, N. A. (2008). *Geometric group testing* [Doctoral dissertation, University of South Carolina].

Pritchard, N. A., & Tebbs, J. M. (2011a). Bayesian inference for disease prevalence using negative binomial group testing. *Biometrical Journal*, *53*(1), 40-56. https://doi.org/10.1002/bimj.201000148

Pritchard, N.A., & Tebbs, J. M. (2011b). Estimating disease prevalence using inverse binomial pooled testing. *Journal of Agricultural, Biological, and Environmental Statistics*, *16*(1), 70-87. https://doi.org/10.1007/s13253-010-0036-4

Rodríguez-Pérez, M. A., Katholi, C. R., Hassan, H. K., & Unnasch, T. R. (2006). Large-scale entomologic assessment of *Onchocerca volvulus* transmission by pool screen PCR in Mexico. *The American Journal of Tropical Medicine and Hygiene*, *74*(6), 1026-1033. https://doi.org/10.4269/ajtmh.2006.74.1026

Rutledge, C. R., Day, J. F., Lord, C. C., Stark, L. M., & Tabachnick, W. J. (2003). West Nile virus infection rates in *Culex nigripalpus (Diptera: Culicidae)* do not reflect transmission rates in Florida. *Journal of Medical Entomology*, *40*(3), 253-258. https://doi.org/10.1603/0022-2585-40.3.253

Sarker, M. S. (2016). *Modern estimation problems in group testing* [Doctoral dissertation, University of South Carolina].

Solomon, T., Ooi, M. H., Beasley D. W., & Mallewa, M. (2003). West Nile *Encephalitis*. *British Medical Journal*, *326*(7394), 865-869. https://doi.org/10.1136/bmj.326.7394.865

Sterrett, A. (1957). On the detection of defective members of large populations. *The Annals of Mathematical Statistics*, *28*(4), 1033-1036. https://doi.org/10.1214/aoms/1177706807

Tamba C. L., Nyongesa K. L., & Mwangi J. W., (2012). Computational Pool-Testing Strategy. *Egerton University Journal*, *11*(1), 51-56.

Tebbs, J. M., McMahan, C. S., & Bilder, C. R. (2013). Two-stage hierarchical group testing for multiple infections with application to the infertility prevention project. *Biometrics, 69*(4), 1064-1073. https://doi.org/10.1111/biom.12080

Tebbs, J., Bilder, C., & Moser, B. (2003). An Empirical Bayes Group-Testing Approach to Estimating Small Proportions. *Communications in Statistics: Theory and Methods*, *32*(5), 983-995. https://doi.org/10.1081/STA-120019957

Thavaselvam, D., & Vijayaraghavan, R. (2010). Biological warfare agents. *Journal of Pharmacy and Bioallied Sciences*, *2*(3), 179. https://doi.org/10.4103/0975-7406.68499

Thompson, K. H. (1962). Estimation of the proportion of vectors in a natural population of insects. *Biometrics*, *18*(4), 568-578. https://doi.org/10.2307/2527902

Thong, A. L. S., & Shan, F. P. (2015). A comparison of several methods for the confidence intervals of negative binomial proportions. *In AIP Conference Proceedings*, *1691*(1), 50013-50021. https://doi.org/10.1063/1.4937095

Turechek, W. W., & Madden, L. V. (2003). A generalized linear modeling approach for characterizing disease incidence in a spatial hierarchy. *Phytopathology*, *93*(4), 458-466. https://doi.org/10.1094/PHYTO.2003.93.4.458

Van, T. T., Miller, J., Warshauer, D. M., Reisdorf, E., Jernigan, D., Humes, R., & Shult, P. A. (2012). Pooling *nasopharyngeal*/throat swab specimens to increase testing capacity for influenza viruses by PCR. *Journal of Clinical Microbiology*, *50*(3), 891-896. https://doi.org/10.1128/JCM.05631-11

Vine, A. E., Lewis, S. M., Dean, A. M., & Brunson, D. (2008). A critical assessment of two-stage group screening through industrial experimentation. *Technometrics*, *50*(1), 15-25. https://doi.org/10.1198/004017007000000489

Wang, D., McMahan, C. S., & Gallagher, C. M. (2015). A general regression framework for group testing data, which incorporates pool dilution effects. *Statistics in Medicine*, *34*(27), 3606-3621. https://doi.org/10.1002/sim.6578

Wanyonyi, R. W., Mwangi, O. W., & Mwangi, C. W. (2021). Re-Testing in Batch Testing Model Based on Quality Control Process for Proportion Estimation. *Open Journal of Statistics, 11*(1), 123-136. https://doi.org/10.4236/ojs.2021.111007

Wanyonyi, R. W., Nyongesa, K. L., & Wasike, A. (2015b). Estimation of Proportion of a Trait by Batch Testing Model in a Quality Control Process. *American Journal of Theoretical and Applied Statistics*, *4*(6), 619-629. https://doi.org/10.11648/j.ajtas.20150406.34

Wanyonyi, R. W., Nyongesa, L. K., & Wasike, A. (2015a). Estimation of proportion of a trait by batch testing with Errors in Inspection in a quality control process. *International Journal of Statistics and Application*, *5*(6), 268-278. https://doi.org/10.11648/j.ajtas.20150406.34

Xie, M., Tatsuoka, K., Sacks, J., & Young, S. (2001). Pool Testing with Blockers and Synergism. *Journal of American Statistical Association*, *96*(453), 92-102. https://doi.org/10.1198/016214501750333009

Xiong, W. (2016). The optimal group size using inverse binomial group testing considering misclassification. *Communications in Statistics-Theory and Methods*, *45*(15), 4600-4610. https://doi.org/10.1080/03610926.2014.923461

Yamamura, K., & Hino, A. (2007). Estimation of the proportion of defective units by using group testing under the existence of a threshold of detection. *Communications in Statistics-Simulation and Computation*, *36*(5), 949-957. https://doi.org/10.1080/03610910701539278

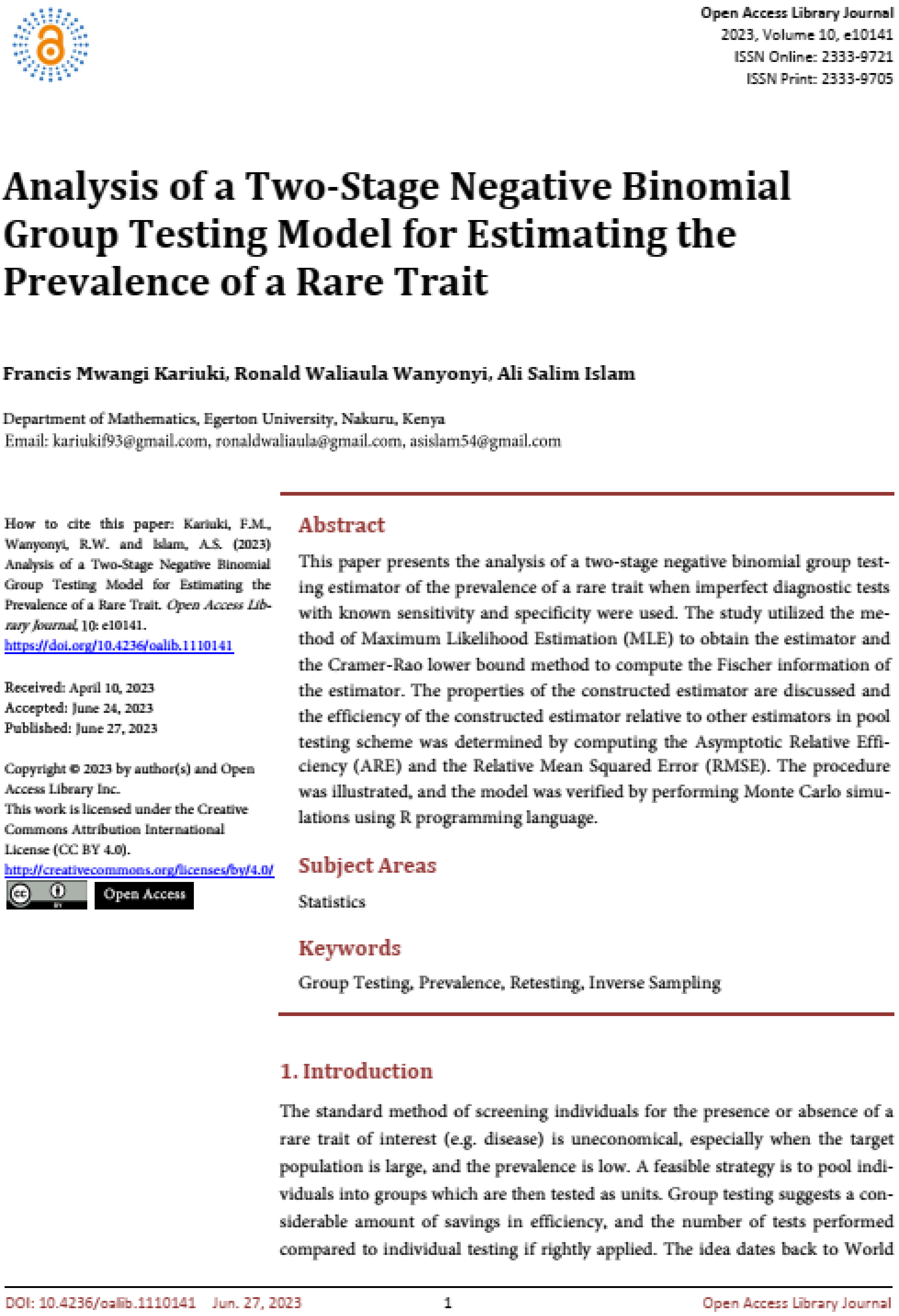
Yu, W., Xu, W., & Zhu, L. (2016). Confidence interval estimation for negative binomial group distribution. *Journal of Statistical Computation and Simulation*, *86*(3), 524-534. https://doi.org/10.1080/00949655.2015.1020807

Zhang, B., Bilder, C. R., & Tebbs, J. M. (2013). Regression analysis for multiple-disease group testing data. *Statistics in Medicine*, *32*(28), 4954-4966. https://doi.org/10.1002/sim.5858

Zilinskas R. (1997). Iraq's Biological Weapons. *Journal of the American Medical Association*. *278*(5), 418-424. https://doi.org/10.1001/jama.1997.03550050080037

APPENDICES

Appendix A: Publication



Open Access Library Journal
2023, Volume 10, e10141
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Analysis of a Two-Stage Negative Binomial Group Testing Model for Estimating the Prevalence of a Rare Trait

Francis Mwangi Kariuki, Ronald Waliaula Wanyonyi, Ali Salim Islam

Department of Mathematics, Egerton University, Nakuru, Kenya
Email: kariukif93@gmail.com, ronaldwaliaula@gmail.com, asislam54@gmail.com

How to cite this paper: Kariuki, F.M., Wanyonyi, R.W. and Islam, A.S. (2023) Analysis of a Two-Stage Negative Binomial Group Testing Model for Estimating the Prevalence of a Rare Trait. *Open Access Library Journal*, 10: e10141.
https://doi.org/10.4236/oalib.1110141

Received: April 10, 2023
Accepted: June 24, 2023
Published: June 27, 2023

Abstract

This paper presents the analysis of a two-stage negative binomial group testing estimator of the prevalence of a rare trait when imperfect diagnostic tests with known sensitivity and specificity were used. The study utilized the method of Maximum Likelihood Estimation (MLE) to obtain the estimator and the Cramer-Rao lower bound method to compute the Fischer information of the estimator. The properties of the constructed estimator are discussed and the efficiency of the constructed estimator relative to other estimators in pool testing scheme was determined by computing the Asymptotic Relative Efficiency (ARE) and the Relative Mean Squared Error (RMSE). The procedure was illustrated, and the model was verified by performing Monte Carlo simulations using R programming language.

Subject Areas

Statistics

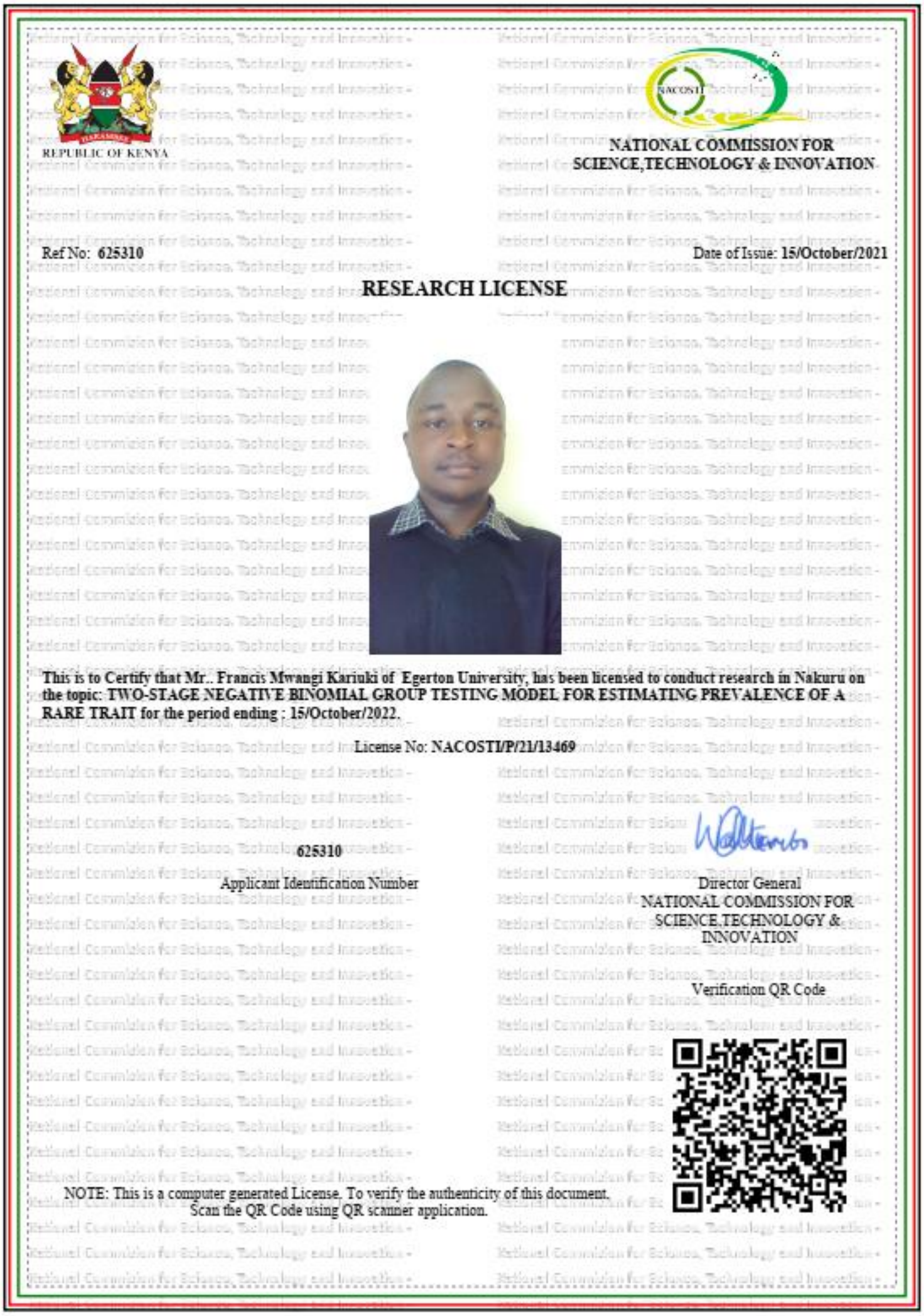
Keywords

Group Testing, Prevalence, Retesting, Inverse Sampling

1. Introduction

The standard method of screening individuals for the presence or absence of a rare trait of interest (e.g. disease) is uneconomical, especially when the target population is large, and the prevalence is low. A feasible strategy is to pool individuals into groups which are then tested as units. Group testing suggests a considerable amount of savings in efficiency, and the number of tests performed compared to individual testing if rightly applied. The idea dates back to World

Appendix B: Research Permit



REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **625310**

Date of Issue: **15/October/2021**

RESEARCH LICENSE

This is to Certify that Mr.. Francis Mwangi Kariuki of Egerton University, has been licensed to conduct research in Nakuru on the topic: TWO-STAGE NEGATIVE BINOMIAL GROUP TESTING MODEL FOR ESTIMATING PREVALENCE OF A RARE TRAIT for the period ending : 15/October/2022.

License No: **NACOSTI/P/21/13469**

625310

Applicant Identification Number

Director General

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

Appendix C: Key R codes

Notation in the R-code

ci	Confidence Interval
k	pool size
n	waiting parameter (the predetermined number of pools that tested positive on a retest)
N	Number of simulations
p	individual prevalence
Sens	The sensitivity of a test
Spec	The specificity of a test
t	Total number of pools tested until *n* positive pools that tested positive on a retest was observed after a retest.
Toler	Tolerance

Key R codes for MLE

```
P.estimate=function(n, k, p, sens, spec){
  pro= (((1-sens)^2)*(1-p)^k) + (spec^2)*(1-(1-p)^k)
  t=c()
  MLE= c(); g=c(); th=c(); out=c()
  N=10000
  for (i in 1:N) {
    t[i] = rnbinom(1, n, pro)+n
    th[i]=(((spec^2)-n/t[i])/((spec^2)-(1-sens)^2))^(1/k)
    g[i]=n/t[i]
    if(t[i]==n){
      MLE[i]= 1
    }
    else if((spec^2)<=g[i]){
      MLE[i]=1
    }
    else if(th[i]>1){
      MLE[i]=0
    }
    else{
```

```
    MLE[i] = 1-(((spec^2)-n/t[i])/((spec^2)-(1-sens)^2))^(1/k)
     }
  }
  return(mean(MLE)) }
```

Key R codes for Bias and MSE

```
MSE. Bias = function (toler, n, k, p, sens, spec){
  tstar=c();g=c();Bias.p=c();mse.p=c(); t=c();pmft=c();MLE=c()
  r= (spec^2)-(1-sens)^2
  prob= (((1-sens)^2)*(1-p)^k) + (spec^2)*(1-(1-p)^k)
  tolcheck =1-toler
  tstar= qnbinom(tolcheck,n, prob=prob)
  th= c();bias.p=c();mse.p=c(); diffl=c()
  t=seq(n, tstar+n, by=1)
  for (i in seq_along(t)) {
    pmft[i]=dnbinom((t[i]-n), n, prob=prob)
    th[i]=(((spec^2)-n/t[i])/((spec^2)-(1-sens)^2))^(1/k)
    g[i]=n/t[i]
    if(t[i]==n){
      MLE[i]= 1
    }
    else if((spec^2)<=g[i]){
      MLE[i]=1
    }
    else if(th[i]>1){
      MLE[i]=0
    }
    else{
      MLE[i] = 1-(((spec^2)-n/t[i])/((spec^2)-(1-sens)^2))^(1/k)
    }
    diffl[i]=abs(MLE[i]-p)
    Bias.p[i]=diffl[i]*pmft[i]
    mse.p[i]=(MLE[i]-p)^2 *(pmft[i])
    BIAS=sum(Bias.p)*10^4
    MSE=sum(mse.p)*10^4
    Output= as.data.frame (cbind(BIAS, MSE=MSE))
  }
  return(Output)
}
```

Key R codes for the Confidence Interval of the Estimator

```
confidence.Ret=function(n, k, p, sens, spec, gamma=0.05){
  pro= (((1-sens)^2)*(1-p)^k) + (spec^2)*(1-(1-p)^k)
  con = 1-gamma/2
  zalph = qnorm(con,0,1)
  t=c()
  p.hat= c()
  g=c()
  N=10000
  var.phat=c();  FisherInfo=c(); ul=c();ll=c();th=c()
  for (i in 1:N) {
    t[i] = rnbinom(1, n, pro)+n
    th[i]=(((spec^2)-n/t[i])/((spec^2)-(1-sens)^2))^(1/k)
    g[i]=n/t[i]
    if (t[i]==n){
      p.hat[i]= 1
      ul[i]=1
      ll[i]=0
      var.phat[i]=0
    }
    else if((spec^2)<=g[i]){
      p.hat[i]=1
      ul[i]=1
      ll[i]=0
      var.phat[i]=0
    }
    else if(th[i]>=1){
      MLE[i]=0
      FisherInfo[i]=n*(k^2)*(1-p.hat[i])^(2*k-2)*((spec^2)-(1-sens)^2)^2 * 1/((1-((1-sens)^2)
*(1-p.hat[i])^k)-(spec^2)*(1-(1-p.hat[i])^k))*1/((((1-spec)^2) *(1-p.hat[i])^k)+ (spec^2)*(1-
(1-p.hat[i])^k))^2
      var.phat[i]=(FisherInfo[i])^-1
      ul[i]=p.hat[i] + zalph*sqrt(var.phat[i])
         ll[i]= p.hat[i] - zalph*sqrt(var.phat[i])
    }
    else {
      p.hat[i] = 1-(((spec^2)-n/t[i])/((spec^2)-(1-sens)^2))^(1/k)
      FisherInfo[i]=n*(k^2)*(1-p.hat[i])^(2*k-2)*((spec^2)-(1-sens)^2)^2 * 1/((1-((1-sens)^2)
*(1-p.hat[i])^k)-(spec^2)*(1-(1-p.hat[i])^k))*1/((((1-spec)^2) *(1-p.hat[i])^k)+ (spec^2)*(1-
(1-p.hat[i])^k))^2
      var.phat[i]=(FisherInfo[i])^-1
      ul[i]=p.hat[i] + zalph*sqrt(var.phat[i])
      ll[i]= p.hat[i] - zalph*sqrt(var.phat[i])
    }
  }
  out=as.data.frame(cbind(T=t, MLE=p.hat, ll=ll, ul=ul, zalph, var.phat))
  ll.conf=mean(out$ll)
  ul.conf=mean(out$ul)
  ci=cbind(LL=ll.conf, UL=ul.conf, length.dif=ul.conf-ll.conf)
  ci
}
```