Topics in Group Testing with Multiple Infections

Peijie Hou

University of South Carolina

5-2017

Follow this and additional works at: https://scholarcommons.sc.edu/etd

Part of the Arts and Humanities Commons, and the Statistics and Probability Commons

Recommended Citation


This Open Access Dissertation is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact dillarda@mailbox.sc.edu.
Topics in Group Testing With Multiple Infections

by

Peijie Hou

Bachelor of Science
Soochow University, 2008
Master of Science
Miami University, Oxford, 2011

Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in
Statistics
College of Arts and Sciences
University of South Carolina
2017

Accepted by:
Joshua M. Tebbs, Major Professor
Dewei Wang, Major Professor
Lianming Wang, Committee Member
Alexander C. McLain, Committee Member
Cheryl Addy, Vice Provost and Dean of the Graduate School
ACKNOWLEDGMENTS

I want to express my deepest gratitude to my advisor Dr. Joshua Tebbs, for all his mentoring during my Ph.D. study. Dr. Tebbs is not only an excellent adviser, who patiently and passionately guided me through every step of academic research, but also a good friend who cares about my success in every aspect. His attitude toward research and life has inspired me to be who I am today, and for that, I am eternally grateful.

I would like to give my heartfelt thanks to Dr. Dewei Wang, who kindly agreed to be my co-advisor when I was stuck on one of my dissertation projects. Dr. Wang’s thoughtful leading has widely broadened my research horizon. Thanks to his unwaivering support, my dissertation research was made steady and solid progress.

I also want to thank my committee members Dr. Lianming Wang and Dr. Alexander McLain for carefully reviewing my dissertation. Their insightful comments and suggestions have improved the completeness and the quality of this work.

I am grateful to the Department of Statistics, for the continuous financial support during my Ph.D. study, as well as the support for attending statistical conferences, which has proven to be important for this dissertation and my career. I would also like to thank all faculty members in the department for guidance and encouragement. Without your support, I cannot accomplish this much as a graduate student.

Finally, I would like to give my special thanks to my parents, Xu Hou and Huimin Yang, and my wife, Songqiao Huang, for their endless and unconditional love. They are always there for me in any circumstances. For that, I am forever indebted to them.
ABSTRACT

Group testing, dating back to the early 1940s, was first proposed to screen for syphilis among US inductees during World War II (Dorfman, 1943). Since then, the benefits of reducing testing costs by employing group testing have been demonstrated in many areas, such as drug discovery, genetics, and infectious disease testing. Traditionally, statistical research in group testing has largely been motivated by applications involving a single infection. With the recent development of multiplex assays that can diagnose multiple infections simultaneously, generalizing the existing group testing literature to incorporate multiple infections is a natural and necessary next step. This dissertation consists of three research projects that extend group testing to multiple infections. In Chapter 2 and Chapter 3, we propose two different testing algorithms to accommodate the use of multiplex assays. Compared to the two-stage hierarchical group testing algorithms currently employed by the Infertility Prevention Project (IPP) in Iowa, our algorithms are proven to confer significant cost savings. In Chapter 4, we propose a semi-parametric regression framework to estimate individual-level marginal probability of infection from multiple infection group testing data. The performance of our testing algorithms and estimation techniques is evaluated through numerical study, simulation, and application to chlamydia and gonorrhea data collected by five states as part of the IPP.
# Table of Contents

Acknowledgments ........................................ iii

Abstract .................................................. iv

List of Tables .............................................. viii

List of Figures ............................................ xiii

Chapter 1 Introduction .................................... 1
  1.1 Literature Review ..................................... 1
  1.2 Outline ............................................... 8

Chapter 2 Hierarchical group testing for multiple infections . 10
  2.1 Introduction ......................................... 11
  2.2 Notation and Assumptions ............................. 13
  2.3 Operating Characteristics ............................. 15
  2.4 Comparisons ......................................... 21
  2.5 Region X Infertility Prevention Project Data .............. 27
  2.6 Discussion .......................................... 30

Chapter 3 Array testing with multiplex assays ............... 33
  3.1 Introduction ......................................... 34
Table 2.1 Expected number of tests per individual \( n^{-1}_1E(T(S)) \) when \( S_{e_j}^{(s)} = 0.95 \), \( S_{p_{e_j}}^{(s)} = 0.99 \), and number of stages \( S \in \{2, 3, 4, 5, 6\} \). The column labeled “Optimal” gives the configuration of \( n_1, n_2, \ldots, n_S \) that minimizes \( n^{-1}_1E(T(S)) \). The percent reduction in \( n^{-1}_1E(T(S)) \) when compared to \( n^{-1}_1E(T^{(2)}) \) is provided. The expected proportion of correct classifications \( n^{-1}_1E(C(S)) \) is also shown; see the discussion at the end of Section 2.4. The maximum allowable master pool size is 100. . . . . . . . . . . . . . . . . . . . . . . . . 23

Table 2.2 Expected number of tests per individual \( n^{-1}_1E(T(S)) \) when \( S_{e_j}^{(s)} = 0.95 \), \( S_{p_{e_j}}^{(s)} = 0.99 \), and number of stages \( S \in \{2, 3, 4, 5, 6\} \). The column labeled “Halving” gives the master pool size for the optimal halving algorithm. The percent reduction in \( n^{-1}_1E(T(S)) \) when compared to \( n^{-1}_1E(T^{(2)}) \) is provided. The expected proportion of correct classifications \( n^{-1}_1E(C(S)) \) is also shown; see the discussion at the end of Section 2.4. The maximum allowable master pool size is 100. . . . . . . . . . . . . . . . . . . . . . . . . 24

Table 2.3 Region X 2011 chlamydia and gonorrhea data for swab specimen. Average number of tests (sample standard deviation, SD) from \( B = 5000 \) sets of pools for 2-, 3-, and 4-stage hierarchical algorithms. The optimal configuration is determined by minimizing \( n^{-1}_1E(T(S)) \) using the 2010 prevalences; see Appendix A.6. The percent reduction in the average number of tests is also shown. The maximum allowable master pool size is 20. . . . . . . . . . . . . . . . . . . . . . . . . 29

Table 2.4 Region X 2011 chlamydia and gonorrhea data for urine specimen. Average number of tests (sample standard deviation, SD) from \( B = 5000 \) sets of pools for 2-, 3-, and 4-stage hierarchical algorithms. The optimal configuration is determined by minimizing \( n^{-1}_1E(T(S)) \) using the 2010 prevalences; see Appendix A.6. The percent reduction in the average number of tests is also shown. The maximum allowable master pool size is 20. . . . . . . . . . . . . . . . . . . . . . . . . 31
Table 3.1 Unbalanced prevalences. The maximum allowable pool size for IPP is 100 and for AT is $20 \times 20$. For each case, the optimal configuration $n^*$, the corresponding efficiency, $P_S$, and PPV are calculated with the formulas provided in Tebbs et al. (2013) and Appendix B.1, for IPP and AT, respectively. $PS_p$ and NPV calculations are in Appendix B.3. 42

Table 3.2 Balanced prevalences. $p_{00}$ and $p_{10}$ are the same as in the cases identified in Table 3.1, but with (nearly) same $p_{10}$ and $p_{01}$. The maximum allowable pool size for IPP is 100 and for AT is $20 \times 20$. For each case, the optimal configuration $n^*$, the corresponding efficiency, $P_S$, and PPV are calculated with the formulas provided in Tebbs et al. (2013) and Appendix B.1, for IPP and AT, respectively. $PS_p$ and NPV calculations are in Appendix B.3. 44

Table 3.3 Extremely rare prevalences. The maximum allowable pool size for IPP is 100 and for ATM is $10 \times 10$. For each case, the optimal configuration $n^*$, the corresponding efficiency, $P_S$, and PPV are calculated with the formulas provided in Tebbs et al. (2013) and Appendix B.2, for algorithm IPP and ATM, respectively. $PS_p$ and NPV calculations are in Appendix B.3. 45

Table 3.4 Region X IPP female data summary. The sample sizes $N$ are given for each state/specimen stratum. The prevalence estimates from year 2010 are used to determine the optimal configuration of IPP, AT and ATM for 2011. 48

Table 3.5 Classification results for 2011. For each state/specimen stratum, the sample size $N$ and the number of tests, which are averaged over $B = 1000$ replications, are included. The optimal configuration $n^*$ is determined using the 2010 prevalences. The savings when compared to the optimal IPP algorithm is also shown. The averaged classification accuracy measure are given in Appendix B.3. 49

Table 4.1 Simulation results for 500 data sets, each with $N = 3000$ individuals, when $(X_1, X_2, X_3)^T \sim \text{MVN}_3 (0, \Omega)$, where $\Omega_{ij} = 0.5^{i-j}$. Clayton Copula is used to generate $\eta(\cdot, \cdot)$, i.e., $\eta(x_{ij}^T \beta_1, x_{ij}^T \beta_2) = \left[ \max \left\{ (x_{ij}^T \beta_1)^{-0.5} + (x_{ij}^T \beta_2)^{-0.5} - 1; 0 \right\} \right]^{-2}$. The columns labeled “IPP” give the results for group testing, and the columns labeled “Individual” give the results for individual testing. Note that $S_{e:k} = 0.95$, $S_{p:k} = 0.95$, for $k = 1, 2$. 68
Table 4.2  Simulation results for 500 data sets, each with $N = 3000$ individuals, when $S_{e,k} = 0.95$, $S_{p,k} = 0.95$, for $k = 1, 2$, $(X_1, X_2, X_3)^T \sim \text{MVN}_3(0, \Omega)$, $\Omega_{ij} = 0.5^{|i-j|}$. The curve $\eta(\cdot, \cdot)$ is $\eta(x^T \beta_1, x^T \beta_2) = \eta_1(x^T \beta_1)\eta_2(x^T \beta_2)\sin^2 \left(2\pi(x^T \beta_1 + x^T \beta_2)\right)$. The columns labeled “IPP” give the results for group testing, and the columns labeled “Individual” give the results for individual testing. The column labeled “Optimal” gives the configuration of $n_1, n_2, \ldots, n_S$ that minimizes $n_1^{-1}E(T(S))$. The column labeled “Halving” gives the master pool size for the optimal halving algorithm. The percent reduction in $n_1^{-1}E(T(S))$ when compared to $n_1^{-1}E(T(2))$ is provided. The expected proportion of correct classifications $n_1^{-1}E(C(S))$ is also shown; see the discussion at the end of Section 2.4. The maximum allowable master pool size is 100. 

Table 4.3  Simulation results for 500 data sets, each with $N = 3000$ individuals, when $X_1 \sim N(0, 1), X_2 \sim \text{Bern}(0.6)$, and $X_3 \sim \text{Bern}(0.2)$. Clayton Copula is used to generate $\eta(\cdot, \cdot)$, i.e., $\eta(x^T \beta_1, x^T \beta_2) = \max \left\{(x^T \beta_1)^{-0.5} + (x^T \beta_2)^{-0.5} - 1; 0\right\}^{-2}$. The columns labeled “IPP” give the results for group testing, and the columns labeled “Individual” give the results for individual testing. Note that $S_{e,k} = 0.95$, $S_{p,k} = 0.95$, for $k = 1, 2$. 

Table 4.4  Simulation results for 500 data sets, each with $N = 3000$ individuals, when $S_{e,k} = 0.95$, $S_{p,k} = 0.95$, for $k = 1, 2$, $X_1 \sim N(0, 1), X_2 \sim \text{Bern}(0.6)$, and $X_3 \sim \text{Bern}(0.2)$. The curve $\eta(\cdot, \cdot)$ is $\eta(x^T \beta_1, x^T \beta_2) = \eta_1(x^T \beta_1)\eta_2(x^T \beta_2)\sin^2 \left(2\pi(x^T \beta_1 + x^T \beta_2)\right)$. The columns labeled “IPP” give the results for group testing, and the columns labeled “Individual” give the results for individual testing. 

Table 4.5  2009 Nebraska female swab estimation results. The columns labeled “IPP” give the results for group testing, and the columns labeled “Individual” give the results for individual testing. The mean estimates from “IPP” and “Individual” and standard error estimates from “IPP” are averaged over 500 data sets. 

Table A.1  Expected number of tests per individual $n_1^{-1}E(T(S))$ when $S_{e,j} = (0.95)^{1+0.5(S-s)}$, $S_{p,j} = 0.99$, and $S \in \{2, 3, 4, 5, 6\}$. The column labeled “Optimal” gives the configuration of $n_1, n_2, \ldots, n_S$ such that minimizes $n_1^{-1}E(T(S))$. The column labeled “Halving” gives the master pool size for the optimal halving algorithm. The percent reduction in $n_1^{-1}E(T(S))$ when compared to $n_1^{-1}E(T(2))$ is provided. The expected proportion of correct classifications $n_1^{-1}E(C(S))$ is also shown; see the discussion at the end of Section 2.4. The maximum allowable master pool size is 100. 

Table A.2  Classification accuracy measures when $S_{e,j} = 0.95$, $S_{p,j} = 0.99$, and number of stages $S \in \{2, 3, 4, 5, 6\}$. These are calculated from our expressions in Section 2.3 and Appendix A.3. The optimal algorithm is the same as that identified in Table 2.1.
Table A.3 Classification accuracy measures when $S_{ej} = 0.95$, $S_{pj} = 0.99$, and number of stages $S \in \{2, 3, 4, 5, 6\}$. These are calculated from our expressions in Section 2.3 and Appendix A.3. The optimal halving algorithm is the same as that identified in Table 2.2. The master pool size is $n_1$.

Table A.4 Expected number of tests per individual $n_1^{-1}E(T^{(S)})$ when $S_{ej}^{(s)} = 0.95$, $S_{pj}^{(s)} = 0.99$, and number of stages $S \in \{2, 3, 4, 5, 6\}$. The column labeled “Optimal” gives the configuration of $n_1, n_2, ..., n_S$ that maximizes $E(C^{(S)})/E(T^{(S)})$; see Malinovsky et al. (2016). The column labeled “Halving” gives the master pool size for the optimal halving algorithm. The percent reduction in $n_1^{-1}E(T^{(S)})$ when compared to $n_1^{-1}E(T^{(2)})$ is provided. The expected proportion of correct classifications $n_1^{-1}E(C^{(S)})$ is also shown. The maximum allowable master pool size is 100.

Table A.5 Region X IPP data summary (females only). Estimates are calculated from the individual test results during 2010-2011 assuming no testing error. The stratum sample sizes $N$ are also given. The 2010 prevalence estimates are used to determine the optimal hierarchical configuration for 2011 analysis shown in Table 2.3 and Table 2.4.

Table A.6 Region X IPP 2011 chlamydia and gonorrhea data. Classification accuracy (averaged over $B = 5000$ sets of pools) for 2-, 3-, and 4-stage hierarchical algorithms under the optimal configuration; see Table 2.3 and Table 2.4.

Table B.1 Values of PS$_{pk}$ and NPV$_k$ for the cases identified in Table 3.1 when $S_{ek} = 0.95$ and $S_{pj} = 0.99$. These are calculated from expressions in Tebbs et al. (2013) and Section B.1.

Table B.2 Values of PS$_{pk}$ and NPV$_k$ for the cases identified in Table 3.2 when $S_{ek} = 0.95$ and $S_{pj} = 0.99$. These are calculated from expressions in Tebbs et al. (2013) and Section B.1.

Table B.3 Values of PS$_{pk}$ and NPV$_k$ for the cases identified in Table 3.3 when $S_{ek} = 0.95$ and $S_{pj} = 0.99$. These are calculated from expressions in Tebbs et al. (2013), Section B.1 and Section B.2.
Table B.4  Classification accuracy measures for 2011 chlamydia and gonorrhea data collected in Region X IPP. For each state/specimen stratum, the sample size $N$, and the classification accuracy measures, which are averaged over $B = 1000$ replications, are included. The optimal configuration $n^*$, the averaged number of tests, and the savings compared to the optimal IPP algorithm is given in Table 3.5.
List of Figures

Figure 2.1 Hierarchical algorithm with $S = 4$ stages and master pool size $n_1 = 12$. Pools that test positively for at least one infection are split into subpools. Pools that test negatively for both infections are not split further. The last stage is individual testing where final diagnoses are made. The maximum number of pools tested in stage $s$ is $n_1/n_s$. 14

Figure 2.2 Optimal number of stages $S$ when $S_e^{(s)} = 0.95$ and $S_p^{(s)} = 0.99$. The maximum allowable master pool size is 100. In the lower left corner of each subfigure, we did not show values of $S$ larger than 6 to avoid crowding. Values of $\eta_1$ and $\eta_2$ in the white regions (barely detectable in the $\rho = 0.10$ subfigure) are not possible because correlations for binary random variables are restricted. Note that “$S = 1$” corresponds to individual testing. 25

Figure 2.3 Simulation study for the third case in Table 2.1 and Table 2.2 with $p_{00} = 0.97$, $S_e^{(s)} = 0.95$, and $S_p^{(s)} = 0.99$. Boxplots of the number of tests per individual are constructed from $B = 5000$ replications under the optimal and halving configurations. 26

Figure 3.1 A $5 \times 5$ array. $R_i$ is the vector of testing responses for the $i$th row, which contains individuals $I_{i1}, \ldots, I_{i5}$. $C_j$ is the vector of testing responses for the $j$th column, which contains individuals $I_{1j}, \ldots, I_{5j}$. 37

Figure 3.2 Optimal algorithm when $S_e^{(s)} = 0.95$ and $S_p^{(s)} = 0.99$. The maximum allowable pool size for algorithm IPP, AT, and ATM is 100, 20 $\times$ 20, and 10 $\times$ 10, respectively. Values in the white region (in $\rho = 0.10, 0.25$, and 0.50 sub-figures) are not possible because correlation for binary responses are restricted. Note that “IND” denotes the individual testing. 46
Figure A.1 Simulation study for the first and second cases in Table 2.1 and Table 2.2 with $S_{c_j}^{(s)} = 0.95$ and $S_{p_j}^{(s)} = 0.99$. Boxplots of the number of tests per individual are constructed from $B = 5000$ replications under the optimal and halving group configurations shown in Table 2.1 and Table 2.2. Top left: First case in Table 2.1 with $p_{00} = 0.90$. Top right: First case in Table 2.2 with $p_{00} = 0.95$. Bottom left: Second case in Table 2.1 with $p_{00} = 0.95$. Bottom right: Second case in Table 2.2 with $p_{00} = 0.95$.

Figure A.2 Simulation study for the fourth and fifth cases in Table 2.1 and Table 2.2. Boxplots of the number of tests per individual are constructed from $B = 5000$ replications under the optimal and halving group configurations shown in Table 2.1 and Table 2.2. Top left: Fourth case in Table 2.1 with $p_{00} = 0.99$. Top right: Fourth case in Table 2.2 with $p_{00} = 0.99$. Bottom left: Fifth case in Table 2.1 with $p_{00} = 0.999$. Bottom right: Fifth case in Table 2.2 with $p_{00} = 0.999$.

Figure A.3 Optimal number of stages $S$ when $S_{c_j}^{(s)} = 0.95$, $S_{p_j}^{(s)} = 0.99$, and optimal configurations are determined from maximizing $E(C^{(S)})/E(T^{(S)})$; see Malinovsky et al. (2016). The maximum allowable master pool size is 100. In the lower left corner of each subfigure, we did not show values of $S$ larger than 6 to avoid crowding. Values of $\eta_1$ and $\eta_2$ in the white regions (barely detectable in the $\rho = 0.10$ subfigure) are not possible because correlations for binary random variables are restricted. Note that “$S = 1$” corresponds to individual testing.

Figure A.4 Boxplots of the 5000 simulated values of $T^{(S)}$, shown cross-classified by specimen type (swab/urine) and state (AK, ID, OR, WA). The average number of tests (and sample standard deviation) in each of these 8 strata are shown in Table 2.3 and Table 2.4.

Figure A.5 Boxplots for the number of tests in Region X overall during 2011. These boxplots are formed by aggregating the number of tests expended over specimen type and state.
CHAPTER 1
INTRODUCTION

1.1 LITERATURE REVIEW

This section offers a brief literature review of group testing. We begin by familiarizing readers with the origin of group testing, and we summarize its development in the past seven decades.

A key factor in screening a large population of individuals for a binary characteristic is the testing mechanism. When the characteristic, say disease status, is rare, screening individuals one by one (i.e., individual testing) could be burdensome and expensive. An alternative method was proposed by Dorfman (1943), who introduced group testing as a method to screen for syphilis among US soldiers for syphilis. Instead of performing a Wasserman-type blood test on each prospective inductee to reveal the presence or absence of syphilitic antigen, Dorfman envisioned that multiple individual blood specimens could be pooled and tested at once. If a pool tested negatively, it indicated the pool did not contain syphilitic antigen. Hence, all individuals within the pool would be classified as negative with one test. On the other hand, if a pool tested positively, at least one of the specimens contained syphilitic antigen and further retesting would be required to identify them. We refer to the procedures used to identify positive individuals after initial group tests as a “decoding algorithm.” Dorfman’s algorithm simply retests all individuals in pools that test positively.

Since Dorfman’s seminal work, researchers have been dedicated to exploring dif-
different aspects of group testing to further realize its potential benefits. In general, research in group testing has proceeded along two different paths: case identification (classification) and estimation. The goal of case identification is to classify all individuals as either positive or negative while reducing testing costs as much as possible. The goal of estimation is to estimate the population-level prevalence or subject-specific probabilities when covariate information is available. In the following sections, we summarize the development of group testing for both case identification and estimation.

CASE IDENTIFICATION

Research in case identification has been fruitful since Dorfman’s original proposal. Although his initial decoding algorithm, i.e., retesting every individual belonging to a positive pool, is widely used in practice because of its simplicity, numerous competitive decoding algorithms have been proposed to further reduce the number of tests needed and/or to improve classification accuracy. The cost of a decoding algorithm is typically evaluated through its efficiency, that is, the expected number of tests per specimen required to identify all positive individuals. When the assay used for screening is not perfect, that is, the probability of a positive pool testing positively (sensitivity) and/or the probability of a negative pool testing negatively (specificity) is less than 1, it is also necessary to characterize the classification accuracy of an algorithm. In group testing, the classification accuracy measures that are generally of interest include but are not limited to: pooling sensitivity (denoted as PS_e), the probability of a positive individual being diagnosed as positive; pooling specificity (PS_p), the probability of a negative individual being diagnosed as negative; pooling positive predictive value (PPV), the probability an individual is truly positive given it is diagnosed as positive; and pooling negative predictive values (NPV), the probability an individual is truly negative given it is diagnosed as negative. These measurements
arise naturally in the group testing literature for case identification.

Depending on how individuals are assigned to initial groups, decoding procedures can be divided into two categories: hierarchical and non-hierarchical. Hierarchical algorithms entail assigning individual specimens to non-overlapping pools initially and then dividing positive pools into smaller, non-overlapping subpools. These sub-pools are then tested and divided again until all the individuals are classified. Non-hierarchical algorithms assign individuals to overlapping pools so that each individual falls in multiple pools. In both types, we refer to the number of sequential steps required by an algorithm as a \textit{stage}.

Dorfman’s procedure is the simplest hierarchical algorithm because it only requires two stages; pooled testing in the first stage and individual testing in the second. This procedure was later extended by Sterrett (1957) to reduce the expected number of tests per specimen. Rather than testing every individual in a positive pool, Sterrett proposed to test individuals randomly until the first positive one was identified, and remaining untested individuals would be tested as a whole. If this pool tested negatively, the procedure would stop and all individuals therein would be declared as negative; otherwise, the process would repeat until all individuals are classified. When compared to Dorfman’s algorithm, Sterrett showed that his modified algorithm could increase cost savings from 80\% to 86\% in a low prevalence setting (for a prevalence of 0.01). Litvak et al. (1994) discussed the use of a multiple stage hierarchical algorithm for HIV screening where a positive pool is repeatedly divided into two equally (or almost equally) sized subpools. Their higher-stage conceptualization is appealing in that, not only can a larger reduction in the number of tests be realized, it can also lead to a significant increase in PPV, which had been considered as the only drawback of group testing in Brennan (1991). In order to improve NPV, Litvak et al. (1994) also proposed the use of “confirmatory tests” on pools that initially tested negatively.

Array-based group testing, a non-hierarchical algorithm, was initially employed
in genetics; see, e.g., Evans and Lewis (1989), Barillot et al. (1991), Amemiya et al. (1992), and Bruno et al. (1995). Phatarfod and Sudbury (1994) later introduced array testing as a method to screen for infectious diseases. In its simplest form, individuals are initially assigned to the cells of an array. Array testing is categorized as non-hierarchical because groups are formed by pooling all individuals that reside in the same row and column, that is, each individual belongs to one row and one column. Phatarfod and Sudbury (1994) envisioned that testing rows and columns would provide the indication of possible positive individuals, for example, an individual at the intersection of a positive row and a positive column is highly likely to be positive. Thus, they structured their decoding process with two stages: at stage 1, test all the rows and columns; at stage 2, test all specimens located at these intersections individually. Berger et al. (2000) extended two-stage array testing to allow for higher dimensions. Like Phatarfod and Sudbury (1994), their work does not allow for test error, i.e., they assume a perfect assay is available. This assumption, however, is highly suspect in practice, especially in the context of infectious disease screening. Kim et al. (2007) relaxed this assumption and derived the efficiency and classification accuracy for two-dimensional array testing in the presence of test error. Later, Kim and Hudgens (2009) generalized this work for higher-dimensional array testing. Compared to their hierarchical alternatives, they showed that array testing can generally improve certain aspects of classification accuracy, without a sacrifice in efficiency.

The aforementioned work assumes that each individual has a common probability of being positive, which is often unrealistic, especially when covariate information can provide insight on the true disease status of individuals. Although exploiting population heterogeneity has attracted attention in research for estimation, it has been shown that including this information can increase the efficiency for case identification as well. Bilder et al. (2010) proposed a modification of the multi-stage algorithm
in Sterrett (1957) to account for individual covariate information. Their algorithm proceeds in the same manner as Sterrett (1957), except that individuals within a positive pool are retested according to their ordered risk probabilities. McMahan et al. (2012a) extended Dorfman’s algorithm to a heterogeneous setting. They took advantage of varying probabilities and proposed “greedy” optimization algorithms to determine the optimal pool sizes that minimize the expected number of tests. Black et al. (2012) modified the halving algorithm in Litvak et al. (1994). They suggested dividing a positive pool into two equally sized subpools according to individuals’ ordered probabilities. Individuals below the median are assigned to one subpool and individuals above the median are assigned to the other subpool; they refer to this as “ordered halving”. McMahan et al. (2012b) gave an informative version of array testing, in which they proposed two probability-based square array testing schemes. For a complete description of informative screening procedures, see Bilder and Tebbs (2012).

Historically, group testing research for case identification has been largely motivated by applications involving a single infection. However, the infectious disease literature is now replete with applications involving the use of “multiplex” assays that detect multiple infections at once. For example, it is well documented that nucleic acid testing (NAT), a popular method to simultaneously detect HIV-1 RNA, HCV RNA, and HBV DNA, can improve blood safety by reducing the window period between infection and serologic detection (McCormick et al., 2006; El Ekiaby et al., 2010; Xiao et al., 2013; FDA, 2013). In addition, several testing centers as part of the nationally implemented Infertility Prevention Project (IPP) use the Aptima Combo 2 Assay to screen pooled and individual specimens for the presence of chlamydia and gonorrhea. This test, which is also based on nucleic acid amplification technology, can discriminate between both infections in a single run. With multiplex assays gaining popularity, it is therefore important to generalize group testing algorithms
for use with multiple infections. The first literature in case identification that deals with multiple infections is Tebbs et al. (2013), who proposed a two-stage hierarchical algorithm based on the ongoing screening practice for chlamydia and gonorrhea as part of the IPP. Their groundbreaking work shows that a larger reduction in testing costs can be realized through testing for multiple infections simultaneously, and it can provide more precise prevalence estimates when compared to individual testing.

Estimation

A separate goal in group testing arises when one is interested in estimating disease positivity. Today, this would be described as the “estimation problem” in group testing. The first frequentist approach for estimation used maximum likelihood. Thompson (1962) first raised this problem to estimate the proportion of insect vectors that are capable of transmitting a pathogen. Thompson proposed to cage $k$ insects with each test plant that would later be tested for the pathogen. If positive, at least one insect is a vector; otherwise, none are vectors. Under this setting, if a randomly-selected insect has probability $p$ of being a vector, then a plant would remain non-infected with probability $(1 - p)^k$. The maximum likelihood estimator (MLE) of $p$ and its asymptotic distribution was given in Thompson (1962), under the assumptions of no inspection error, common $p$, independence among insects being examined, and no retesting. Sobel and Elashoff (1975) proposed an alternative MLE which takes retesting information into account. Swallow (1985) pointed out that the MLE of $p$ in Thompson (1962) is positively biased except when the group size is 1. In response to this work, Burrows (1987) suggested an alternative estimator that has uniformly superior bias and mean square error properties when compared to the MLE of $p$. Hughes-Oliver and Swallow (1994) presented an adaptive method to estimate $p$ which requires an a priori value $p_0$ to determine the pool size for the first stage. They showed that their method improves estimation over the non-adaptive procedures,
even when a poor $p_0$ is used. Hung and Swallow (1999) proposed two dilution effect models and a serial correlation model to address the situation where the assumption of perfect testing and independent individuals is violated. More recently, Tebbs et al. (2013) developed an EM algorithm to estimate prevalence of multiple infections in the presence of test error. Assuming the population is homogeneous, Warasi et al. (2016) proposed a Bayesian approach to estimate the prevalence of chlamydia and gonorrhea, as well as sensitivity and specificity of the assay being used for screening, using multiple infection data collected by the Nebraska IPP.

Since $p$ is usually very small in the applications involving group testing, this prior information advanced the development of Bayesian approaches. Chaubey and Li (1995) proposed two Bayes estimators of $p$ that possibly lead to smaller error when compared to the traditional MLE. For the first estimator, they used a two-parameter beta prior distribution for $p$, the probability of an individual being positive; for the second estimator, they specified a prior distribution for $1 - (1 - p)^k$, the probability of a group being positive. However, the performance of their methods highly depends on the choice of user-defined hyperparameters. To eliminate this subjectivity, Tebbs et al. (2003) pointed out that hyperparameters can be estimated with the observed data and proposed a parametric empirical Bayes estimator using a one-parameter beta prior. This work was further explored by Bilder and Tebbs (2005) in the context of vector-transfer experiments, in which they presented improved and easier-to-interpret Bayes estimators and formulated credible intervals.

Farrington (1992) first introduced the idea of modeling covariate-specific probabilities with group testing data. In this work, he suggested using a generalized linear model (GLM) with log-log link function to model individual probabilities, assuming individuals within the same group have identical covariate values. This restrictive assumption is also made in Tu et al. (1999), who proposed a Bayesian approach for prevalence estimation in a heterogeneous population. Vansteelandt et al. (2000)
generalized Farrington (1992) by allowing for heterogeneous covariates in the pools and arbitrary link functions. Xie (2001) implemented an expectation-maximization (EM) algorithm to find maximum likelihood estimates by treating the individual true statuses as latent variables. This framework is flexible because it can be used with different classes of regression functions, even outside of a GLM setting, and it takes retesting information into consideration. To assess the effect of pooling composition on parameter estimation, Bilder and Tebbs (2009) compared two non-homogeneous pooling strategies, “random” and “different,” to both homogeneous pooling and individual testing. Chen et al. (2009) regarded covariate effects as random and described a maximum likelihood method to fit mixed-effects models. More recently, Delaigle and Meister (2011) and Delaigle and Hall (2012) proposed non-parametric regression approaches. Their work was designed to incorporate a single continuous covariate and to model data obtained from the initial pool responses. Wang et al. (2014) described a general semi-parametric framework for including multiple covariates, as well as data arising from decoding. Their method falls in the category of a single-index model, which is a frequently used method to avoid the so-called “curse of dimensionality”. In the context of multiple infections, Zhang et al. (2013) developed a GEE approach to estimate individual-level probability of each infection, using only master pool testing responses. Their work also estimates a common correlation between two infections among individuals.

1.2 Outline

The overarching theme of this dissertation is to generalize the existing group testing literature to incorporate multiple infections. With modern disease screening practices, the utilization of assays that test for multiple infections has attracted considerable attention due to its effectiveness. However, statistical research in this area is notably underdeveloped. We believe that new group testing algorithms and estimation
procedures for multiple infections would be practically and statistically meaningful.

Chapter 2 and Chapter 3 focus on case identification. In Chapter 2, we generalize the two-stage hierarchical group testing algorithms for two infections described in Tebbs et al. (2013) to a general $S$-stage setting, where $S \geq 2$. We derive the operating characteristics (i.e., efficiency, accuracy) of our algorithms within a stochastic process framework by viewing the pool decoding process as a finite-state time-inhomogeneous Markov chain. We demonstrate that a higher-stage hierarchical algorithm can offer significant cost savings when infections of interest are rare. In Chapter 3, we propose two- and three-stage array-based group testing algorithms, extending their one infection analogues described in Kim et al. (2007). We illustrate the performance of our algorithms by comparing them with the two-stage hierarchical algorithm currently used in the Iowa IPP. Chapter 4 focuses on estimation. We propose a new semiparametric regression framework to estimate individual-level probability of each infection. Our framework is appealing in that it incorporates retesting information and has the flexibility to account for all types of correlation between two infections. In addition, interpretable marginal inference can be made for separate infections through our joint modeling approach, which is usually the main interest. We illustrate the performance of our framework through simulation and a data application.
Chapter 2

Hierarchical Group Testing for Multiple Infections

Summary: Group testing, where individuals are tested initially in pools, is widely used to screen a large number of individuals for rare diseases. Triggered by the recent development of assays that detect multiple infections at once, screening programs now involve testing individuals in pools for multiple infections simultaneously. Tebbs, McMahan, and Bilder (2013, Biometrics) recently evaluated the performance of a two-stage hierarchical algorithm used to screen for chlamydia and gonorrhea as part of the Infertility Prevention Project in the United States. In this article, we generalize this work to accommodate a larger number of stages. To derive the operating characteristics of higher-stage hierarchical algorithms with more than one infection, we view the pool decoding process as a time-inhomogeneous, finite-state Markov chain. Taking this conceptualization enables us to derive closed-form expressions for the expected number of tests and classification accuracy rates in terms of transition probability matrices. When applied to chlamydia and gonorrhea testing data from four states (Region X of the United States Department of Health and Human Services), higher-stage hierarchical algorithms provide, on average, an estimated 11 percent reduction in the number of tests when compared to two-stage algorithms. For applications with rarer infections, we show theoretically that this percentage reduction can be much larger.
2.1 Introduction

Group testing, also known as pooled testing, was proposed by Dorfman (1943) as a strategy to screen military recruits for syphilis during World War II. Dorfman envisioned that instead of testing each recruit’s blood specimen separately, multiple specimens could be pooled together and tested at once. Individuals from negative pools would be declared negative, and specimens from positive pools would be retested individually to identify which recruits had contracted syphilis. Over 70 years later, pooling biospecimens through group testing is commonplace in a variety of infectious disease settings. This is especially true in large-scale screening programs where, because of cost constraints or other physical limitations, there are restrictions on the number of tests that can be performed.

Dorfman’s motivation for using group testing was to reduce testing costs while still identifying all syphilitic-positive recruits. Today, this would be described as the “case identification problem,” because the goal is to identify all positive individuals among all individuals tested. Dorfman’s approach to case identification can be viewed as a two-stage hierarchical algorithm; i.e., non-overlapping pools are tested in the first stage and individuals from positive pools are tested in the second. When the disease prevalence is small, higher-stage algorithms have proven to be useful at further reducing the number of tests needed. For example, motivated by HIV testing in North Carolina, Pilcher et al. (2005) use a three-stage algorithm where individuals are first tested in a master pool of size 90. If positive, 9 non-overlapping subpools of size 10 are tested in the second stage, and individual testing is used to resolve all positive subpools in the third stage. Sherlock et al. (2007), in their survey of HIV screening practices in the United States, describe how variations of this three-stage algorithm are used in Atlanta, Los Angeles, San Francisco, and Seattle. In other applications, Kleinman et al. (2005) propose a three-stage algorithm to screen blood donors for HBV in the United States and Quinn et al. (2000) implement a four-stage algorithm.
for HIV testing in India.

Group testing research for case identification has been largely motivated by applications involving a single infection, such as HIV. However, large-scale sexually transmitted disease screening practices are rapidly moving towards the use of “multiplex assays,” that is, assays that detect multiple infections at once. For example, as part of national screening programs in the United States, several federally funded testing centers use the Aptima Combo 2 Assay (Hologic/Gen-Probe, Inc.), a nucleic acid amplification test that simultaneously detects the presence of chlamydia and gonorrhea in pooled and individual specimens (Jirsa, 2008; Lewis et al., 2012). For screening blood banks, the United States Food and Drug Administration (FDA) and the more recent infectious disease testing literature points to the development of multiplex assays that detect HIV, HBV, and HCV in pools while being able to discriminate against each one (Xiao et al., 2013; FDA, 2013). With the ongoing development of new assays and testing platforms that accommodate multiple disease screening, generalizing group testing algorithms for use with multiple infections is an important next step.

In this article, we develop $S$-stage hierarchical algorithms for multiple infections, where $S \geq 2$. Our goal is to generalize Tebbs et al. (2013), who characterized the performance of Dorfman’s two-stage ($S = 2$) algorithm for two infections. In Section 2.2, we introduce notation and state assumptions. In Section 2.3, we derive expressions for the expected number of tests and classification accuracy probabilities in a general $S$-stage hierarchical algorithm. This is accomplished by viewing the testing process from within a Markov chain framework, allowing us to characterize performance succinctly using transition probability matrices. In Section 2.4, we discuss different pool splitting strategies and show that higher-stage algorithms can be far more cost efficient than two-stage algorithms. In Section 2.5, we use chlamydia and gonorrhea testing data collected in Alaska, Idaho, Oregon, and Washington to illus-
trate the benefits of implementing higher-stage algorithms with multiple diseases. In Section 2.6, we provide a summary discussion.

To mitigate the complexity of the notation used in this article, we restrict attention herein to two infections (e.g., chlamydia and gonorrhea, etc.). We use Appendix A to show how one can quickly generalize our derivations to handle three or more infections as needed.

2.2 NOTATION AND ASSUMPTIONS

Our work is motivated by the recent development of multiplex assays that test for multiple infections. Some multiplex assays are non-discriminating; i.e., a positive result means only that at least one infection is detected. For example, the cobas TaqScreen MPX Test (Roche, Inc.) screens plasma specimens for HIV, HBV, and HCV in pools of size up to 96, but it does not determine which virus(es) is(are) detected (Ohhashi et al., 2010). On the other hand, assays are described as discriminating when upon application a diagnosis for each infection is provided separately. Most multiplex assays based on nucleic acid amplification technology used for chlamydia/gonorrhea detection discriminate between the two infections in swab and urine specimens (Gaydos et al., 2010; CDC, 2014); as noted earlier, the Aptima Combo 2 Assay is an example. For three infections, the Procleix Ultrio Assay (Hologic/Gen-Probe, Inc.) discriminates among HIV, HBV, and HCV in plasma/serum pools of size up to 16. In this article, we assume that a discriminating assay is used each time a specimen is tested (pool or individual) and that one such assay is used throughout the testing process.

An $S$-stage hierarchical algorithm begins with testing $n_1$ individuals in a master pool at stage 1. Let $n_s$ denote the pool size at the $s$th stage, where $s = 1, 2, ..., S - 1$ and $n_S = 1$. If a pool at the $s$th stage tests positively for at least one infection (excluding at stage $S$), it is split into $n_s/n_{s+1}$ subpools and each subpool is tested.
Any pool or subpool that tests negatively for both infections is not split further, and its members are declared negative for both infections. Individual testing is used in stage $S$ where final diagnoses are made. Figure 2.1 depicts the complete version of an $S = 4$ stage algorithm with master pool size $n_1 = 12$ and subpool sizes $n_2 = 6$, $n_3 = 2$, and $n_4 = 1$ at stages 2, 3, and 4, respectively.

We assume $n_s/n_{s+1}$ is a positive integer for $s = 1, 2, ..., S - 1$; i.e., pool sizes are equal within a given stage. Denote the $l$th individual by $I_l$, for $l = 1, 2, ..., n_1$. Let $\tilde{Y}_{lj} = 1$ if individual $I_l$ is truly positive for the $j$th infection, $\tilde{Y}_{lj} = 0$ otherwise, for $j = 1, 2$. We assume $\mathbf{Y}_l = (\tilde{Y}_{l1}, \tilde{Y}_{l2})'$ are independent and identically distributed with probability mass function $\text{pr}(\tilde{Y}_{l1} = \tilde{y}_1, \tilde{Y}_{l2} = \tilde{y}_2) = p_{00}^{(1-\tilde{y}_1)(1-\tilde{y}_2)}p_{10}^{(1-\tilde{y}_1)\tilde{y}_2}p_{01}^{\tilde{y}_1(1-\tilde{y}_2)}p_{11}^{\tilde{y}_1\tilde{y}_2}$, for $\tilde{y}_1, \tilde{y}_2 \in \{0, 1\}$, where $p_{00} + p_{10} + p_{01} + p_{11} = 1$. Because of potential misclassification arising from assay error, the $\tilde{Y}_l$'s are best regarded as latent. Let $G_{s,i}$ denote the $i$th pool at the $s$th stage whose true status is denoted by $\tilde{Z}_{s,i} = (\tilde{Z}_{s,i1}, \tilde{Z}_{s,i2})'$, for $s = 1, 2, ..., S$ and $i = 1, 2, ..., n_1/n_s$. At the $s$th stage, the true pool statuses $\tilde{Z}_{s,ij}$ are determined by the true statuses of those individuals within $G_{s,i}$; i.e., $\tilde{Z}_{s,ij} = 1$ if pool
$G_{s,i}$ contains at least one positive individual for the $j$th infection, $\tilde{Z}_{s,ij} = 0$ otherwise. Note that “pools” $G_{S,i}$ tested at stage $S$ contain only one individual. Finally, let $\theta_{n_s,\tilde{z}_1\tilde{z}_2}$ denote the probability a pool of size $n_s$ has true statuses $\tilde{z}_1 \in \{0, 1\}$ and $\tilde{z}_2 \in \{0, 1\}$ for the first and second infection, respectively. In Appendix A.1, we show that $\theta_{n_s,00} = p_{00}^{n_s}$, $\theta_{n_s,10} = (p_{10} + p_{00})^{n_s} - p_{00}^{n_s}$, and $\theta_{n_s,01} = (p_{01} + p_{00})^{n_s} - p_{00}^{n_s}$.

Let $S_{e,j}^{(s)}$ and $S_{p,j}^{(s)}$ denote the assay sensitivity and specificity, respectively, for the $j$th infection at the $s$th stage of testing, for $j = 1, 2$ and $s = 1, 2, \ldots, S$, and let $Z_{s,i} = (Z_{s,i1}, Z_{s,i2})'$ denote the vector of (potentially incorrect) testing outcomes for pool $G_{s,i}$. We assume all testing outcomes are mutually independent, conditional on the true statuses of the specimens being tested. This type of assumption is pervasive in the group testing literature for single infections in the presence of testing error (Litvak et al., 1994; Kim et al., 2007; Kim and Hudgens, 2009) and is used to derive relevant quantities in closed form. For further discussion on our assumptions with multiple infections, see Section 2.6. To characterize the decoding process as a Markov chain, we utilize the notion of an “ancestor pool.” For pool $G_{s,i}$ at stage $s$, denote its ancestor pool at stage $s' < s$ by $G_{s,i}^{(s')}$, for $s' = 1, 2, \ldots, s - 1$. We also use the term “parent pool” when referring to the ancestor pool at the previous stage. For example, consider pool $G_{3,2}$ in Figure 2.1, which is the second pool tested in the third stage. Both $G_{1,1}$ and $G_{2,1}$ are ancestor pools of $G_{3,2}$ and can be labeled as $G_{3,2}^{(1)}$ and $G_{3,2}^{(2)}$, respectively. Also, the master pool $G_{1,1}$ is the parent pool of $G_{2,1}$, which is the parent pool of $G_{3,2}$.

2.3 OPERATING CHARACTERISTICS

EXPECTED NUMBER OF TESTS

In an $S$-stage algorithm, a pool at stage $s + 1$, $s = 1, 2, \ldots, S - 1$, is tested only when its parent pool in stage $s$ tests positively for at least one infection. Let $T_{s+1}$ denote the number of tests expended at stage $s + 1$ so that $E(T_{s+1}) = (n_1/n_{s+1})P(Z_{s,i1} + Z_{s,i2} >$
0), for \( s = 1, 2, ..., S - 1 \), a result established in Appendix A.1. Let \( T^{(S)} \) denote the number of tests needed to classify all individuals in a master pool when using \( S \) stages. Including the master pool test and then summing over the stages, the expected value of \( T^{(S)} \) is given by

\[
E(T^{(S)}) = 1 + \sum_{s=1}^{S-1} \left( \frac{n_1}{n_{s+1}} \right) \text{pr}(Z_{s,i1} + Z_{s,i2} > 0).
\] (2.1)

The challenging part of Equation (2.1) is calculating \( \text{pr}(Z_{s,i1} + Z_{s,i2} > 0) \), the probability that pool \( G_{s,i} \) in stage \( s \) tests positively. We use a Markov chain conceptualization of the decoding process to calculate this probability, as we now describe.

If pool \( G_{s,i} \) tests positively for at least one infection, then each of its ancestor pools \( G_{s',i}^{(s')} \), \( s' = 1, 2, ..., s - 1 \), must have as well. Therefore, calculating \( \text{pr}(Z_{s,i1} + Z_{s,i2} > 0) \) for \( G_{s,i} \) requires information on all of its ancestor pools’ true statuses. At any stage, each pool has four possible true statuses, denoted by “00,” “10,” “01,” and “11.” Traversing from the master pool \( G_{s,i}^{(1)} \) to pool \( G_{s,i} \) in stage \( s \) admits a potentially large number of paths, and it is not practical to keep track of the probability of each one on a case-by-case basis. To simplify the problem, we conceptualize the true status path of \( G_{s,i}^{(1)}, G_{s,i}^{(2)}, ..., G_{s,i} \) as a Markov chain with possible states in \( \Omega = \{00, 10, 01, 11\} \). The Markov property is satisfied because transition probabilities involving true statuses depend only on those at the previous state.

To illustrate this last point, refer again to Figure 2.1. Suppose the true status of the master pool \( G_{1,1} \) is “11,” the true status of the stage 2 pool \( G_{2,1} \) is “10,” and the true status of the stage 3 pool \( G_{3,2} \) is “00.” In other words, the true status process starts in state 11, transitions to state 10 in stage 2, and then transitions to state 00 in stage 3. Given the true status of \( G_{2,1} \), the true status of \( G_{1,1} \) does not provide additional information about the true status of \( G_{3,2} \). For this specific path realization,
the joint probability can be calculated as

\[
\Pr(\tilde{Z}'_{3,2} = (0, 0), \tilde{Z}'_{2,1} = (1, 0), \tilde{Z}'_{1,1} = (1, 1))
= \Pr(\tilde{Z}'_{3,2} = (0, 0)|\tilde{Z}'_{2,1} = (1, 0)) \Pr(\tilde{Z}'_{2,1} = (1, 0)|\tilde{Z}'_{1,1} = (1, 1)) \Pr(\tilde{Z}'_{1,1} = (1, 1)).
\]

(2.2)

Note that \(\Pr(\tilde{Z}'_{3,2} = (0, 0)|\tilde{Z}'_{2,1} = (1, 0))\) and \(\Pr(\tilde{Z}'_{2,1} = (1, 0)|\tilde{Z}'_{1,1} = (1, 1))\) in Equation (2.2) can be viewed as “one-step” transition probabilities associated with the true status process. The probability \(\Pr(\tilde{Z}'_{1,1} = (1, 1)) = \theta_{n_1, 11}\) describes the initial state of the process.

To generalize this discussion; i.e., so that we can account for all possible paths, define \(M = \text{diag}(\theta_{n_1, 00}, \theta_{n_1, 10}, \theta_{n_1, 01}, \theta_{n_1, 11})\) and

\[
\pi^{(t)} = \begin{pmatrix}
\pi^{(t)}_{00 \rightarrow 00} & \pi^{(t)}_{00 \rightarrow 10} & \pi^{(t)}_{00 \rightarrow 01} & \pi^{(t)}_{00 \rightarrow 11} \\
\pi^{(t)}_{10 \rightarrow 00} & \pi^{(t)}_{10 \rightarrow 10} & \pi^{(t)}_{10 \rightarrow 01} & \pi^{(t)}_{10 \rightarrow 11} \\
\pi^{(t)}_{01 \rightarrow 00} & \pi^{(t)}_{01 \rightarrow 10} & \pi^{(t)}_{01 \rightarrow 01} & \pi^{(t)}_{01 \rightarrow 11} \\
\pi^{(t)}_{11 \rightarrow 00} & \pi^{(t)}_{11 \rightarrow 10} & \pi^{(t)}_{11 \rightarrow 01} & \pi^{(t)}_{11 \rightarrow 11}
\end{pmatrix}.
\]

The matrix \(M\) contains probabilities corresponding to the initial state of the true status process (i.e., for the master pool in stage 1). The entries in \(\pi^{(t)}\) are of the form \(\pi^{(t)}_{A \rightarrow B}\) and give the probability that the parent pool \(G^{(t)}_{t+1, i}\) in stage \(t\) transitions from state \(A\) to state \(B\) with its subpool \(G^{(t)}_{t+1, i}\) in stage \(t + 1\). For example,

\[
\pi^{(t)}_{10 \rightarrow 00} = \Pr(\tilde{Z}_{t+1, i}^{(t)} = (0, 0)|\tilde{Z}_{t+1, i}^{(t)} = (1, 0)) = \theta_{n_z, 10}^{-1}\theta_{n_{t+1, 00}}\theta_{n_z - n_{t+1, 10}},
\]

where \(\tilde{Z}_{t+1, i}^{(t)} = (\tilde{Z}_{t+1, i, 1}^{(t)}, \tilde{Z}_{t+1, i, 2}^{(t)})^{\top}\) denotes the true status of \(G^{(t)}_{t+1, i}\). In Appendix A.1, we derive expressions for each transition probability in \(\pi^{(t)}\). Because the transition matrix \(\pi^{(t)}\) characterizes the true status process, it is lower triangular. Note also that \(\pi^{(t)}\) changes from stage to stage because different stages use different pool sizes.

In the language of Markov processes, the chain identified by the true status paths of \(G^{(1)}_{s, i}, G^{(2)}_{s, i}, \ldots, G_{s, i}\) is therefore best described as time-inhomogeneous.
Joint probabilities for all possible true status paths are collected in the entries of $C = M\pi^{(1)} \pi^{(2)} \cdots \pi^{(s-1)}$. However, this matrix does not account for misclassification (which can occur at any stage), so we must augment the matrix to incorporate it. Recall that if the $s$th stage pool $G_{s,i}$ tests positively for at least one infection, then each of $G_{s,i}^{(1)}$, $G_{s,i}^{(2)}$, ..., $G_{s,i}^{(s-1)}$ must have too, even if one or more of these pools is truly negative. Therefore, we need a matrix “operator” that, at any stage, allows us to diagnose both truly positive and truly negative pools as positive for at least one infection. Under our assumptions,

$$P^{(s)} = \text{diag}(1 - S^{(s)}_{p,1} S^{(s)}_{p,2}, 1 - S^{(s)}_{e,1} S^{(s)}_{e,2}, 1 - S^{(s)}_{p,1} S^{(s)}_{e,2}, 1 - S^{(s)}_{e,1} S^{(s)}_{e,2}),$$

where $S^{(s)}_{e,j} = 1 - S^{(s)}_{e,j}$ and $S^{(s)}_{p,j} = 1 - S^{(s)}_{p,j}$ for $j = 1, 2$, is the matrix that does this at stage $s$, $s = 1, 2, ..., S - 1$. To understand what role $P^{(s)}$ plays, take, for example, the initial state matrix $M$ and post-multiply it by $P^{(1)}$ to form $MP^{(1)}$. The $(1,1)$ entry in $MP^{(1)}$, which is $\theta_{n_1,00}(1 - S^{(1)}_{p,1} S^{(1)}_{p,2})$, gives the probability a truly negative master pool (in stage 1) is incorrectly diagnosed as positive for at least one infection. Other diagonal entries in $MP^{(1)}$ have analogous interpretations, and the matrix $\pi^{(t)}P^{(t+1)}$ summarizes similar diagnosis calculations at stage $t + 1$, for $t = 1, 2, ..., s - 1$. Because pools can be diagnosed correctly or incorrectly at any stage, joint probabilities for all paths where $G_{s,i}$ tests positively for at least one infection are collected in the entries of $D = MP^{(1)}\pi^{(1)}P^{(2)}\pi^{(2)}P^{(3)} \cdots \pi^{(s-1)}P^{(s)}$.

The quadratic form $\mathbf{1}'_4 D \mathbf{1}_4$, where $\mathbf{1}'_4 = (1, 1, 1, 1)$, then adds these probabilities to obtain $\text{pr}(Z_{s,i1} + Z_{s,i2} > 0)$.

Updating our expression in Equation (2.1), we can write the expected number of tests as

$$E(T^{(S)}) = 1 + \sum_{s=1}^{S-1} \left( \frac{n_1}{n_{s+1}} \right) \mathbf{1}'_4 MP^{(1)} \prod_{t=0}^{s-1} (\pi^{(t)}P^{(t+1)}) \mathbf{1}_4,$$

where $\pi^{(0)} = (P^{(1)})^{-1}$. We include the $t = 0$ term in Equation (2.3) only so that our expression for $E(T^{(S)})$ remains correct when $S = 2$. In this case, $E(T^{(2)}) =$
$1 + n_1 \mathbf{1}' \mathbf{MP}^{(1)} \mathbf{1}_4$ reduces to Equation (1) in Tebbs et al. (2013) for two-stage Dorfman algorithms. We call $n_1^{-1} E(T^{(S)})$ the expected number of tests \textit{per individual}; this measure allows us to compare the efficiency of hierarchical algorithms using different values of $n_1$ and $S$. It is straightforward to extend Equation (2.3) to $J > 2$ infections. This is done by making obvious modifications to $\Omega$, $\pi^{(t)}$, $\mathbf{M}$, and $\mathbf{P}^{(s)}$, and then changing $\mathbf{1}_4$ to $\mathbf{1}_{2J}$. Details are provided in Appendix A.2.

**Classification Accuracy**

To complete our characterization of hierarchical algorithms for multiple infections, we derive accuracy measures commonly cited in the case identification literature. For the $j$th infection, define the \textit{pooling sensitivity} as

\[ \text{PS}_{e,j} = \text{pr}(Z_{S,ij} = 1 | \tilde{Z}_{S,ij} = 1), \]

that is, the probability an individual is classified as positive for the $j$th infection given that the individual is truly positive for the $j$th infection. The \textit{pooling specificity} $\text{PS}_{p,j}$ is defined analogously for truly negative individuals being classified negatively. An individual is classified negatively if and only if it is not classified positively in stage $S$; therefore, $\text{PS}_{p,j} = 1 - \text{pr}(Z_{S,ij} = 1 | \tilde{Z}_{S,ij} = 0)$. Deriving expressions for $\text{PS}_{e,j}$ and $\text{PS}_{p,j}$ is possible by again viewing the decoding process from within our Markov chain framework. We now illustrate this with $\text{PS}_{e,1}$ when $S > 2$.

Consider the true status path of $\mathcal{G}^{(1)}_{S,i}, \mathcal{G}^{(2)}_{S,i}, \ldots, \mathcal{G}_{S,i}$, but now, conditional on the event that each pool in this sequence contains a common individual $(\mathcal{G}_{S,i})$ that is truly positive for the first infection. For $t = 1, 2, \ldots, S - 1$, let $\tilde{Z}_{-S,i}^{(t)}$ denote the true status of pool $\mathcal{G}^{(t)}_{S,i}$ after individual $\mathcal{G}_{S,i}$ is removed. The joint probability of the true status path of $\mathcal{G}^{(1)}_{S,i}, \mathcal{G}^{(2)}_{S,i}, \ldots, \mathcal{G}_{S,i}$, conditional on the event $\{\tilde{Z}_{S,i1} = 1\}$, can be found by calculating

\[
\text{pr}(\tilde{Z}_{-S,i}^{(1)} = \tilde{z}_1, \tilde{Z}_{-S,i}^{(2)} = \tilde{z}_2, \ldots, \tilde{Z}_{-S,i}^{(S-1)} = \tilde{z}_{S-1}, \tilde{Z}_{S,i} = (1, \tilde{z}_2)' | \tilde{Z}_{S,i1} = 1) = \text{pr}(\tilde{Z}_{S,i} = (1, \tilde{z}_2)' | \tilde{Z}_{S,i1} = 1) \text{pr}(\tilde{Z}_{-S,i}^{(1)} = \tilde{z}_1, \tilde{Z}_{-S,i}^{(2)} = \tilde{z}_2, \ldots, \tilde{Z}_{-S,i}^{(S-1)} = \tilde{z}_{S-1}), \quad (2.4)
\]
where \( \tilde{z}'_1, \tilde{z}'_2, \ldots, \tilde{z}'_{S-1} \in \{(0,0), (1,0), (0,1), (1,1)\} \) and \( \tilde{z}_2 \in \{0,1\} \). The first probability on the right-hand side of Equation (2.4) is \( p_{1z_2}/(p_{10} + p_{11}) \). The second probability is calculated by recognizing the Markov structure of \( G^{(1)}_{S,i}, G^{(2)}_{S,i}, \ldots, G^{(S-1)}_{S,i} \) that emerges after removing \( G_{S,i} \). That is, the same conceptualization we exploited in calculating \( E(T^{(S)}) \) applies and probabilities of the form \( \text{pr}(\tilde{z}_{-S,i}^{(1)}) = \tilde{z}_1, \tilde{z}_{S,i}^{(2)} = \tilde{z}_2, \ldots, \tilde{z}_{S-1}^{(S-1)} = \tilde{z}_{S-1} \) are collected in the entries of \( C_{-1} = M_{-1}\pi_{-1}^{(1)}\pi_{-1}^{(2)} \cdots \pi_{-1}^{(S-2)} \). The matrices \( M_{-1} \) and \( \pi_{-1}^{(t)} \) are the same as \( M \) and \( \pi^{(t)} \) in Section 3.1, respectively, except that all pool sizes are reduced by one.

To complete our derivation, all that remains is to incorporate the effect of misclassification that can occur at any stage. Misclassification can arise due to either infection, so the two values of \( \tilde{z}_2 \in \{0,1\} \) in Equation (2.4) must be treated separately. If \( \tilde{z}_2 = 0 \), then \( G^{(t)}_{S,i} \) must be truly positive for the first infection, because \( \tilde{Z}_{S,i}^{(t)} = 1 \) by assumption, and the second infection’s true status is determined by \( \tilde{Z}_{S,i}^{(t)} \). If \( \tilde{z}_2 = 1 \), then each pool in the sequence \( G^{(1)}_{S,i}, G^{(2)}_{S,i}, \ldots, G^{(S-1)}_{S,i} \) must be truly positive for both infections. To cover both cases, respectively, we define the two matrix operators \( P_{+-}^{(s)} = \text{diag}(1 - S_{e:1}^{(s)}S_{e:2}^{(s)}, 1 - S_{e:1}^{(s)}S_{e:2}^{(s)}, 1 - S_{e:1}^{(s)}S_{e:2}^{(s)}, 1 - S_{e:1}^{(s)}S_{e:2}^{(s)} \) and \( P_{++}^{(s)} = (1 - S_{e:1}^{(s)}S_{e:2}^{(s)})I_4 \), where \( I_4 \) is the \( 4 \times 4 \) identity matrix. The matrices \( P_{+-}^{(s)} \) and \( P_{++}^{(s)} \) then augment \( C_{-1} \) accordingly for the two values of \( \tilde{z}_2 \in \{0,1\} \) in the same way \( P^{(s)} \) augmented \( C \) in previous section. Adding up the probabilities for all transition paths, we obtain

\[
\text{PS}_{e:1} = \left( \frac{p_{10}}{p_{10} + p_{11}} \right) I_4M_{-1}P_{+-}^{(1)} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)}P_{+-}^{(t+1)})I_4S_{e:1}^{(S)} + \left( \frac{p_{11}}{p_{10} + p_{11}} \right) I_4M_{-1}P_{++}^{(1)} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)}P_{++}^{(t+1)})I_4S_{e:1}^{(S)}. \tag{2.5}
\]

The additional “\( S_{e:1}^{(S)} \)” in the expression for \( \text{PS}_{e:1} \) accounts for the final diagnosis at stage \( S \) where individual testing occurs.

The preceding derivation also applies when \( S = 2 \); i.e., for the Dorfman-type algorithm in Tebbs et al. (2013). The only difference is that \( \prod_{t=1}^{S-2} (\pi_{-1}^{(t)}P_{+-}^{(t+1)}) \) and
\[ \Pi_{i=1}^{S-2} (\pi_{-i}^{(t)} P_{++}^{(t+1)}) \] in Equation (2.5) are replaced by identity matrices. Furthermore, as shown in Appendix A.3 and A.4, general expressions for \( \text{PS}_{e2}, 1 - \text{PS}_{p1}, \text{and } 1 - \text{PS}_{p2} \) all possess the same form as \( \text{PS}_{e1} \); i.e., each one can be written as a convex combination of two quadratic forms. Each quantity is derived by exploiting the Markov structure of \( G_{S,i}^{(1)}, G_{S,i}^{(2)}, \ldots, G_{S,i}^{(S-1)} \) that arises after removing one individual. This structure remains regardless of the number of infections considered, so generalizing these expressions when \( J > 2 \) is also straightforward.

Two additional measures of classification accuracy are the pooling positive predictive value and the pooling negative predictive value. For the \( j \)th infection, these are given by

\[
\text{PPV}_j = \frac{\eta_j \text{PS}_{e_j}}{\eta_j \text{PS}_{e_j} + (1 - \eta_j)(1 - \text{PS}_{p_j})} \quad \text{and} \quad \text{NPV}_j = \frac{(1 - \eta_j)\text{PS}_{p_j}}{(1 - \eta_j)\text{PS}_{p_j} + \eta_j(1 - \text{PS}_{e_j})},
\]

respectively, where \( \eta_1 = p_{10} + p_{11} \) and \( \eta_2 = p_{01} + p_{11} \) are the marginal probabilities. In words, \( \text{PPV}_j \) (\( \text{NPV}_j \)) gives the probability that an individual is truly positive (negative) for the \( j \)th infection given that the individual has been classified positively (negatively) for the \( j \)th infection. Expressions for \( \text{PPV}_j \) and \( \text{NPV}_j \) are found by using Bayes’ Rule.

2.4 Comparisons

We now compare hierarchical algorithms that use a different number of stages. For an \( S \)-stage algorithm, we first identify the optimal configuration of \( n_1, n_2, \ldots, n_S \) for given values of \( p_{00}, p_{10}, p_{01}, \text{and } p_{11} \). In this article, we define the “optimal” configuration as the one that minimizes \( n_1^{-1} E(T^{(S)}) \), the expected number of tests per individual, subject to the constraint that \( (n_1, n_2, \ldots, n_S)' \) resides in

\[
\emptyset = \{(n_1, n_2, \ldots, n_S)': n_s/n_{s+1} \in \mathbb{N}_{>1}, \ s = 1, 2, \ldots, S - 1; \ n_S = 1\},
\]

where \( \mathbb{N}_{>1} = \{2, 3, \ldots, \} \). The condition \( n_s/n_{s+1} \in \mathbb{N}_{>1} \) simply ensures that pool sizes will be common within a given stage. Because extremely large pool sizes are rarely
seen in the infectious disease testing literature, we assume the master pool size $n_1$ is no larger than 100. This restriction was also used by Kim and Hudgens (2009) who evaluated the utility of higher-stage array group testing algorithms for single infections. For us, this restriction puts a constraint on the space of possible configurations and allows us to identify the optimal one using a direct search. Hierarchical algorithms which implement halving; i.e., $n_s/n_{s+1} = 2$, for $s = 1, 2, \ldots, S - 2$ and $n_S = 1$, arise as a special case. Halving algorithms for single infections were highlighted by Litvak et al. (1994) and Black et al. (2012).

In Table 2.1 and Table 2.2, we calculate the expected number of tests per individual for different values of $S$ under different configurations of $p_{00}$, $p_{10}$, $p_{01}$, and $p_{11}$ with $S_{e,j}^{(s)} = 0.95$ and $S_{p,j}^{(s)} = 0.99$, for $j = 1, 2$ and $s = 1, 2, \ldots, S$. To evaluate the performance of algorithms with different levels of disease prevalence, we let $p_{00} \in \{0.90, 0.95, 0.97, 0.99, 0.999\}$ and vary the other probabilities accordingly. Values of $p_{00} = 0.90, 0.95$ were chosen to be consistent with our chlamydia and gonorrhea application in Section 2.5. Values of $p_{00} = 0.99, 0.999$ were chosen to emulate what would occur when the two infections are very rare (e.g., HIV-1 and HIV-2, etc.). For each setting, we calculate the overall optimal testing configuration by minimizing $n_1^{-1}E(T^{(S)})$ (Table 2.1) and, separately, the master pool size that corresponds to the most efficient use of halving (Table 2.2). We kept $S_{e,j}^{(s)} = 0.95$ and $S_{p,j}^{(s)} = 0.99$ constant across the stages for simplicity. Proper assay calibration and/or the adjustment of dilution ratios would be needed to make this assumption reasonable; see McMahan et al. (2013) and the references therein. Appendix A.5 contains additional results where $S_{e,j}^{(s)}$ varies across stages.

Our calculations in Table 2.1 and Table 2.2 show that as the combined disease prevalence decreases ($p_{00}$ increases), higher-stage algorithms for multiple infections can markedly reduce the value of $n_1^{-1}E(T^{(S)})$. For example, when $p_{00} = 0.97$ and the marginal disease probabilities $\eta_1 = p_{10} + p_{11}$ and $\eta_2 = p_{01} + p_{11}$ are each 0.02
The maximum allowable master pool size is 100. The configuration of \( n_1, n_2, ..., n_S \) that minimizes \( n_1^{-1}E(T(S)) \) is also shown; see the discussion at the end of Section 2.4. The expected proportion of correct classifications \( n_1^{-1}E(C(S)) \) is also shown.

<table>
<thead>
<tr>
<th>( S )</th>
<th>Optimal</th>
<th>( n_1^{-1}E(T(S)) )</th>
<th>% Reduction</th>
<th>( n_1^{-1}E(C(S)) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_{00} = 0.90 )</td>
<td>2</td>
<td>4 : 1</td>
<td>0.593</td>
<td>--</td>
</tr>
<tr>
<td>( p_{10} = 0.05 )</td>
<td>3</td>
<td>9 : 3 : 1</td>
<td>0.569</td>
<td>4.0</td>
</tr>
<tr>
<td>( p_{01} = 0.04 )</td>
<td>4</td>
<td>99 : 9 : 3 : 1</td>
<td>0.577</td>
<td>2.7</td>
</tr>
<tr>
<td>( p_{11} = 0.01 )</td>
<td>5</td>
<td>90 : 45 : 9 : 3 : 1</td>
<td>0.595</td>
<td>-0.3</td>
</tr>
<tr>
<td>( p_{00} = 0.95 )</td>
<td>2</td>
<td>5 : 1</td>
<td>0.433</td>
<td>--</td>
</tr>
<tr>
<td>( p_{10} = 0.03 )</td>
<td>3</td>
<td>9 : 3 : 1</td>
<td>0.371</td>
<td>14.3</td>
</tr>
<tr>
<td>( p_{01} = 0.01 )</td>
<td>4</td>
<td>18 : 6 : 3 : 1</td>
<td>0.370</td>
<td>14.5</td>
</tr>
<tr>
<td>( p_{11} = 0.01 )</td>
<td>5</td>
<td>90 : 18 : 6 : 3 : 1</td>
<td>0.377</td>
<td>12.9</td>
</tr>
<tr>
<td>( p_{00} = 0.97 )</td>
<td>2</td>
<td>7 : 1</td>
<td>0.345</td>
<td>--</td>
</tr>
<tr>
<td>( p_{10} = 0.01 )</td>
<td>3</td>
<td>16 : 4 : 1</td>
<td>0.273</td>
<td>20.9</td>
</tr>
<tr>
<td>( p_{01} = 0.01 )</td>
<td>4</td>
<td>27 : 9 : 3 : 1</td>
<td>0.260</td>
<td>24.6</td>
</tr>
<tr>
<td>( p_{11} = 0.01 )</td>
<td>5</td>
<td>36 : 12 : 6 : 3 : 1</td>
<td>0.264</td>
<td>23.5</td>
</tr>
<tr>
<td>( p_{00} = 0.99 )</td>
<td>2</td>
<td>11 : 1</td>
<td>0.209</td>
<td>--</td>
</tr>
<tr>
<td>( p_{10} = 0.004 )</td>
<td>3</td>
<td>25 : 5 : 1</td>
<td>0.135</td>
<td>35.4</td>
</tr>
<tr>
<td>( p_{01} = 0.004 )</td>
<td>4</td>
<td>48 : 12 : 4 : 1</td>
<td>0.117</td>
<td>44.0</td>
</tr>
<tr>
<td>( p_{11} = 0.002 )</td>
<td>5</td>
<td>81 : 27 : 9 : 3 : 1</td>
<td>0.112</td>
<td>46.4</td>
</tr>
<tr>
<td>( p_{00} = 0.999 )</td>
<td>2</td>
<td>33 : 1</td>
<td>0.081</td>
<td>--</td>
</tr>
<tr>
<td>( p_{10} = 0.0004 )</td>
<td>3</td>
<td>99 : 11 : 1</td>
<td>0.032</td>
<td>60.5</td>
</tr>
<tr>
<td>( p_{01} = 0.0004 )</td>
<td>4</td>
<td>96 : 24 : 6 : 1</td>
<td>0.024</td>
<td>70.4</td>
</tr>
<tr>
<td>( p_{11} = 0.0002 )</td>
<td>5</td>
<td>96 : 48 : 16 : 4 : 1</td>
<td>0.023</td>
<td>71.6</td>
</tr>
<tr>
<td>( p_{00} = 0.9999 )</td>
<td>2</td>
<td>33 : 1</td>
<td>0.081</td>
<td>--</td>
</tr>
<tr>
<td>( p_{10} = 0.00004 )</td>
<td>3</td>
<td>99 : 11 : 1</td>
<td>0.032</td>
<td>60.5</td>
</tr>
<tr>
<td>( p_{01} = 0.00004 )</td>
<td>4</td>
<td>96 : 24 : 6 : 1</td>
<td>0.024</td>
<td>70.4</td>
</tr>
<tr>
<td>( p_{11} = 0.00002 )</td>
<td>5</td>
<td>96 : 48 : 16 : 4 : 1</td>
<td>0.023</td>
<td>71.6</td>
</tr>
</tbody>
</table>

The third case in Table 2.1, the optimal hierarchical algorithm uses \( S = 4 \) stages (with pool sizes \( n_1 = 27, n_2 = 9, n_3 = 3, \) and \( n_4 = 1 \)) and confers a 24.6% reduction in the expected number of tests per individual when compared to the optimally sized Dorfman algorithm from Tebbs et al. (2013). The optimal halving algorithm in this same setting uses \( S = 5 \) stages (with master pool size \( n_1 = 24 \)) and confers a 23.2% reduction when compared to the best Dorfman algorithm. Those cases involving rarer infections (i.e., \( p_{00} = 0.99, 0.999 \)) provide even larger reductions. To provide a panoptic examination, we display in Figure 2.2 the best number of stages
allowable master pool size is 100.

compared to pool size for the optimal halving algorithm. The percent reduction in there is a sizeable subset of the parameter space for which higher-stage designs are in Figure 2.2 identify the number of stages \( \eta \) to use when the marginal disease probabilities \( \eta_1 \) and \( \eta_2 \) range from 0.001 to 0.20, \( S^{(s)}_{e_{ij}} = 0.95 \) and \( S^{(s)}_{p_{ij}} = 0.99 \), and the correlation between the true disease statuses \( \rho = \text{corr}(\bar{Y}_{1i}, \bar{Y}_{12}) \) is fixed at \( \rho = 0.10 \) and \( \rho = 0.25 \). At each configuration of \( \eta_1 \) and \( \eta_2 \), the optimal hierarchical algorithm is determined for each \( S \geq 2 \), and the regions in Figure 2.2 identify the number of stages \( S \) that minimizes \( n_1^{-1}E(T^{(S)}) \). Clearly, there is a sizeable subset of the parameter space for which higher-stage designs are more efficient than those that use only two stages.

To better understand how hierarchical algorithms will perform in practice, we

<table>
<thead>
<tr>
<th>( S )</th>
<th>Halving</th>
<th>( n_1^{-1}E(T^{(S)}) )</th>
<th>% Reduction</th>
<th>( n_1^{-1}E(C^{(S)}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_{00} = 0.90 )</td>
<td>2</td>
<td>( n_1 = 4 )</td>
<td>0.593</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>( n_1 = 6 )</td>
<td>0.574</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>( n_1 = 12 )</td>
<td>0.595</td>
<td>-0.3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>( n_1 = 24 )</td>
<td>0.620</td>
<td>-4.6</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>( n_1 = 48 )</td>
<td>0.637</td>
<td>-7.4</td>
</tr>
<tr>
<td>( p_{00} = 0.95 )</td>
<td>2</td>
<td>( n_1 = 5 )</td>
<td>0.433</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>( n_1 = 8 )</td>
<td>0.385</td>
<td>11.1</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>( n_1 = 12 )</td>
<td>0.373</td>
<td>13.9</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>( n_1 = 24 )</td>
<td>0.381</td>
<td>12.0</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>( n_1 = 48 )</td>
<td>0.392</td>
<td>9.5</td>
</tr>
<tr>
<td>( p_{00} = 0.97 )</td>
<td>2</td>
<td>( n_1 = 7 )</td>
<td>0.345</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>( n_1 = 10 )</td>
<td>0.289</td>
<td>16.2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>( n_1 = 16 )</td>
<td>0.269</td>
<td>22.0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>( n_1 = 24 )</td>
<td>0.265</td>
<td>23.2</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>( n_1 = 32 )</td>
<td>0.272</td>
<td>21.2</td>
</tr>
<tr>
<td>( p_{00} = 0.99 )</td>
<td>2</td>
<td>( n_1 = 11 )</td>
<td>0.209</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>( n_1 = 16 )</td>
<td>0.156</td>
<td>25.4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>( n_1 = 24 )</td>
<td>0.131</td>
<td>37.3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>( n_1 = 32 )</td>
<td>0.118</td>
<td>43.5</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>( n_1 = 48 )</td>
<td>0.113</td>
<td>45.9</td>
</tr>
<tr>
<td>( p_{00} = 0.999 )</td>
<td>2</td>
<td>( n_1 = 33 )</td>
<td>0.081</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>( n_1 = 48 )</td>
<td>0.046</td>
<td>43.2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>( n_1 = 68 )</td>
<td>0.034</td>
<td>58.0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>( n_1 = 96 )</td>
<td>0.027</td>
<td>66.7</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>( n_1 = 96 )</td>
<td>0.023</td>
<td>71.6</td>
</tr>
</tbody>
</table>
Figure 2.2: Optimal number of stages $S$ when $S_{ej}^{(s)} = 0.95$ and $S_{p,j}^{(s)} = 0.99$. The maximum allowable master pool size is 100. In the lower left corner of each subfigure, we did not show values of $S$ larger than 6 to avoid crowding. Values of $\eta_1$ and $\eta_2$ in the white regions (barely detectable in the $\rho = 0.10$ subfigure) are not possible because correlations for binary random variables are restricted. Note that “$S = 1$” corresponds to individual testing.

Conducted a simulation study to assess the variability in the number of tests expended on a per-individual basis. For each parameter configuration in Table 2.1 and Table 2.2, we first generated the true infection statuses of 100,000 individuals according to the specified cell probabilities. This sample size was chosen to be comparable to our data application in Section 2.5. Under each optimal and halving configuration in Table 2.1 and Table 2.2, we assigned our 100,000 individuals to pools, performed our hierarchical algorithms using $S_{ej}^{(s)} = 0.95$ and $S_{p,j}^{(s)} = 0.99$, and recorded the number of tests per individual. This process was repeated $B = 5000$ times for each design. For the third case in Table 2.1 and Table 2.2 where $p_{00} = 0.97$, Figure 2.3 displays boxplots of 5000 values of the number of tests per individual for each number of stages $S$. One notes that the variation in the number of tests per individual for this case is fairly constant across the values of $S$ and that higher-stage algorithms are always preferred. Similarly constructed figures for the other four parameter configurations in Table 2.1 and Table 2.2 are provided in Appendix A.5.
Figure 2.3: Simulation study for the third case in Table 2.1 and Table 2.2 with $p_{00} = 0.97$, $S_{e;j} = 0.95$, and $S_{p;j} = 0.99$. Boxplots of the number of tests per individual are constructed from $B = 5000$ replications under the optimal and halving configurations.

Finally, a comparison of the classification accuracy measures derived in Section 2.3 is given in Appendix A.5 under the same settings as in Table 2.1 and Table 2.2. This comparison shows that pooling sensitivity $PS_{e;j}$ decreases as the number of stages $S$ increases, as expected, but not as rapidly as it would in $S$-stage hierarchical algorithms for single infections where the pooling sensitivity equals $\prod_{s=1}^{S} S_{e;j}^{(s)}$. In fact, provided that $S_{e;j}^{(s)} < 1$ for $s = 1, 2, ..., S$, one can show algebraically that $PS_{e;j} > \prod_{s=1}^{S} S_{e;j}^{(s)}$ for all $S \geq 2$, an important additional benefit of using hierarchical algorithms with multiplex assays. Also, the pooling positive predictive value $PPV_{j}$ increases in higher-stage algorithms for multiple infections, substantially so when both infections are rare. Values of $PS_{p;j}$ and $NPV_{j}$ remain fairly constant across values of $S$.

We conclude this section with a remark. While we have used the expected number of tests per individual $n^{-1}_1 E(T^{(S)})$ to determine optimal group configurations in this section, other objective functions which incorporate classification accuracy could be used. Based on the recommendations of anonymous referees, we have also determined optimal configurations in this section by maximizing $E(C^{(S)}) / E(T^{(S)})$, where $C^{(S)}$...
denotes the number of individuals correctly classified in a master pool tested in $S$ stages. This type of objective function was recommended by Malinovsky et al. (2016) for single infections and two-stage testing. In Appendix A.5, we use our Markov chain framework to derive $E(C_{(S)})$ for multiple infections with any number of stages, and we reproduce Table 2.1, Table 2.2 and Figure 2.2 using the configurations obtained from maximizing $E(C_{(S)})/E(T_{(S)})$. For the cases we considered in this section, there is very strong agreement between the two approaches.

### 2.5 Region X Infertility Prevention Project Data

The Infertility Prevention Project (IPP) was a national program that started in 1988 and was implemented in all 50 states. The purpose of the program was to screen individuals for chlamydia and gonorrhea in high-risk populations and to offer treatment services for those who were infected. Chlamydia and gonorrhea are two of the most common sexually transmitted diseases in the United States with approximately 1.6 million new infections reported each year (CDC, 2014). The IPP, which was funded by the Department of Health and Human Services (HHS) and overseen by the Centers for Disease Control and Prevention (CDC), was discontinued in 2013 after the Affordable Care Act was passed. This has since forced STD clinics and public health laboratories nationwide to rely on other sources of external funding (e.g., private health insurance, Medicaid, etc.) for the purpose of screening these same high-risk populations. As a result, public-health officials have experienced increased pressure to be mindful of testing costs (JSI Research & Training Institute, Inc., 2013).

Because chlamydia and gonorrhea remain moderately rare even in higher-risk populations, our higher-stage hierarchical algorithms emerge as excellent candidates to further reduce the number of tests. Public health laboratories in multiple states have used two-stage Dorfman algorithms with multiplex assays to screen for chlamydia and gonorrhea (Jirsa, 2008; Lewis et al., 2012), and Tebbs et al. (2013) show this provides
a sizeable reduction in the number of tests when compared to individual testing. Our goal is to determine if higher-stage algorithms (i.e., $S > 2$) can provide additional savings. To accomplish this, we use chlamydia and gonorrhea data collected from HHS Region X during 2010-2011. Region X consists of four states, Alaska, Idaho, Oregon, and Washington, and our data set contains about 260,000 individual testing results for both chlamydia and gonorrhea among these states (roughly 130,000 individuals each year). Because approximately 99% of the testing results were obtained from using the Aptima Combo 2 Assay, we focus on these individuals in our analysis.

To illustrate the potential use of higher-stage algorithms, we use female specimens only. Male subjects are more likely to be tested only when they exhibit symptoms of infection (e.g., painful urination, etc.), resulting in much higher positivity rates and therefore making higher-stage testing less attractive. On the other hand, females are routinely screened as part of annual health examinations and visits to family-planning health centers. In Appendix A.6, we provide the observed prevalences for the 107,463 females tested in 2010, cross-classified by specimen type (swab/urine) and state within Region X. We also provide values of the Aptima Combo 2 Assay sensitivity and specificity for each infection; these values were taken from the most recent product literature available at the manufacturer’s website.

Using the 103,690 females tested in 2011, we investigate the performance of hierarchical algorithms with $S = 2$, 3, and 4 stages. For each state and within specimen type, we randomly assign the 2011 individuals to master pools under the optimal testing configuration which we determine using the 2010 prevalences. In doing so, we set the maximum allowable master pool size at 20, because documented applications of group testing for chlamydia and gonorrhea do not use pool sizes larger than this. In order to measure classification accuracy, we treat the 2011 individuals’ responses as the “true” statuses; we then test and decode pools ourselves by simulating test outcomes using the assay accuracies reported for the Aptima Combo 2 Assay at each
### Table 2.3: Region X 2011 chlamydia and gonorrhea data for swab specimen. Average number of tests (sample standard deviation, SD) from $B = 5000$ sets of pools for 2-, 3-, and 4-stage hierarchical algorithms. The optimal configuration is determined by minimizing $n_1^{-1}E(T^{(S)})$ using the 2010 prevalences; see Appendix A.6. The percent reduction in the average number of tests is also shown. The maximum allowable master pool size is 20.

<table>
<thead>
<tr>
<th>State</th>
<th># Stages</th>
<th>Configuration</th>
<th># Tests (SD)</th>
<th>% Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>$S = 2$</td>
<td>5 : 1</td>
<td>1509.9 (30.7)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>$S = 3$</td>
<td>9 : 3 : 1</td>
<td>1343.7 (31.2)</td>
<td>11.0</td>
</tr>
<tr>
<td></td>
<td>$S = 4$</td>
<td>18 : 6 : 3 : 1</td>
<td>1352.4 (38.5)</td>
<td>10.4</td>
</tr>
<tr>
<td>Idaho</td>
<td>$S = 2$</td>
<td>5 : 1</td>
<td>3938.0 (49.9)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>$S = 3$</td>
<td>9 : 3 : 1</td>
<td>3511.3 (51.1)</td>
<td>10.8</td>
</tr>
<tr>
<td></td>
<td>$S = 4$</td>
<td>18 : 6 : 3 : 1</td>
<td>3516.0 (66.0)</td>
<td>10.7</td>
</tr>
<tr>
<td>Oregon</td>
<td>$S = 2$</td>
<td>5 : 1</td>
<td>19633.1 (108.9)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>$S = 3$</td>
<td>9 : 3 : 1</td>
<td>17322.5 (112.1)</td>
<td>11.8</td>
</tr>
<tr>
<td></td>
<td>$S = 4$</td>
<td>18 : 6 : 3 : 1</td>
<td>17272.5 (140.2)</td>
<td>12.0</td>
</tr>
<tr>
<td>Washington</td>
<td>$S = 2$</td>
<td>5 : 1</td>
<td>10497.1 (80.4)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>$S = 3$</td>
<td>9 : 3 : 1</td>
<td>9199.5 (81.0)</td>
<td>12.4</td>
</tr>
<tr>
<td></td>
<td>$S = 4$</td>
<td>18 : 6 : 3 : 1</td>
<td>9162.6 (103.8)</td>
<td>12.7</td>
</tr>
</tbody>
</table>

This entire procedure was repeated $B = 5000$ times to include multiple sets of possible pools and to average over the effects of simulation.

For each state in Region X, Table 2.3 and 2.4 display the number of tests expended for female subjects during 2011 (averaged over the 5000 implementations) and, for higher-stage algorithms, the percent reduction in the average number of tests when compared to $S = 2$, for swab and urine specimen, respectively. Boxplots of the 5000 simulated values of $T^{(S)}$, shown cross-classified by specimen type (swab/urine) and state (AK, ID, OR, WA), are given in Appendix A.6. Our results suggest that using higher-stage hierarchical algorithms in all four states would be highly beneficial. For example, for females tested using swabs in Alaska, a three-stage algorithm (with pool sizes $n_1 = 9$, $n_2 = 3$, and $n_3 = 1$) confers an 11.0% reduction in the average number of tests when compared to the best two-stage algorithm from Tebbs et al. (2013). This same reduction for swabs is 10.8%, 11.8%, and 12.4% for Idaho, Oregon, and Washington, respectively. Note that higher-stage gains are smaller when testing urine specimens because the 2011 marginal infection rates are slightly larger (see Appendix...
A.6); however, the corresponding three-stage gains still do range from 5.9-10.5%. There are even a few instances in Table 2.3 where an optimal four-stage algorithm is the most efficient (i.e., swab testing in Oregon and Washington). However, four-stage gains for these data are small when compared to the best three-stage algorithms.

Overall, our analysis demonstrates that moving from two-stage to three-stage hierarchical algorithms would be preferred for Region X and in other regions where the marginal infection rates of chlamydia and gonorrhea are similar. Among the 103,690 Region X females tested in 2011, implementing the optimal two-stage algorithm from Tebbs et al. (2013) requires 53,231 tests on average, calculated by summing across the states and specimen types in Table 2.3 and Table 2.4. Optimal three-stage hierarchical algorithms require 47,412 tests on average, an overall 11% reduction and a savings of over 5,800 tests. Finally, we use Appendix A.6 to display the classification accuracy results from our investigation. There is a loss in pooling sensitivity for both infections as the number of stages increases, which is expected for any hierarchical procedure; however, this loss is often minor for gonorrhea. On the other hand, higher-stage algorithms provide larger positive predictive values for both infections.

2.6 Discussion

We have introduced $S$-stage hierarchical group testing algorithms for multiple infections, simultaneously generalizing Tebbs et al. (2013) and the extensive literature on hierarchical algorithms for single infections. Our operating characteristic derivations exploit a novel conceptualization of the decoding process by viewing testing results as error-laden realizations of a Markov chain, and our analysis of the IPP data from Region X illustrates the benefit of using higher-stage algorithms for chlamydia and gonorrhea detection. To help practitioners implement the methodology in this article, we have made R programs available at the Biometrics website on Wiley Online Library and at www.chrisbilder.com/grouptesting. Our programs will determine optimal
Table 2.4: Region X 2011 chlamydia and gonorrhea data for urine specimen. Average number of tests (sample standard deviation, SD) from $B = 5000$ sets of pools for 2-, 3-, and 4-stage hierarchical algorithms. The optimal configuration is determined by minimizing $n_1^{-1}E(T^{(S)})$ using the 2010 prevalences; see Appendix A.6. The percent reduction in the average number of tests is also shown. The maximum allowable master pool size is 20.

<table>
<thead>
<tr>
<th>State</th>
<th># Stages</th>
<th>Configuration</th>
<th># Tests (SD)</th>
<th>% Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>$S = 2$</td>
<td>4 : 1</td>
<td>2615.6 (31.5)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>$S = 3$</td>
<td>9 : 3 : 1</td>
<td>2460.4 (42.1)</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td>$S = 4$</td>
<td>18 : 6 : 3 : 1</td>
<td>2512.0 (54.1)</td>
<td>4.0</td>
</tr>
<tr>
<td>Idaho</td>
<td>$S = 2$</td>
<td>5 : 1</td>
<td>2253.4 (34.8)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>$S = 3$</td>
<td>9 : 3 : 1</td>
<td>2047.1 (37.4)</td>
<td>9.2</td>
</tr>
<tr>
<td></td>
<td>$S = 4$</td>
<td>18 : 6 : 3 : 1</td>
<td>2082.2 (48.5)</td>
<td>7.6</td>
</tr>
<tr>
<td>Oregon</td>
<td>$S = 2$</td>
<td>4 : 1</td>
<td>4459.0 (39.2)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>$S = 3$</td>
<td>9 : 3 : 1</td>
<td>4073.1 (52.7)</td>
<td>8.7</td>
</tr>
<tr>
<td></td>
<td>$S = 4$</td>
<td>18 : 6 : 3 : 1</td>
<td>4134.2 (68.3)</td>
<td>7.3</td>
</tr>
<tr>
<td>Washington</td>
<td>$S = 2$</td>
<td>5 : 1</td>
<td>8324.5 (66.2)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>$S = 3$</td>
<td>9 : 3 : 1</td>
<td>7454.5 (70.2)</td>
<td>10.5</td>
</tr>
<tr>
<td></td>
<td>$S = 4$</td>
<td>18 : 6 : 3 : 1</td>
<td>7521.1 (90.1)</td>
<td>9.7</td>
</tr>
</tbody>
</table>

The assumptions we have made in this article regarding the testing outcomes do not affect our Markov chain calculations because these calculations refer to the underlying true status process. Therefore, relaxing any of these assumptions should be possible by modifying the misclassification operators $P^{(s)}$, $P^{(s)}_{++}$ and $P^{(s)}_{+-}$ (Section 2.3), and those in Appendix A.3 and A.4. For example, one assumption we made was that testing responses are conditionally independent given the true statuses of all pools tested. This is certainly reasonable when misclassification is driven primarily by factors related to test implementation; however, it may not be reasonable otherwise.

We also implicitly assumed that $S^{(s)}_{e_{ij}}$ and $S^{(s)}_{p_{ij}}$ for one infection in stage $s$ do not depend on the true status of the other infection, an assumption that requires the multiplex assay used to possess adequate discriminating power. Future research in group testing
could investigate ways to avoid making either or both assumptions. McMahan et al. (2013) provide one way to relax the conditional independence assumption when additional biomarker information is available for each group testing response. Albert and Dodd (2004) provide an excellent summary of this issue when individual testing is used.

Finally, the merger of group testing for multiple infections and Markov processes brings with it exciting opportunities to investigate other case identification algorithms. For example, it should be possible to extend the $S$-stage array procedures in Berger et al. (2000) and Kim and Hudgens (2009) to allow for multiple infections using the framework outlined in this article. This extension would be more difficult because individuals are placed in overlapping pools; however, the underlying Markov chain structure for the true status decoding process still remains. We also believe that multiple-disease algorithms could be developed to incorporate risk factor information (e.g., age, race, number of sexual partners, etc.) on each individual. Bilder and Tebbs (2012) provide a review of recently proposed “informative” algorithms involving single infections. The approach outlined in Section 2.3 of this article could serve as a starting point towards generalizing their work.
Chapter 3

Array Testing with Multiplex Assays

Summary: In the single disease case, when compared to hierarchical group testing algorithms, such as Dorfman testing, array-based methods can not only lower testing costs but also overcome the higher false negative predictive value issue caused by the decrease in sensitivity. Motivated by the use of array-based algorithms for single infections, and the recent implement of multiplex assays with group testing to diagnose multiple infections simultaneously, we propose two new array-based group testing algorithms for multiplex assays, and we further compare them with previously-studied hierarchical algorithms for multiple infections. The operating characteristics of the proposed testing procedures, including the expected number of tests and classification accuracy probabilities, are derived. We illustrate the performance of array algorithms using simulation and by applying them to chlamydia and gonorrhea data collected in four states as part of the Infertility Prevention Project.
3.1 Introduction

When screening a population for low prevalence diseases, testing specimens in pools can be far more cost efficient than testing specimens individually. Individuals in pools that test negatively can be classified as negative, and individuals in pools that test positively can be retested to determine which ones are positive. Testing pooled specimens, which is known as group testing (or pooled testing), has a long history dating back to Dorfman (1943), who used group testing to screen United States military recruits for syphilis during World War II. Today, group testing is routinely used to screen blood donations for HIV, HBV, and HCV in the United States and in other developed nations (Mine et al., 2003; Seed et al., 2005; Schmidt et al., 2010; O’Brien et al., 2012; Stramer et al., 2013). It also arises in screening applications for other infections, including chlamydia and gonorrhea (Tebbs et al., 2013), West Nile virus (Busch et al., 2005), and influenza (Edouard et al., 2015).

Statistical research in group testing splits into two main areas: estimation and case identification. This article focuses on the latter. Case identification algorithms in group testing can generally be grouped into one of two categories: hierarchical and non-hierarchical. A hierarchical algorithm uses master pools that are non-overlapping, and positive pools are resolved in stages by splitting each one into smaller non-overlapping subpools. Dorfman’s original proposal was to accomplish this in two stages; i.e., master pools are tested in the first stage and individuals (from positive pools) are tested in the second. When the disease prevalence is low, increasing the number of stages can further reduce the number of tests needed. For example, Pilcher et al. (2005) use a three-stage algorithm for HIV testing in North Carolina with a master pool of size 90, nine second-stage subpools of size 10, and individual testing in the third stage. Although higher-stage hierarchical algorithms can be more efficient than two-stage algorithms, they are more complex and can also increase certain probabilities of misclassification (Kim et al., 2007).
Array testing, also known as matrix pooling, is the most common type of non-hierarchical algorithm. In array testing procedures (in two dimensions), individual specimens are initially assigned to an array consisting of rows and columns. Row and column master pools are tested in the first stage, and individuals not classified as negative after the first stage are retested in the second. Phatarfod and Sudbury (1994) introduced array testing for infectious disease screening purposes in the absence of testing error. Kim et al. (2007) and Westreich et al. (2008) offered thorough comparisons of array testing and hierarchical algorithms while allowing for imperfect assays. In other work, Hudgens and Kim (2011) determined optimal configurations for square arrays, McMahan et al. (2012b) extended array testing to heterogeneous populations, and Lendle et al. (2012) generalized two-dimensional array testing to account for correlated responses. Kim and Hudgens (2009) examined array testing in higher dimensions where, geometrically, one can envision that rows and columns are tested across multiple planes or hyperplanes. Martin et al. (2013) implemented a three-dimensional version of this algorithm for HIV testing in New Jersey.

In this article, we extend the use of array testing to test for multiple diseases simultaneously. Our work is motivated by the recent development of “multiplex assays,” which can detect multiple infections in a single application. As noted earlier, blood banks in Japan, Australia, Germany, Canada, and the United States already apply group testing with multiplex assays for viral infections, and several statewide screening programs in the United States use group testing with multiplex assays for chlamydia and gonorrhea (Lewis et al., 2012). However, despite their increased use, statistical research in case identification with multiplex assays is limited. Tebbs et al. (2013) derived the operating characteristics of two-stage hierarchical algorithms for two diseases, motivated by current screening practices in Iowa as part of the Infertility Prevention Project (IPP). For the remainder of this article, we refer to this algorithm as Iowa IPP algorithm.
Building on the hierarchical multiplex algorithms proposed in Tebbs et al. (2013), we introduce new array testing case identification procedures for multiple diseases. In Section 3.2, we propose a two-stage array-based group testing algorithm to account for two infections. In Section 3.3, we state the assumptions made throughout this article and define the operating characteristics used to evaluate the performance of the proposed algorithms. In Section 3.4, we provide numerical evidence to demonstrate the effectiveness of our new procedure, and introduce a modified version of the proposed two-stage algorithm when diseases are extremely rare. In Section 3.5, we illustrate the benefits of the proposed algorithms with the individual testing data collected by four states (Region X of the United States Department of Health and Human Services) as part of the IPP. Finally, we conclude the paper with a brief discussion in Section 3.6. All derivation details and additional numerical results are included in Appendix B.

3.2 Array-Based Group Testing Algorithm

In this section, we introduce an array-based group testing algorithm with two infections, denoted throughout by AT. Similar to the Iowa IPP algorithm, AT consists of two stages of testing, where we assume a single discriminating assay is used throughout. It begins by randomly assigning individuals to the cells of a square array of size $n \times n$, where $n$ is a positive integer and commonly larger than two. We denote the individual assigned to the $(i,j)$ cell by $I_{ij}$ for $i,j = 1, \ldots, n$. At the first stage, AT proceeds by testing $2n$ pools, where each pool is formed by mixing $n$ individual specimens in the same row or the same column. Let $R_i$ and $C_j$ denote the vectors of the testing responses of the $i$th row pool and the $j$th column pool, respectively, i.e., $R_i = (R_{i1}, R_{i2})'$ and $C_j = (C_{j1}, C_{j2})'$, where for $k = 1, 2$, $R_{ik} = 1$ ($C_{jk} = 1$) if the $i$th row pool ($j$th column pool) tests positively for the $k$th infection, and $R_{ik} = 0$ ($C_{jk} = 0$) otherwise. See Figure 3.1 for an example that illustrates this notation when
At the second stage, AT partitions all individuals in the array into one of two classes: those who need to be further retested individually (denoted by $\mathcal{M}_+$) and those who do not (denoted by $\mathcal{M}_-$). Determining which class an individual belongs to depends on the testing responses of all row and column pools for each infection. Individual $I_{ij}$ is assigned to $\mathcal{M}_+$ if either

(i) the $i$th row and the $j$th column pool test positively for a same infection, or

(ii) the $i$th row (the $j$th column) pool tests positively for an infection while all the column (row) pools test negatively for the same infection.
When the assay has no testing errors, individuals who are truly positive only exist in the cells described as in scenario (i). When the assay is not perfect, scenario (ii) may happen and deliver contradictory information. For example, if the \(i\)th row tests positively for the first infection, it suggests that there is at least one individual in this row who is truly positive for this infection; however, the negative test results for all columns indicates that all the individuals in this row are truly negative for this infection. To resolve this ambiguity, we choose to retest all the individuals in this row separately. This is similar to the array procedure designed for a single infection in Kim et al. (2007). Mathematically, we can express

\[
M_+ = \{ I_{ij} : T_{ij1}^{(AT)} + T_{ij2}^{(AT)} \geq 1 \},
\]

and \(M_- = M^c_+\), where, for \(k = 1, 2,\)

\[
T_{ijk}^{(AT)} = \begin{cases} 
1, & \text{if } R_{ik} = 1, C_{jk} = 1; \\
1, & \text{if } R_{ik} = 1, \sum_{j=1}^{n} C_{jk} = 0; \\
1, & \text{if } \sum_{i=1}^{n} R_{ik} = 0, C_{jk} = 1; \\
0, & \text{otherwise.}
\end{cases}
\]

Once the classes \(M_+\) and \(M_-\) are identified, AT classifies individuals in \(M_+\) by testing their specimens individually. Individuals in \(M_-\) are classified as negative for both infections without further testing.

### 3.3 Assumptions and Operating Characteristics

**Assumptions**

We now list several assumptions we make for the remainder of this article. These assumptions were also used in Tebbs et al. (2013). Discussions regarding our assumptions are given in this section and in Section 3.6.

**Assumption 1:** Let \(\widetilde{Y}_{ij} = (\widetilde{Y}_{ij1}, \widetilde{Y}_{ij2})'\) denote the true infection status of \(I_{ij}\), that is, \(\widetilde{Y}_{ij1} = 1\) if \(I_{ij}\) is truly positive for the \(k\)th infection, for \(i, j = 1, 2, \ldots, n\), and
$k = 1, 2; $ and 0 otherwise. We assume that the $Y_{ij}$’s are identically and independent
distributed multinomial random vectors with probability mass function $\Pr(Y_{ij1} =$ $\bar{y}_1, Y_{ij2} = \bar{y}_2) = p_{00}^{(1-\bar{y}_1)(1-\bar{y}_2)} p_{10}^{\bar{y}_1(1-\bar{y}_2)} p_{01}^{(1-\bar{y}_1)\bar{y}_2} p_{11}^{\bar{y}_1\bar{y}_2}$, for $\bar{y}_1, \bar{y}_2 \in \{0, 1\}$, where $p_{00} + p_{10} +$ $p_{01} + p_{11} = 1$.

Note that the cell probabilities $p_{00}, p_{10}, p_{01}$ and $p_{11}$ in Assumption 1 characterize
the population level prevalence; e.g., $p_{10}$ is the probability that an individual is truly
positive for the first infection and negative for the second infection, etc. Also note
that due to classification errors, the $Y_{ij}$’s cannot be observed directly nor deduced
from $Y_{ij}$’s; therefore, they are regarded as latent.

**Assumption 2**: The probability a pool tests positively for the $k$th infection condi-
tional on that it contains at least one individual who is truly positive for the $k$th
infection equals a known constant $S_{e_k}$, where $k = 1, 2$.

We refer to $S_{e_k}$ as the assay sensitivity for the $k$th infection. This assumption
implies that the assay sensitivity does not depend on the pool size nor the number
of truly positive specimens therein.

**Assumption 3**: The probability a pool tests negatively for the $k$th infection condi-
tional on that all specimens therein are truly negative for the $k$th infection
equals a known constant $S_{p_k}$, where $k = 1, 2$.

We refer to $S_{p_k}$ as the assay specificity for the $k$th infection. This assumption
implies that the assay specificity does not depend on the pool size.

**Assumption 4**: Conditional on the true statuses of any pooled (or individual)
specimen being tested, the test responses for both infections are mutually indepen-
dent.

For instance, we denote the vector of true infection statuses of the $i$th row pool
and the $j$th column pool by $\tilde{R}_i = (\tilde{R}_{i1}, \tilde{R}_{i2})'$ and $\tilde{C}_j = (\tilde{C}_{j1}, \tilde{C}_{j2})'$, respectively, where
for $k = 1, 2$, $\tilde{R}_{ik} = I(\sum_{j=1}^n Y_{ijk} > 0)$ and $\tilde{C}_{jk} = I(\sum_{i=1}^n Y_{ijk} > 0)$ with $I(\cdot)$ being the
indicator function. That is, $\tilde{R}_{ik} = 1$ ($\tilde{C}_{jk} = 1$) if the $i$th row ($j$th column) contains at least one individual who is truly positive for the $k$th infection and $\tilde{R}_{ik} = 0$ ($\tilde{C}_{jk} = 0$) otherwise. Then, under Assumption 4, $\Pr\{\mathbf{R}_1 = (x_1, x_2), \mathbf{C}_1 = (x_3, x_4) \mid \tilde{R}_1 = (y_1, y_2), \tilde{C}_1 = (y_3, y_4)\} = \Pr(R_{11} = x_1 \mid \tilde{R}_{11} = y_1) \Pr(R_{12} = x_2 \mid \tilde{R}_{12} = y_2) \Pr(C_{11} = x_3 \mid \tilde{C}_{11} = y_3) \Pr(C_{12} = x_4 \mid \tilde{C}_{12} = y_4)$ for any $x_1, \ldots, x_4, y_1, \ldots, y_4 \in \{0, 1\}$.

**Operating Characteristics**

As in Kim et al. (2007) and Tebbs et al. (2013), we evaluate an arbitrary group testing algorithm $\mathcal{A}$ by the following operating characteristics (in italics). *Efficiency*, denoted by $\text{EFF}(\mathcal{A})$, is defined as the expected number of tests per specimen for algorithm $\mathcal{A}$ to classify all individuals as either positive or negative for both infections. A small value of $\text{EFF}(\mathcal{A})$ indicates a large saving that algorithm $\mathcal{A}$ can achieve on average. Because we assume the assay is not perfect, which might potentially lead to the misclassification of individual specimens, it is also important to evaluate the classification accuracy of our algorithms. We define *pooling sensitivity (specificity)* for the $k$th infection, denoted by $\text{PS}_{e:k}$ ($\text{PS}_{p:k}$), to be the probability an individual is classified as positive (negative) for this infection by the algorithm given that it is truly positive (negative). A high value of $\text{PS}_{e:k}$ ($\text{PS}_{p:k}$) indicates that the algorithm produces a low expected false negative (positive) classification rate for the $k$th infection. Further, *pooling positive (negative) predictive value* for the $k$th infection, denoted by $\text{PPV}_k$ ($\text{NPV}_k$), is defined to be the probability an individual is truly positive (negative) for this infection given that it has been categorized as positive (negative) by the algorithm. Through the use of Bayes’ Rule, we can relate $\text{PPV}_k$ and $\text{NPV}_k$ to $\text{PS}_{e:k}$ and $\text{PS}_{p:k}$ for any algorithm; i.e., for $k = 1, 2$,

$$\text{PPV}_k = \frac{\pi_k \text{PS}_{e:k}}{\pi_k \text{PS}_{e:k} + (1 - \pi_k)(1 - \text{PS}_{p:k})} \quad \text{and} \quad \text{NPV}_k = \frac{(1 - \pi_k) \text{PS}_{p:k}}{(1 - \pi_k) \text{PS}_{p:k} + \pi_k (1 - \text{PS}_{e:k})},$$

where $\pi_1 = p_{10} + p_{11}$ and $\pi_2 = p_{01} + p_{11}$ are the population prevalence of the first and second infection, respectively.
Closed-form expressions for the operating characteristics of AT are obtained under the assumptions presented in Section 3.3. One can envision that these expressions depend on the assay sensitivity and specificity for each infection ($S_{e,k}$’s and $S_{p,k}$’s), the array size $n$, and the cell probabilities $p_{00}$, $p_{10}$ and $p_{01}$. Unfortunately, the formulas are not very friendly and thus are given in Appendix B.1.

3.4 Numerical Studies

In this section, we compare algorithm AT with the Iowa IPP algorithm. The operating characteristics of the IPP algorithm have been derived in Tebbs et al. (2013). For each algorithm, we first identify the “optimal” configuration for given values of the $S_{e,k}$’s, $S_{p,k}$’s, and the individual probabilities $p_{00}$, $p_{10}$ and $p_{01}$. To this end, we rewrite AT and IPP as functions of pool sizes in each stage; i.e., as $AT(n : 1)$ and $IPP(n : 1)$, respectively, where $n$ for AT is the number of rows (columns) of the square array and for IPP is the size of the master pool. The optimal configuration of each algorithm is defined as the value of $n$ that minimizes the efficiency; i.e., for each $\mathcal{A} \in \{AT(n : 1), IPP(n : 1)\}$, the best configuration is chosen to be $n^* = \arg\min_{n>1} \text{Eff}(\mathcal{A})$. Due to potential dilution (Quinn et al., 2000) in pools of large sizes, we set the maximum allowable size of the master pool in algorithm IPP to be no larger than 100; similar restrictions have also been used in Kim et al. (2007), Kim and Hudgens (2009), and Tebbs et al. (2013). For AT, we restrict the array size to be no larger than $20 \times 20$. This size is chosen because a size too small might not be able to fully exploit its potential to save costs, and a size too large does not only have the risk of diluting the specimens, but also create a longer waiting time for specimens to be processed. With these restrictions, it enables us to use direct search to find the optimal configuration $n^*$ for each algorithm among all possible configurations.

In Table 3.1, we consider six different combinations of the cell probabilities $p_{00}$, $p_{10}$, $p_{01}$, for $S_{e,k} = 0.95$ and $S_{p,k} = 0.99$, $k = 1, 2$. Cases (iii) and (iv) were chosen to
Table 3.1: Unbalanced prevalences. The maximum allowable pool size for IPP is 100 and for AT is $20 \times 20$. For each case, the optimal configuration $n^*$, the corresponding efficiency, $\text{PS}_e$ and PPV are calculated with the formulas provided in Tebbs et al. (2013) and Appendix B.1, for IPP and AT, respectively. $\text{PS}_p$ and NPV calculations are in Appendix B.3.

<table>
<thead>
<tr>
<th>Case</th>
<th>Cell probabilities</th>
<th>$\mathcal{A}$</th>
<th>$n^*$</th>
<th>Eff($\mathcal{A}$)</th>
<th>$\text{PS}_{e,1}$</th>
<th>$\text{PS}_{e,2}$</th>
<th>PPV$_1$</th>
<th>PPV$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>$p_{00} = 0.870$, $p_{10} = 0.123$</td>
<td>IPP</td>
<td>4</td>
<td>0.668</td>
<td>0.905</td>
<td>0.931</td>
<td>0.975</td>
<td>0.613</td>
</tr>
<tr>
<td></td>
<td>$p_{01} = 0.004$, $p_{11} = 0.003$</td>
<td>AT</td>
<td>6</td>
<td>0.659</td>
<td>0.863</td>
<td>0.939</td>
<td>0.981</td>
<td>0.674</td>
</tr>
<tr>
<td>(ii)</td>
<td>$p_{00} = 0.890$, $p_{10} = 0.104$</td>
<td>IPP</td>
<td>4</td>
<td>0.617</td>
<td>0.905</td>
<td>0.932</td>
<td>0.973</td>
<td>0.607</td>
</tr>
<tr>
<td></td>
<td>$p_{01} = 0.003$, $p_{11} = 0.003$</td>
<td>AT</td>
<td>7</td>
<td>0.598</td>
<td>0.863</td>
<td>0.939</td>
<td>0.977</td>
<td>0.647</td>
</tr>
<tr>
<td>(iii)</td>
<td>$p_{00} = 0.918$, $p_{10} = 0.078$</td>
<td>IPP</td>
<td>4</td>
<td>0.540</td>
<td>0.905</td>
<td>0.930</td>
<td>0.971</td>
<td>0.565</td>
</tr>
<tr>
<td></td>
<td>$p_{01} = 0.002$, $p_{11} = 0.002$</td>
<td>AT</td>
<td>8</td>
<td>0.505</td>
<td>0.862</td>
<td>0.939</td>
<td>0.974</td>
<td>0.600</td>
</tr>
<tr>
<td>(iv)</td>
<td>$p_{00} = 0.930$, $p_{10} = 0.066$</td>
<td>IPP</td>
<td>5</td>
<td>0.503</td>
<td>0.904</td>
<td>0.922</td>
<td>0.962</td>
<td>0.552</td>
</tr>
<tr>
<td></td>
<td>$p_{01} = 0.003$, $p_{11} = 0.001$</td>
<td>AT</td>
<td>9</td>
<td>0.456</td>
<td>0.861</td>
<td>0.931</td>
<td>0.971</td>
<td>0.618</td>
</tr>
<tr>
<td>(v)</td>
<td>$p_{00} = 0.950$, $p_{10} = 0.047$</td>
<td>IPP</td>
<td>5</td>
<td>0.431</td>
<td>0.904</td>
<td>0.923</td>
<td>0.959</td>
<td>0.549</td>
</tr>
<tr>
<td></td>
<td>$p_{01} = 0.002$, $p_{11} = 0.001$</td>
<td>AT</td>
<td>10</td>
<td>0.374</td>
<td>0.862</td>
<td>0.932</td>
<td>0.969</td>
<td>0.621</td>
</tr>
<tr>
<td>(vi)</td>
<td>$p_{00} = 0.970$, $p_{10} = 0.028$</td>
<td>IPP</td>
<td>7</td>
<td>0.342</td>
<td>0.905</td>
<td>0.929</td>
<td>0.939</td>
<td>0.485</td>
</tr>
<tr>
<td></td>
<td>$p_{01} = 0.001$, $p_{11} = 0.001$</td>
<td>AT</td>
<td>14</td>
<td>0.276</td>
<td>0.862</td>
<td>0.932</td>
<td>0.959</td>
<td>0.588</td>
</tr>
</tbody>
</table>

mimic the joint prevalence of chlamydia and gonorrhea data for female swab and urine specimens collected in four states of HHS Region X; see Section 3.5 for greater details. One may notice that in both cases, $p_{10}$ is much larger than $p_{01}$, which is consistent with the fact that chlamydia is much more prevalent than gonorrhea. We refer to this pattern as “unbalanced prevalences”. Under this scenario, Cases (i)–(ii) and Cases (v)–(vi) were chosen to illustrate the performance of these algorithms when two infections are moderately rare ($p_{00} \in \{0.870, 0.890\}$) and very rare ($p_{00} \in \{0.950, 0.970\}$), respectively. For each case, we first identify the optimal configuration $n^*$ of each algorithm and then calculate the operating characteristics described in Section 3.3 under the optimal configuration accordingly. Table 3.1 contains the efficiency (Eff($\mathcal{A}$)), the pooling sensitivities ($\text{PS}_{e,1}$ and $\text{PS}_{e,2}$) and the positive predictive values (PPV$_1$ and PPV$_2$). The differences in pooling specificity and negative predictive values between the two algorithms are negligible and thus are shown in Appendix B.3.

Our calculation of the efficiency in Table 3.1 shows that AT reduces the expected number of tests per individual in all considered scenarios. For example, the third case (used to emulate chlamydia and gonorrhea prevalence in Region X) shows that
the optimal AT algorithm uses $n^* = 8$ and confers a 6.5% reduction when compared to the optimal IPP algorithm; when $p_{00}$ increases to 0.970 (case (vi) in Table 3.1), AT uses $n^* = 14$ and the reduction increases to 19.4%. Besides cost savings, we have also observed that AT is very competitive to IPP in terms of classification accuracy. For example, values of PPV$_1$ and PPV$_2$ under the optimal AT algorithm are higher than those under the optimally sized IPP algorithm.

For pooling sensitivity, one commonly seen behavior in group testing for a single infection is that, cost savings usually lead to a loss in pooling sensitivity (Kim et al., 2007). This phenomenon still exists in our situation. One can see that PS$_{e1}$ of AT are less than those for the IPP. However, it is surprising to observe that AT produces a higher PS$_{e2}$ in all the considered examples. Note that this improvement is achieved under a cost reduction. This is not a coincidence. Because the assay being used in AT can diagnose both infections simultaneously in a single run, the individuals who should be retested for the infection with higher prevalence are also tested for the other infection as a byproduct. AT creates a larger chance for this to happen than the IPP algorithm; in other words, more individuals who are positive for the infection with lower prevalence can get retested with AT, and thus lead to a higher PS$_{e2}$. We view this as a unique feature of AT with multiplex assays that could be potentially helpful for current screening practices.

Table 3.2 shows the results for cases with balanced marginal prevalences. Cell probabilities $p_{00}$ and $p_{11}$ are exactly the same as those considered in Table 3.1, while $p_{10}$ and $p_{01}$ were chosen to be (nearly) the same. One can see that, when the marginal prevalences are close to each other, the savings achieved by AT when compared to IPP is now even larger. For example, Case (iii) in Table 3.2 presents a 18.2% savings for AT while in Table 3.1 the savings for AT is 6.5%. However, the unusual increase in PS$_{e2}$ is no longer observed.

We now propose a modification of AT when infections are extremely rare. The
Table 3.2: Balanced prevalences. $p_{00}$ and $p_{10}$ are the same as in the cases identified in Table 3.1, but with (nearly) same $p_{10}$ and $p_{01}$. The maximum allowable pool size for IPP is 100 and for AT is $20 \times 20$. For each case, the optimal configuration $n^*$, the corresponding efficiency, $PS_e$, and PPV are calculated with the formulas provided in Tebbs et al. (2013) and Appendix B.1, for IPP and AT, respectively. $PS_p$ and NPV calculations are in Appendix B.3.

<table>
<thead>
<tr>
<th>Case</th>
<th>Cell probabilities</th>
<th>$A$</th>
<th>$n^*$</th>
<th>$\text{Eff}(A)$</th>
<th>$PS_{e1}$</th>
<th>$PS_{e2}$</th>
<th>$PPV_1$</th>
<th>$PPV_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>$p_{00} = 0.870, p_{10} = 0.064$</td>
<td>IPP</td>
<td>4</td>
<td>0.670</td>
<td>0.913</td>
<td>0.913</td>
<td>0.945</td>
<td>0.944</td>
</tr>
<tr>
<td></td>
<td>$p_{01} = 0.063, p_{11} = 0.003$</td>
<td>AT</td>
<td>7</td>
<td>0.588</td>
<td>0.876</td>
<td>0.876</td>
<td>0.961</td>
<td>0.960</td>
</tr>
<tr>
<td>(ii)</td>
<td>$p_{00} = 0.890, p_{10} = 0.054$</td>
<td>IPP</td>
<td>4</td>
<td>0.619</td>
<td>0.912</td>
<td>0.912</td>
<td>0.943</td>
<td>0.942</td>
</tr>
<tr>
<td></td>
<td>$p_{01} = 0.053, p_{11} = 0.003$</td>
<td>AT</td>
<td>8</td>
<td>0.532</td>
<td>0.874</td>
<td>0.875</td>
<td>0.956</td>
<td>0.955</td>
</tr>
<tr>
<td>(iii)</td>
<td>$p_{00} = 0.918, p_{10} = 0.040$</td>
<td>IPP</td>
<td>4</td>
<td>0.541</td>
<td>0.910</td>
<td>0.910</td>
<td>0.938</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td>$p_{01} = 0.040, p_{11} = 0.002$</td>
<td>AT</td>
<td>9</td>
<td>0.443</td>
<td>0.873</td>
<td>0.873</td>
<td>0.953</td>
<td>0.953</td>
</tr>
<tr>
<td>(iv)</td>
<td>$p_{00} = 0.930, p_{10} = 0.034$</td>
<td>IPP</td>
<td>5</td>
<td>0.505</td>
<td>0.911</td>
<td>0.911</td>
<td>0.924</td>
<td>0.924</td>
</tr>
<tr>
<td></td>
<td>$p_{01} = 0.034, p_{11} = 0.002$</td>
<td>AT</td>
<td>10</td>
<td>0.401</td>
<td>0.872</td>
<td>0.872</td>
<td>0.949</td>
<td>0.949</td>
</tr>
<tr>
<td>(v)</td>
<td>$p_{00} = 0.950, p_{10} = 0.030$</td>
<td>IPP</td>
<td>5</td>
<td>0.429</td>
<td>0.908</td>
<td>0.908</td>
<td>0.931</td>
<td>0.928</td>
</tr>
<tr>
<td></td>
<td>$p_{01} = 0.029, p_{11} = 0.001$</td>
<td>AT</td>
<td>13</td>
<td>0.301</td>
<td>0.868</td>
<td>0.868</td>
<td>0.958</td>
<td>0.956</td>
</tr>
<tr>
<td>(vi)</td>
<td>$p_{00} = 0.970, p_{10} = 0.015$</td>
<td>IPP</td>
<td>7</td>
<td>0.342</td>
<td>0.909</td>
<td>0.910</td>
<td>0.888</td>
<td>0.881</td>
</tr>
<tr>
<td></td>
<td>$p_{01} = 0.014, p_{11} = 0.001$</td>
<td>AT</td>
<td>16</td>
<td>0.234</td>
<td>0.868</td>
<td>0.869</td>
<td>0.936</td>
<td>0.932</td>
</tr>
</tbody>
</table>

Examples considered in Table 3.1 and Table 3.2 are moderately rare to very rare infections; in other words, the array is likely to contain a few positive individuals. When infections are even rarer, such as HIV, HBV, and HCV, it is possible that all the specimens in the array are negative for the infections. To avoid introducing potential unnecessary tests, we begin by adding a master pool testing at the beginning of AT. More specifically, instead of testing rows and columns at the first stage, the new algorithm proceeds to test the master pool which is constructed by mixing all specimens in the array. If the test results are negative for both infections, all individuals in the array are declared as negative; otherwise, the algorithm proceeds as in AT. We denote this three-stage algorithm by ATM. Mathematical expressions of ATM and all the derivations of its operating characteristics are presented in Appendix B.2.

Table 3.3 presents the operating characteristics of ATM and IPP under their optimal configurations. To emphasize the impact of the group size on the performance of ATM, we write ATM as ATM($n^2 : n : 1$), where $n^2$, $n$, and 1 are the group sizes at each stage. Since ATM and IPP both start with master pool testing, we restrict
Table 3.3: Extremely rare prevalences. The maximum allowable pool size for IPP is 100 and for ATM is 10 × 10. For each case, the optimal configuration \( n^* \), the corresponding efficiency, \( PS_e \) and PPV are calculated with the formulas provided in Tebbs et al. (2013) and Appendix B.2, for algorithm IPP and ATM, respectively. \( PS_p \) and NPV calculations are in Appendix B.3.

<table>
<thead>
<tr>
<th>Case</th>
<th>Cell probabilities</th>
<th>( A )</th>
<th>( n^* )</th>
<th>Eff (( A ))</th>
<th>( PS_{e1} )</th>
<th>( PS_{e2} )</th>
<th>PPV 1</th>
<th>PPV 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>( p_{00} = 0.9900, p_{10} = 0.0048 )</td>
<td>IPP</td>
<td>11</td>
<td>0.209</td>
<td>0.909</td>
<td>0.909</td>
<td>0.807</td>
<td>0.807</td>
</tr>
<tr>
<td></td>
<td>( p_{01} = 0.0048, p_{11} = 0.0004 )</td>
<td>ATM</td>
<td>5</td>
<td>0.149</td>
<td>0.899</td>
<td>0.899</td>
<td>0.972</td>
<td>0.972</td>
</tr>
<tr>
<td>(ii)</td>
<td>( p_{00} = 0.9930, p_{10} = 0.0033 )</td>
<td>IPP</td>
<td>13</td>
<td>0.178</td>
<td>0.908</td>
<td>0.908</td>
<td>0.775</td>
<td>0.770</td>
</tr>
<tr>
<td></td>
<td>( p_{01} = 0.0034, p_{11} = 0.0003 )</td>
<td>ATM</td>
<td>6</td>
<td>0.119</td>
<td>0.898</td>
<td>0.899</td>
<td>0.967</td>
<td>0.966</td>
</tr>
<tr>
<td>(iii)</td>
<td>( p_{00} = 0.9960, p_{10} = 0.0019 )</td>
<td>IPP</td>
<td>17</td>
<td>0.140</td>
<td>0.909</td>
<td>0.909</td>
<td>0.706</td>
<td>0.706</td>
</tr>
<tr>
<td></td>
<td>( p_{01} = 0.0019, p_{11} = 0.0002 )</td>
<td>ATM</td>
<td>7</td>
<td>0.084</td>
<td>0.899</td>
<td>0.899</td>
<td>0.959</td>
<td>0.959</td>
</tr>
<tr>
<td>(iv)</td>
<td>( p_{00} = 0.9990, p_{10} = 0.0004 )</td>
<td>IPP</td>
<td>33</td>
<td>0.081</td>
<td>0.911</td>
<td>0.913</td>
<td>0.524</td>
<td>0.478</td>
</tr>
<tr>
<td></td>
<td>( p_{01} = 0.0005, p_{11} = 0.0001 )</td>
<td>ATM</td>
<td>10</td>
<td>0.036</td>
<td>0.901</td>
<td>0.904</td>
<td>0.939</td>
<td>0.926</td>
</tr>
</tbody>
</table>

the maximum allowable array size for algorithm ATM to be 10 × 10 to provide a fair comparison. One notices that with the help of master pool testing, ATM can significantly reduce the testing expenditures when compared to the IPP algorithm for infections that are extremely rare and also increase the PPV’s. For example, in Case (iv) of Table 3.3, the optimal ATM algorithm can save roughly 55% the number of tests when compared to the optimal IPP algorithm, and the PPV’s for ATM are almost twice as large. The pooling sensitivities for ATM, on the other hand, are lower than those for IPP, but the difference is minimal.

To provide a comprehensive comparison between the array-based testing algorithms (AT and ATM) with the IPP procedure, we display in Figure 3.2 the best algorithm to use when the marginal prevalences \( \pi_1 \) and \( \pi_2 \) range from 0.001 to 0.2, \( S_{e:k} = 0.95 \), \( S_{p:k} = 0.99 \), for \( k = 1, 2 \), and the correlation between the true statuses of two infections \( \rho = \text{corr}(\tilde{Y}_{ij1}, \tilde{Y}_{ij2}) \) is fixed at 0, 0.1, 0.25 and 0.5. For each combination of \( \pi_1 \), \( \pi_2 \), and \( \rho \), the best algorithm was chosen to be the one which produces the smallest efficiency. Figure 3.2 shows that AT is preferable for a sizable subset of parameter space where the infection prevalences are moderate, and ATM is preferred when one or both prevalences are extremely low.
Figure 3.2: Optimal algorithm when $S_{e,k} = 0.95$ and $S_{p,k} = 0.99$. The maximum allowable pool size for algorithm IPP, AT, and ATM is $100$, $20 \times 20$, and $10 \times 10$, respectively. Values in the white region (in $\rho = 0.10, 0.25,$ and $0.50$ sub-figures) are not possible because correlation for binary responses are restricted. Note that “IND” denotes the individual testing.

3.5 Applications

Recall that the four states in HHS Region X, which include Alaska, Idaho, Oregon, and Washington, use individual testing to screen for chlamydia and gonorrhea as part of the IPP. Our data consists of the individual screening results collected by them during 2010 and 2011, roughly 100,000 results each year. In our analysis, we treat the individual classification results as the “true statuses” and then incorporate the characteristics of the Aptima Combo 2 Assay to simulate group testing process.
Using individual level data to emulate group testing data is a common approach in the group testing literature (McMahan et al., 2012a; McMahan et al., 2012b; Tebbs et al., 2013). Doing so also allows us to numerically evaluate the classification accuracy of our algorithms by comparing the emulated test responses with the “true statuses.” Because group testing in the IPP has been applied to female specimens, we only include these data in this study.

Table 3.4 provides a summary of the data for each year, including the sample size $N$ and the estimated cell probabilities $\hat{p}_{00}, \hat{p}_{10}, \hat{p}_{01},$ and $\hat{p}_{11},$ stratified by state and specimen type. This helps us acknowledge the population heterogeneity among the states, and the differences in characteristics of the Aptima Combo 2 Assay on swab and urine specimens. The assay characteristics for both specimens are provided in the product literature of Gen-Probe, Inc., available at www.fda.gov. The assay sensitivities for chlamydia and gonorrhea are 0.942 and 0.992, and the assay specificities are 0.976 and 0.987, respectively, for female swab specimens. For female urine specimens, the sensitivities are 0.947 and 0.913 and the specificities are 0.989 and 0.993, respectively.

To perform our analysis, we view the data collected in 2010 as training data, and use them to determine the optimal configuration, $n^\ast$, for each of $A \in \{\text{IPP, AT, ATM}\}$. Then, we implement three group testing algorithms with the 2011 individuals by randomly assigning them to the optimally-configured pools or arrays. Based on the assay characteristics, we simulate all testing responses that would occur under each algorithm and record the number of tests used and the final classification results for each individual (which are used to calculate the classification accuracy). This process is repeated $B = 1000$ times. The average results for the number of tests are summarized in Table 3.5. The averaged classification accuracies are given in Appendix B.3.

In Table 3.5, $\overline{T}$ denotes the average number of tests performed to diagnose all individuals in each stratum. One can easily see that the two array testing algorithms
Table 3.4: Region X IPP female data summary. The sample sizes \( N \) are given for each state/specimen stratum. The prevalence estimates from year 2010 are used to determine the optimal configuration of IPP, AT and ATM for 2011.

<table>
<thead>
<tr>
<th>State</th>
<th>Year</th>
<th>Swab Size</th>
<th>Prevalence</th>
<th>Urine Prevalence</th>
<th>Urine Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>2010</td>
<td>( \hat{p}_{00} = 0.941 )</td>
<td>( \hat{p}_{01} = 0.002 )</td>
<td>( \hat{p}_{00} = 0.905 )</td>
<td>( \hat{p}_{01} = 0.006 )</td>
</tr>
<tr>
<td>Alaska</td>
<td>2011</td>
<td>( \hat{p}_{00} = 0.929 )</td>
<td>( \hat{p}_{01} = 0.003 )</td>
<td>( \hat{p}_{00} = 0.905 )</td>
<td>( \hat{p}_{01} = 0.004 )</td>
</tr>
<tr>
<td>Idaho</td>
<td>2010</td>
<td>( \hat{p}_{00} = 0.932 )</td>
<td>( \hat{p}_{01} = 0.001 )</td>
<td>( \hat{p}_{00} = 0.930 )</td>
<td>( \hat{p}_{01} = 0.001 )</td>
</tr>
<tr>
<td>Idaho</td>
<td>2011</td>
<td>( \hat{p}_{00} = 0.926 )</td>
<td>( \hat{p}_{01} = 0.001 )</td>
<td>( \hat{p}_{00} = 0.918 )</td>
<td>( \hat{p}_{01} = 0.002 )</td>
</tr>
<tr>
<td>Oregon</td>
<td>2010</td>
<td>( \hat{p}_{00} = 0.938 )</td>
<td>( \hat{p}_{01} = 0.001 )</td>
<td>( \hat{p}_{00} = 0.919 )</td>
<td>( \hat{p}_{01} = 0.003 )</td>
</tr>
<tr>
<td>Oregon</td>
<td>2011</td>
<td>( \hat{p}_{00} = 0.932 )</td>
<td>( \hat{p}_{01} = 0.002 )</td>
<td>( \hat{p}_{00} = 0.920 )</td>
<td>( \hat{p}_{01} = 0.002 )</td>
</tr>
<tr>
<td>Washington</td>
<td>2010</td>
<td>( \hat{p}_{00} = 0.942 )</td>
<td>( \hat{p}_{01} = 0.002 )</td>
<td>( \hat{p}_{00} = 0.938 )</td>
<td>( \hat{p}_{01} = 0.001 )</td>
</tr>
<tr>
<td>Washington</td>
<td>2011</td>
<td>( \hat{p}_{00} = 0.935 )</td>
<td>( \hat{p}_{01} = 0.003 )</td>
<td>( \hat{p}_{00} = 0.928 )</td>
<td>( \hat{p}_{01} = 0.002 )</td>
</tr>
</tbody>
</table>

The proposed algorithm can outperform the IPP algorithm currently used in the state of Iowa in terms of cost savings. For example, to screen the 21,134 female swab specimens in the state of Washington, one needs to carry out 10,497.6 tests with the optimal IPP algorithm. This number reduces to 9,355.5 when algorithm AT is employed, an approximate 11% savings. ATM does not perform as well as it
Table 3.5: Classification results for 2011. For each state/specimen stratum, the sample size \( N \) and the number of tests, which are averaged over \( B = 1000 \) replications, are included. The optimal configuration \( n^* \) is determined using the 2010 prevalences. The savings when compared to the optimal IPP algorithm is also shown. The averaged classification accuracy measure are given in Appendix B.3.

<table>
<thead>
<tr>
<th>State</th>
<th>Specimen</th>
<th>Algorithm</th>
<th>( T(n^*) )</th>
<th>% Saving</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>Swab ((N = 2910))</td>
<td>IPP</td>
<td>1510.5 (5)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>1373.7 (10)</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>1354.6 (10)</td>
<td>10.3</td>
</tr>
<tr>
<td></td>
<td>Urine ((N = 4558))</td>
<td>IPP</td>
<td>2614.4 (4)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>2491.8 (8)</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>2464.0 (8)</td>
<td>5.8</td>
</tr>
<tr>
<td>Idaho</td>
<td>Swab ((N = 7470))</td>
<td>IPP</td>
<td>3935.3 (5)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>3626.3 (9)</td>
<td>7.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>3549.9 (9)</td>
<td>9.8</td>
</tr>
<tr>
<td></td>
<td>Urine ((N = 4168))</td>
<td>IPP</td>
<td>2253.4 (5)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>2127.5 (9)</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>2086.1 (9)</td>
<td>7.4</td>
</tr>
<tr>
<td>Oregon</td>
<td>Swab ((N = 38654))</td>
<td>IPP</td>
<td>19632.5 (5)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>17723.4 (9)</td>
<td>9.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>17356.3 (9)</td>
<td>11.6</td>
</tr>
<tr>
<td></td>
<td>Urine ((N = 8381))</td>
<td>IPP</td>
<td>4459.3 (4)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>4165.6 (8)</td>
<td>6.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>4115.7 (8)</td>
<td>7.7</td>
</tr>
<tr>
<td>Washington</td>
<td>Swab ((N = 21134))</td>
<td>IPP</td>
<td>10497.6 (5)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>9355.5 (10)</td>
<td>10.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>9212.5 (10)</td>
<td>12.2</td>
</tr>
<tr>
<td></td>
<td>Urine ((N = 16415))</td>
<td>IPP</td>
<td>8320.6 (5)</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>7660.6 (9)</td>
<td>7.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>7531.9 (9)</td>
<td>9.5</td>
</tr>
</tbody>
</table>

does in the cases identified in Table 3.3, which is expected because chlamydia and gonorrhea are not extremely rare infections. However, it still provides additional 1.3% savings when compared to AT (which translates into hundreds of tests). The same pattern can be found in all states. Overall, using the optimal IPP algorithm, it takes 53223.6 tests to screen all 103,690 female specimens in four states; while using the optimal AT and ATM algorithms requires 48524.4 (8.8% saving) and 47671 tests (10.4% saving). Because programs like IPP are implemented nationwide, one would expect these savings to multiply quickly. In terms of classification accuracy (given in Table B.4), there is evidence that the pooling sensitivities for the array algorithms
are somewhat worse for chlamydia, but better for gonorrhea, as expected. On the other hand, the PPVs for array-based algorithms are higher than those of IPP, while the pooling specificity and NPV do not change significantly.

3.6 Discussion

We have proposed two array testing algorithms with two infections under the assumption that a single discriminating assay is used throughout. The operating characteristics of these algorithms have been derived, and their performances are illustrated through a comparison with the hierarchical alternative that is used in the Iowa IPP; see Tebbs et al. (2013). By applying our methods to chlamydia and gonorrhea data collected by the four states in federal HHS Region X, we show that the proposed algorithms are attractive pooling strategies for multiple infections because they can confer significant additional cost savings and an increase in classification accuracy. Furthermore, because multiplex assays are now also used with diseases with extreme low prevalence, such as HIV, HBV, and HCV screening practice carried out by the American Red Cross, one can expect the benefits of employing our three-stage version, ATM, to be significant.

This work could be extended in several ways. First, we only focus herein on two-dimensional square arrays and determine the optimal pool size from these. The assumption of equal number of rows and columns is made to avoid making our notation unnecessarily complex, but it can be easily relaxed to incorporate $m \times n$ arrays. The efficiency gain by relaxing the square array assumption, however, is not expected to be significant. A similar conclusion is found in Berger et al. (2000) for single infection. As for the dimension of the arrays, our two-dimensional conceptualization can be extended to higher-dimensional settings. For example, in a three-dimensional setting, individuals can be placed in a cube and decoded for multiple infections in planar slices, generalizing the work in Kim and Hudgens (2009) for one infection.
Deriving the operating characteristics, of course, would be much more complex, but it could be done in the same spirit as in our two-dimensional case.

We put a constraint on the maximum allowable pool (or array) size when searching for the optimal configuration. Although the constant sensitivity and specificity assumption are common in the group testing literature (and can be tested in practice by using positive and negative controls), it might be unrealistic in practice when there is evidence of assay contamination. For example, as pointed out in Kim et al. (2007), pool size and assay sensitivity are negatively associated when screening for HIV using nucleic acid amplification tests. To take potential dilution into consideration, one possible future extension of our work is to treat the sensitivity and specificity as functions that depend on pool size; see Hwang (1976) and Johnson et al. (1991).

Finally, a more challenging extension of our work would be to exploit population heterogeneity by acknowledging the effect of risk factors on individuals’ positivity. When individuals visit testing centers to submit specimens for examining, they are often asked to provide several covariates, such as age, sexual history, symptom status, etc. It is well documented that one can benefit greatly by incorporating this covariate information into group testing algorithms with one infection. For example, McMahan et al. (2012b) has shown that a significant reduction in the number of tests can be realized if individuals are assigned to the cells of a square array according to their ordered risk probabilities. When testing for multiple infections, although more complex, we believe that taking risk factor information into account could further enhance the cost savings similar to what McMahan et al. (2012b) observed for single infection.
Chapter 4

Semiparametric analysis of multiple-disease group testing data with application to the infertility prevention project

Summary: Group testing has been widely used as a cost-effective means for large-scale infectious disease screening. With the use of multiplex assays which can screen multiple infections at once, the research in group testing is moving toward developing new methodologies to incorporate multiple correlated disease data. Recent advances of group testing estimation literature are mainly treating the population of interest as homogeneous and estimating overall disease prevalence. In this work, we propose a semi-parametric regression framework by using multiple disease data to estimate individual-specific probability of each infection. The approach we took views the correlation among infections as nuisance parameters that can be modeled by nonparametric single index functions, and it also allows us to produce interpretable statistical inference for each infection separately. Further, we investigate the finite sample performance of our proposed methodology through simulation and by applying it to chlamydia and gonorrhea data obtained from Nebraska as part of the Infertility Prevention Project.
4.1 Notation and Assumptions

Recall that in the previous chapters, we introduced the hierarchical algorithm currently used by Iowa for simultaneous detection of chlamydia and gonorrhea as part of the Infertility Prevention Project. We refer to this algorithm as Iowa IPP algorithm. It proceeds in two steps. Firstly, individuals are randomly assigned into non-overlapping pools, each of which is of size \( n \) and then tested by a single discriminating assay, the Aptima Combo 2 Assay (AC2A). In the second step, if a pool tests positive for at least one infection, then each individual within the pool are retested by AC2A separately and the final diagnosis is drawn based on the individual test responses. On the other hand, if a pool tests negative for both infections, then all individuals within the pool are declared as negative.

Denote the \( j \)th individual in the \( i \)th pool as \( I_{ij} \), for \( i = 1, 2, \ldots, I \) and \( j = 1, 2, \ldots, n \). Let \( \tilde{Y}_{ij} = (\tilde{Y}_{ij1}, \tilde{Y}_{ij2})^T \) be a random vector, where \( \tilde{Y}_{ijk} = 1 \) if \( I_{ij} \) is truly positive for the \( k \)th infection; 0 otherwise. For each individual \( I_{ij} \), we have a \((p+1)\)-dimensional vector of covariates available, denoted by \( X_{ij} = (1, X_{ij1}, X_{ij2}, \ldots, X_{ijp})^T \). We assume that \( (\tilde{Y}_{ij}^T, X_{ij}^T)^T \) are independent and identically distributed across all pools and individuals, and \( X_{ij} \) is marginally related to \( \tilde{Y}_{ij} \) through \( \Pr\{\tilde{Y}_{ijk} = 1 | X_{ij} = x_{ij}\} = \eta_k(x_{ij}^T \beta_k) \), for \( k = 1, 2 \), where \( \eta_k(\cdot) \) is a known link function, and \( \beta_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp})^T \) is a \((p+1) \times 1\) vector of regression parameters that need to be estimated.

Denote the \( i \)th pool by \( G_i \), for \( i = 1, 2, \ldots, I \), and its test responses by a binary random vector \( Z_i = (Z_{i1}, Z_{i2})^T \), where \( Z_{ik} = 1 \) if pool \( G_i \) is tested positively for the \( k \)th infection, for \( k = 1, 2 \); and 0 otherwise. Note that in this article, we do not assume the test is perfect, hence the test responses we observe might differ with the true statuses, which we denote by binary random vector \( \tilde{Z}_i = (\tilde{Z}_{i1}, \tilde{Z}_{i2})^T \), where \( \tilde{Z}_{ik} = 1 \) if pool \( G_i \) is truly positive for the \( k \)th infection, for \( k = 1, 2 \); and 0 otherwise. The discrepancy between test responses and true statuses for the \( k \)th infection are described through
assay sensitivity and specificity, denoted by $S_{e.k}$ and $S_{p.k}$, respectively. Similar to Tebbs et al. (2013), we assume these quantities are known constants and do not depend on the pool sizes. In addition, we also assume the test responses are mutually independent conditional on the true statuses.

As noted earlier, if pool $G_i$ is tested positively for at least one infection, $I_{ij}$ will be retested for every $j = 1, 2, \ldots$. Let $Y_{ij} = (Y_{ij1}, Y_{ij2})^T$ denote the individual test responses produced by the retesting of $I_{ij}$, where $Y_{ijk} = 1$ indicated $I_{ij}$ is tested positively for $k$th infection at the second stage; 0 otherwise. For notation simplicity, we define another binary random variable $D_{ijk}$ such that

$$D_{ijk} = \begin{cases} 1, & \text{if } Y_{ijk} = 1; \\ 0, & \text{if } Z_{i1} + Z_{i2} = 0; \\ 0, & \text{if } Z_{i1} + Z_{i2} > 0 \text{ and } Y_{ijk} = 0. \end{cases}$$

That is, $D_{ijk}$ indicates the final diagnosis of the $k$th infection for $I_{ij}$.

4.2 Methodology

The overarching goal of this project is to estimate the individual risk for each infection, i.e., $\eta_1(x_{ij}^T \beta_1)$ and $\eta_2(x_{ij}^T \beta_2)$. To simplify our discussion, we assume $\eta_k(\cdot)$ is an inverse logit function, i.e.,

$$\eta_k(x_{ij}^T \beta_k) = \frac{\exp(x_{ij}^T \beta_k)}{1 + \exp(x_{ij}^T \beta_k)}.$$ 

One must note that in the Iowa IPP algorithm, the retesting of one infection might be triggered by another infection, therefore the correlation between these two infections cannot be neglected. Zhang, Bilder, and Tebbs (2013) proposed a GEE approach to estimate $\beta_1$, $\beta_2$ and the correlation simultaneously. Their work treats the correlation as a common parameter among individuals, and they only consider the master pool testing. In this section, we propose a semi-parametric regression framework which incorporates the additional individual level retesting information to estimate $\beta_1$ and
\( \beta_2 \). Our methods treat the correlation between two infections as nuisance parameters and allows them to differ among individuals. Therefore, we acknowledge the heterogeneous nature among individuals. Our approach puts a nonparametric form on the correlation by \( \eta(x_{ij}^T \beta_1, x_{ij}^T \beta_2) \), i.e.,

\[
\eta(x_{ij}^T \beta_1, x_{ij}^T \beta_2) \equiv \Pr\{\widetilde{Y}_{ij1} = 1, \widetilde{Y}_{ij2} = 1 | X_{ij} = x_{ij}\},
\]

where \( \eta(\cdot, \cdot) \) is a unknown 2-dimensional curve. For simplicity, we define

\[
p_{00(ij)} \equiv \Pr\{\widetilde{Y}_{ij} = (0, 0)^T | X_{ij} = x_{ij}\} = 1 - \eta_1(x_{ij}^T \beta_1) - \eta_2(x_{ij}^T \beta_2) + \eta(x_{ij}^T \beta_1, x_{ij}^T \beta_2),
\]

\[
p_{10(ij)} \equiv \Pr\{\widetilde{Y}_{ij} = (1, 0)^T | X_{ij} = x_{ij}\} = \eta_1(x_{ij}^T \beta_1) - \eta(x_{ij}^T \beta_1, x_{ij}^T \beta_2),
\]

and

\[
p_{01(ij)} \equiv \Pr\{\widetilde{Y}_{ij} = (0, 1)^T | X_{ij} = x_{ij}\} = \eta_2(x_{ij}^T \beta_2) - \eta(x_{ij}^T \beta_1, x_{ij}^T \beta_2).
\]

Further, we let \( Z, Y \), and \( X \) be vectors containing all \( Z_i, Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})^T \), and \( X_i = (X_{i1}, X_{i2}, \ldots, X_{in})^T \). Then the log-likelihood is

\[
l(\theta | Z = z, Y = y, X = x) = \sum_{i=1}^l \log \left( (\gamma_{i,00} S_{p,1} S_{p,2} + \gamma_{i,10} \overline{S}_{e,1} S_{p,2} + \gamma_{i,01} S_{p,1} \overline{S}_{e,2} + \gamma_{i,11} \overline{S}_{e,1} \overline{S}_{e,2}) I\{Z_{i1} + Z_{i2} = 0\} + (\zeta_{i,00}(y_i) S_{p,1}^{1-z_{i1}} \overline{S}_{p,2}^{1-z_{i2}} S_{p,2} + \zeta_{i,10}(y_i) S_{p,1}^{1-z_{i1}} \overline{S}_{e,1} \overline{S}_{e,2} + \zeta_{i,01}(y_i) S_{p,1} \overline{S}_{e,1}^{1-z_{i2}} \overline{S}_{e,2}^{1-z_{i2}} + \zeta_{i,11}(y_i) S_{p,1} \overline{S}_{e,1} \overline{S}_{e,2}^{1-z_{i2}}) I\{Z_{i1} + Z_{i2} > 0\} \right),
\]

where \( I\{\cdot\} \) is the indicator function, \( \theta = (\beta_1^T, \beta_2^T)^T, \gamma_{i,00}, \gamma_{i,10}, \gamma_{i,01}, \) and \( \gamma_{i,11} \) denotes the probability that \( G_i \) has true statuses “00”, “10”, “01”, and “11”, respectively. By Law of Total Probability, we obtain

\[
\gamma_{i,00} = \prod_{j=1}^n p_{00(ij)}, \gamma_{i,10} = \prod_{j=1}^n (p_{00(ij)} + p_{10(ij)}) - \gamma_{i,00}, \gamma_{i,01} = \prod_{j=1}^n (p_{00(ij)} + p_{01(ij)}) - \gamma_{i,00}, \text{ and } \gamma_{i,11} = 1 - \gamma_{i,00} - \gamma_{i,10} - \gamma_{i,01}.
\]
\(\zeta_{i, \tilde{z}_1 \tilde{z}_2}(y_i)\) in Equation (4.1) denotes \(\text{pr}(Y_i = y_i, \tilde{Z}_{i1} = \tilde{z}_1, \tilde{Z}_{i2} = \tilde{z}_2)\) for \(\tilde{z}_1, \tilde{z}_2 \in \{0, 1\}\), and are given by

\[
\zeta_{i,00}(y_i) = \prod_{j=1}^{n} \left\{ p_{00(ij)} S_{p_1}^{1-Y_{i1j}} S_{p_1}^{Y_{i1j}} S_{p_2}^{1-Y_{i2j}} S_{p_2}^{Y_{i2j}} \right\},
\]

\[
\zeta_{i,10}(y_i) = \prod_{j=1}^{n} \left\{ p_{00(ij)} S_{p_1}^{1-Y_{i1j}} S_{p_1}^{Y_{i1j}} S_{p_2}^{1-Y_{i2j}} S_{p_2}^{Y_{i2j}} + p_{10(ij)} S_{e_1}^{1-Y_{i1j}} S_{e_1}^{Y_{i1j}} S_{p_2}^{1-Y_{i2j}} S_{p_2}^{Y_{i2j}} \right\} - \zeta_{i,00}(y_i),
\]

\[
\zeta_{i,01}(y_i) = \prod_{j=1}^{n} \left\{ p_{00(ij)} S_{p_1}^{1-Y_{i1j}} S_{p_1}^{Y_{i1j}} S_{p_2}^{1-Y_{i2j}} S_{p_2}^{Y_{i2j}} + p_{01(ij)} S_{p_1}^{1-Y_{i1j}} S_{p_1}^{Y_{i1j}} S_{p_2}^{1-Y_{i2j}} S_{p_2}^{Y_{i2j}} \right\} - \zeta_{i,00}(y_i),
\]

\[
\zeta_{i,11}(y_i) = \prod_{j=1}^{n} \left\{ p_{00(ij)} S_{p_1}^{1-Y_{i1j}} S_{p_1}^{Y_{i1j}} S_{p_2}^{1-Y_{i2j}} S_{p_2}^{Y_{i2j}} + p_{10(ij)} S_{e_1}^{1-Y_{i1j}} S_{e_1}^{Y_{i1j}} S_{p_2}^{1-Y_{i2j}} S_{p_2}^{Y_{i2j}} + p_{01(ij)} S_{p_1}^{1-Y_{i1j}} S_{e_2}^{Y_{i1j}} S_{p_2}^{1-Y_{i2j}} S_{e_2}^{Y_{i2j}} + \eta(x_{ij}^T \beta_1, x_{ij}^T \beta_2) S_{e_1}^{1-Y_{i1j}} S_{e_1}^{Y_{i1j}} S_{e_2}^{1-Y_{i2j}} S_{e_2}^{Y_{i2j}} \right\} - \zeta_{i,00}(y_i) - \zeta_{i,10}(y_i) - \zeta_{i,01}(y_i),
\]

where \(S_{e_k} = 1 - S_{e_k}\) and \(S_{p_k} = 1 - S_{p_k}\). In the following sections, we explain the methods we used to obtain the parameter estimate \(\hat{\theta}\) and the variance-covariance matrix \(\hat{\Sigma}_\theta\).

**Estimate \(\beta_k\)**

There are three unknown components in the likelihood (4.1), \(\eta(\cdot, \cdot)\) and \(\beta_k\), for \(k = 1, 2\). The first step is to obtain a profiling estimator of the curve \(\eta(\cdot, \cdot)\). Once \(\hat{\eta}(\cdot, \cdot)\) is obtained, we can simply plug it in the log-likelihood to obtain the maximum likelihood estimator \(\hat{\beta}_k\). Our method to get \(\hat{\eta}(\cdot, \cdot)\) consists of three steps.

**Step 1.** We use all group testing data to estimate the population level prevalence; i.e., we treat all individuals as a sample from a homogeneous population and their
true statues are iid from a multinomial model: \( p_{y_1y_2} = \operatorname{pr}(\vec{Y}_{ij} = (y_1, y_2)^T) = p_{y_1y_2} \), for \( y_1, y_2 \in \{0, 1\} \). In other words, \( p_{y_1y_2} \) can also be viewed as

\[
p_{y_1y_2} = E_x \left[ \operatorname{pr}(\vec{Y}_{ij} = (y_1, y_2)^T) = p_{y_1y_2} | X = x \right],
\]

where \( X \) is a vector aggregating all covariates in the sample. Tebbs et al. (2013) proposed an EM algorithm augmented with Gibbs sampler to estimate these quantities. We adopt the same approach in this paper, but with a slight modification by replacing the Gibbs sampler part in the E-step with closed form expression to estimate the conditional expectations. We now outline the EM algorithm, with an emphasis on the E-step. Let \( \vec{Y} \) be the vector that aggregates all the individual true statuses and \( \theta = (p_{00}, p_{10}, p_{01})^T \) is the parameter of interest (homogeneous), then the complete data likelihood under homogeneous assumption is

\[
L_C(\theta) = \prod_{i=1}^n \prod_{j=1}^n \prod_{k=1}^I (1 - \bar{y}_{ij1})(1 - \bar{y}_{ij2}) p_{00} \bar{y}_{ij1}(1 - \bar{y}_{ij2}) p_{10} \bar{y}_{ij2}(1 - p_{00} - p_{10} - p_{01}) \bar{y}_{ij1}\bar{y}_{ij2} \times \prod_{k=1}^I \left( \prod_{i=1}^n S_{\bar{y}_{ik}} \sum_{S_{\bar{y}_{ik}}}^I \left\{ \sum_{j=1}^n \bar{y}_{ijk} > 0 \right\} \prod_{j=1}^n \sum_{S_{\bar{y}_{jk}}}^I \left\{ \sum_{j=1}^n \bar{y}_{ijk} = 0 \right\} \right) \times \left\{ \prod_{j=1}^n S_{\bar{y}_{ijk}} \sum_{S_{\bar{y}_{jk}}}^I \left\{ \sum_{j=1}^n \bar{y}_{ijk} > 0 \right\} \prod_{j=1}^n \sum_{S_{\bar{y}_{jk}}}^I \left\{ \sum_{j=1}^n \bar{y}_{ijk} = 0 \right\} \right\}.
\]

The EM algorithm proceeds as following:

a. Set \( d = 0 \), specify \( \theta^{(0)} \).

b. E-step. Calculate \( \alpha_{00(ij)} \equiv E((1 - \bar{y}_{ij1})(1 - \bar{y}_{ij2}) | Z, Y, \theta^{(d)}) \), \( \alpha_{01(ij)} \equiv E((1 - \bar{y}_{ij1})\bar{y}_{ij2} | Z, Y, \theta^{(d)}) \), and \( \alpha_{10(ij)} \equiv E(\bar{y}_{ij1}(1 - \bar{y}_{ij2}) | Z, Y, \theta^{(d)}) \).

c. M-step. Set \( d = d + 1 \), calculate \( \theta^{(d)} = (p_{00}^{(d)}, p_{10}^{(d)}, p_{01}^{(d)})^T \), where

\[
\begin{align*}
p_{00}^{(d)} &= \frac{1}{nI} \sum_{i=1}^n \sum_{j=1}^I \alpha_{00(ij)}, \quad p_{10}^{(d)} = \frac{1}{nI} \sum_{i=1}^n \sum_{j=1}^I \alpha_{01(ij)}, \quad p_{01}^{(d)} = \frac{1}{nI} \sum_{i=1}^n \sum_{j=1}^I \alpha_{10(ij)}.
\end{align*}
\]

d. Repeat Step b and c until converges, that is, \( \| \theta^{(d+1)} - \theta^{(d)} \| \leq 0.0001 \), where \( \| \cdot \| \) is the \( L_2 \) norm.
Let $\Delta_{y1 y2(ij)}$ denote $\text{pr}(Z_{i1} = z_{i1}, Z_{i2} = z_{i2}, \tilde{Y}_{i1} = y_1, \tilde{Y}_{i1} = y_2)$. We now provide the expressions for the conditional expectations in the E-step.

**Case I:** When $z_{i1} + z_{i2} = 0$, some simple derivations lead to

$$
\Delta_{00(ij)} = S_{e1} S_{e2} Z_{10} \gamma_{00}^{(n)} + S_{e1} S_{e2} p_{00} \gamma_{10}^{(n-1)} + S_{e1} S_{e2} p_{00} \gamma_{01}^{(n-1)} + S_{e1} S_{e2} p_{00} \gamma_{11}^{(n-1)},
$$

$$
\Delta_{10(ij)} = S_{e1} S_{e2} p_{10}^{(d)} + (S_{e1} S_{e2} - S_{e1} S_{e2}) [\gamma_{10}^{(n-1)} + \gamma_{00}^{(n-1)}] p_{10}^{(d)},
$$

$$
\Delta_{01(ij)} = S_{e1} S_{e2} p_{01}^{(d)} + (S_{e1} S_{e2} - S_{e1} S_{e2}) [\gamma_{01}^{(n-1)} + \gamma_{00}^{(n-1)}] p_{01}^{(d)},
$$

and

$$
\Delta_{11(ij)} = S_{e1} S_{e2} p_{11}^{(d)},
$$

where $\gamma_{y1 y2}^{(s)}$ denote the probability that a pool with $s$ has true status “$y_1 y_2$”. It is straightforward to see that $\gamma_{y0}^{(s)} = [p_{00}]^s$, $\gamma_{10}^{(s)} = (p_{00} + p_{10})^s - \gamma_{y0}^{(s)}$, $\gamma_{01}^{(s)} = (p_{00} + p_{01})^s - \gamma_{y0}^{(s)}$, and $\gamma_{11}^{(s)} = 1 - \gamma_{00}^{(s)} - \gamma_{01}^{(s)} - \gamma_{10}^{(s)}$. Note that $\vartheta$ in the E-step takes values of $\vartheta^{(d)}$.

**Case II:** When $z_{i1} + z_{i2} > 0$, we obtain

$$
\Delta_{00(ij)} = a_{00(ij)} \times \left[ S_{p1}^1 S_{p2}^1 S_{e1}^1 S_{e2}^1 Z_{11} Z_{12} Z_{21} Z_{22} b_{00(ij)} + S_{e1} S_{e2} S_{p1}^1 S_{p2}^1 S_{e1} Z_{12} Z_{21} b_{10(ij)}
+ S_{p1}^1 S_{p2}^1 S_{e1} S_{e2} S_{e1} Z_{12} S_{e2} Z_{12} b_{01(ij)} + S_{e1} S_{e2} S_{e1} S_{e2}^1 S_{p1} S_{p2}^1 S_{e2} Z_{21} Z_{12} b_{11(ij)} \right],
$$

$$
\Delta_{10(ij)} = a_{10(ij)} \left[ S_{e1} S_{e2} S_{p2}^1 S_{p2}^1 S_{e2} Z_{11} Z_{12} Z_{21} (b_{00(ij)} + b_{10(ij)})
+ S_{e1} S_{e2} S_{e1} S_{e2}^1 S_{e2} S_{e2} Z_{21} Z_{12} (b_{01(ij)} + b_{11(ij)}) \right],
$$

$$
\Delta_{01(ij)} = a_{01(ij)} \left[ S_{p1}^1 S_{p2}^1 S_{e1} S_{e2} S_{e2} Z_{11} Z_{12} Z_{21} (b_{00(ij)} + b_{01(ij)})
+ S_{e1} S_{e2} S_{e1} S_{e2} S_{e1} S_{e2} Z_{12} Z_{21} (b_{10(ij)} + b_{11(ij)}) \right],
$$

and

$$
\Delta_{11(ij)} = a_{11(ij)} \left[ S_{e1} S_{e2} S_{e1} S_{e2} S_{e1} S_{e2} Z_{11} Z_{12} Z_{21} (b_{00(ij)} + b_{10(ij)} + b_{01(ij)} + b_{11(ij)}) \right],
$$

58
where \( a_{00(ij)} = p_{00} S_{p:1}^{-1} Y_{ij1} S_{p:2}^{-1} Y_{ij2} \), \( a_{10(ij)} = p_{10} S_{e:1}^{-1} Y_{ij1} S_{p:2}^{-1} Y_{ij2} \), and \( a_{01(ij)} = p_{01} S_{p:1}^{-1} Y_{ij1} S_{e:2}^{-1} Y_{ij2} \). In addition, \( b_{00(ij)} = \prod_{l \neq j} S_{p:1}^{-1} Y_{il1} S_{p:2}^{-1} Y_{il2} \). \( b_{01(ij)} = \prod_{l \neq j} S_{p:1}^{-1} Y_{il1} S_{p:2}^{-1} Y_{il2} S_{e:2}^{-1} Y_{il2} \). \( b_{10(ij)} = \prod_{l \neq j} S_{p:1}^{-1} Y_{il1} S_{p:2}^{-1} Y_{il2} S_{p:2}^{-1} Y_{il2} \). \( b_{11(ij)} = \prod_{l \neq j} S_{p:1}^{-1} Y_{il1} S_{p:2}^{-1} Y_{il2} S_{p:2}^{-1} Y_{il2} S_{e:2}^{-1} Y_{il2} \).

In both Case I and Case II, the conditional expectations are given by

\[
\alpha_{00(ij)} = \frac{\Delta_{00(ij)}}{\Delta_{00(ij)} + \Delta_{10(ij)} + \Delta_{01(ij)} + \Delta_{11(ij)}},
\]

\[
\alpha_{10(ij)} = \frac{\Delta_{10(ij)}}{\Delta_{00(ij)} + \Delta_{10(ij)} + \Delta_{01(ij)} + \Delta_{11(ij)}},
\]

and

\[
\alpha_{11(ij)} = \frac{\Delta_{11(ij)}}{\Delta_{00(ij)} + \Delta_{10(ij)} + \Delta_{01(ij)} + \Delta_{11(ij)}}.
\]

This completes the step 1. Note that similar to Tebbs et al. (2013), we found that the initial value of \( \theta^{(0)} \) has minimal effect on the performance of our EM algorithm.

**Step 2.** Based on the estimates \( \hat{p}_{00}, \hat{p}_{10}, \hat{p}_{01}, \hat{p}_{11} \) obtained from step 1, we can estimate \( \Pr(D_{ij1} = d_1, D_{ij2} = d_2 | \bar{Y}_{ij1} = y_1, \bar{Y}_{ij2} = y_2) \), for \( d_1, d_2, y_1, y_2 \in \{0, 1\} \), and \( d_1 + d_2 \neq 0 \). For convenience, we denote the above probability as \( w^{d_1 d_2 y_1 y_2}_{(ij)} \), some simple derivations lead to the following results. When \( y_1 = 1 \) and \( y_2 = 1 \):

\[
w^{d_1 d_2 11}_{(ij)} = (1 - S_{e:1} S_{e:2}) S_{e:1}^{-d_1} S_{e:1}^{-d_1} S_{e:2}^{-d_2} S_{e:2}^{-d_2}.
\]
When $y_1 = 1$ and $y_2 = 0$:

$$w_{d_1d_2}^{10} = S_{e_1}^{d_1} S_{e_1}^{1-d_1} S_{p_2}^{d_2} S_{p_2}^{1-d_2} \left[ (1 - S_{e_1} S_{p_2}) + \left( \gamma_{10}^{(n-1)} + \gamma_{11}^{(n-1)}S_{e_1}(S_{e_2} + S_{p_2} - 1) \right) \right].$$

When $y_1 = 0$ and $y_2 = 1$:

$$w_{d_1d_2}^{01} = S_{p_1}^{d_1} S_{p_1}^{1-d_1} S_{e_2}^{d_2} S_{e_2}^{1-d_2} \left[ (1 - S_{p_1} S_{e_2}) + \left( \gamma_{10}^{(n-1)} + \gamma_{11}^{(n-1)}S_{e_2}(S_{e_1} + S_{p_1} - 1) \right) \right].$$

When $y_1 = 0$ and $y_2 = 0$:

$$w_{d_1d_2}^{00} = S_{p_1}^{d_1} S_{p_1}^{1-d_1} S_{p_2}^{d_2} S_{p_2}^{1-d_2} \left[ (1 - S_{p_1} S_{p_2}) + \gamma_{00}^{(n-1)}(S_{e_1} S_{e_2} - S_{p_1} S_{p_2}) + \gamma_{10}^{(n-1)}S_{e_1}(1 - S_{e_2} - S_{p_2}) + \gamma_{01}^{(n-1)}S_{e_2}(1 - S_{e_1} - S_{p_1}) \right].$$

**Step 3.** Let $\delta_{ij} = (I\{D_{ij1} = 1, D_{ij2} = 1\}, I\{D_{ij1} = 1, D_{ij2} = 0\}, I\{D_{ij1} = 0, D_{ij2} = 1\})^T$, then we have

$$E(\delta_{ij}|X_{ij} = x_{ij}) = W(A + B\eta),$$

where

$$A = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & -1 \\ 0 & 0 & -1 \\ -1 & -1 & 1 \end{pmatrix}, \quad \eta = \begin{pmatrix} \eta_1(x_{ij}^T \beta_1) \\ \eta_2(x_{ij}^T \beta_2) \\ \eta(x_{ij}^T \beta_1, x_{ij}^T \beta_2) \end{pmatrix},$$

and

$$W = \begin{pmatrix} w_{1111} & w_{1110} & w_{1101} & w_{1100} \\ w_{1101} & w_{1100} & w_{1011} & w_{1010} \\ w_{1001} & w_{1000} & w_{0111} & w_{0110} \\ w_{0101} & w_{0100} & w_{0011} & w_{0010} \end{pmatrix}.$$
Denote \( \delta_{ij}^u \equiv (\delta_{ij1}^u, \delta_{ij2}^u, \delta_{ij3}^u)^T = (WB)^{-1}(\delta_{ij} - WA) \), then
\[
E(\delta_{ij3}^u | X_{ij} = x_{ij}) = \eta(x_{ij1}^T \beta_1, x_{ij2}^T \beta_2).
\]
Thus, for any given \( \beta_1 \) and \( \beta_2 \), we obtain a profiling estimator \( \hat{\eta}(\cdot, \cdot) \) through minimizing
\[
\sum_{i=1}^l \sum_{j=1}^n \left[ \delta_{ij3}^u - \eta(x_{ij1}^T \beta_1, x_{ij2}^T \beta_2) \right]^2 \kappa_H(X_{ij1}^T \beta_1 - x_{ij1}^T \beta_1, X_{ij2}^T \beta_2 - x_{ij2}^T \beta_2), \tag{4.2}
\]
for any \( x \), where \( H = (h_1, h_2)^T \) is a bandwidth vector, \( \kappa_H(u, v) = (h_1 h_2)^{-1}K(u/h_1) \)
\( K(v/h_2) \), \( K(\cdot) \) is a univariate kernel function. For simplicity, we use Gaussian kernel, i.e.,
\[
K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \quad \text{for } u \in (-\infty, \infty).
\]
The bandwidth vector \( H \) in objective function 4.2 is chosen by a leaving-one-group-out cross validation. In details, we choose \( \hat{H} \) to maximize log-likelihood (4.1) with \( \eta(\cdot, \cdot) \) being replaced by \( \hat{\eta}^u(\cdot, \cdot) \), where \( \hat{\eta}^u(\cdot, \cdot) \) is obtained through minimizing
\[
\sum_{i=1}^l \sum_{j=1}^n \left[ \delta_{ij3}^u - \eta(x_{ij1}^T \beta_1, x_{ij2}^T \beta_2) \right]^2 \kappa_H(X_{ij1}^T \beta_1 - x_{ij1}^T \beta_1, X_{ij2}^T \beta_2 - x_{ij2}^T \beta_2), \tag{4.3}
\]
Herein \( \beta_1 \) and \( \beta_2 \) in objective function 4.3 are obtained marginally, i.e., we minimize
\[
\sum_{i=1}^l \sum_{j=1}^n \left[ \delta_{ijk} - \frac{\exp(x_{ij1}^T \beta_k)}{1 + \exp(x_{ij1}^T \beta_k)} \right]^2,
\]
for \( k = 1, 2 \).

Several notes are in order. First, the profiling estimator \( \hat{\eta}(x_{ij1}^T \beta_1, x_{ij2}^T \beta_2) \) has closed form. It is given by
\[
\hat{\eta}(x_{ij1}^T \beta_1, x_{ij2}^T \beta_2) = \frac{\sum_{i=1}^l \sum_{j=1}^n \delta_{ij3}^u \kappa_H(X_{ij1}^T \beta_1 - x_{ij1}^T \beta_1, X_{ij2}^T \beta_2 - x_{ij2}^T \beta_2)}{\sum_{i=1}^l \sum_{j=1}^n \kappa_H(X_{ij1}^T \beta_1 - x_{ij1}^T \beta_1, X_{ij2}^T \beta_2 - x_{ij2}^T \beta_2)}.
\]
Second, since \( \hat{\eta}(x_{ij1}^T \beta_1, x_{ij2}^T \beta_2) \) is the estimated probability of \( I_{ij} \) being infected with both infections, it must be truncated between 0 and \( \min\{\eta_1(x_{ij1}^T \beta_1), \eta_2(x_{ij2}^T \beta_2)\} \).
Through step 1 to step 3, we can obtain the profiling estimator \( \hat{\eta}(\cdot, \cdot) \). The last step is to maximize log-likelihood function 4.1 with respect to \( \beta_1 \) and \( \beta_2 \) to obtain the MLE \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \). This can be done easily through any modern statistical software.

**Estimate \( \hat{\sigma}_k \)**

In the above section, we obtain \( \hat{\theta} = (\hat{\beta}_1, \hat{\beta}_2) \). We now estimate variance-covariance matrix for \( \hat{\theta} \).

**Variance-covariance matrix \( \Sigma_{\hat{\theta}} \)**

For notation simplicity, we write the likelihood as

\[
l(\theta) = \sum_{i=1}^{I} \log R(Z_i, Y_i, X_i, \theta, \eta_\theta),
\]

where \( \theta = (\beta_1^T, \beta_2^T)^T \). We write \( \eta(\cdot, \cdot) \) as \( \eta_\theta \) to highlight the fact that it is dependent on \( \theta \). Further, we denote the first derivative of log-likelihood as \( G(\theta) = \frac{\partial l(\theta)}{\partial \theta} \), then

\[
G(\hat{\theta}) = \sum_{i=1}^{I} \frac{\partial}{\partial \theta} R(Z_i, Y_i, X_i, \theta, \eta_\theta) \bigg|_{\theta = \hat{\theta}} = \sum_{i=1}^{I} A_i(\hat{\theta})B_i(\hat{\theta}),
\]

where \( A_i(\hat{\theta}) = R^{-1}(Z_i, Y_i, X_i, \theta, \eta_\theta) \bigg|_{\theta = \hat{\theta}} \) and \( B_i(\hat{\theta}) = \frac{\partial}{\partial \theta} R(Z_i, Y_i, X_i, \theta, \eta_\theta) \bigg|_{\theta = \hat{\theta}} \).

We now derive \( A_i(\hat{\theta}) \) and \( B_i(\hat{\theta}) \).

**Case I:** When \( G_i \) is tested negatively for both infections,

\[
A_i(\theta) = \left( \gamma_{i,00}S_{p:1}S_{p:2} + \gamma_{i,10}S_{e:1}S_{p:2} + \gamma_{i,01}S_{p:1}S_{e:2} + \gamma_{i,11}S_{e:1}S_{e:2} \right)^{-1},
\]

and

\[
B_i(\theta) = (S_{p:1}S_{p:2} - S_{e:1}S_{e:2})\gamma'_{i,00} + (S_{e:1}S_{p:2} - S_{e:1}S_{e:2})\gamma'_{i,10} + (S_{p:1}S_{e:2} - S_{e:1}S_{e:2})\gamma'_{i,01},
\]

where \( \gamma'_{i,00} = \left( \frac{\partial \gamma_{i,00}}{\partial \beta_1}, \frac{\partial \gamma_{i,00}}{\partial \beta_2} \right)^T \), in which \( \frac{\partial \gamma_{i,00}}{\partial \beta_k} = \sum_{j=1}^{n} r_k(\beta_k) \prod_{j' \neq j} p_{00(i,j)} \) and \( r_k(\beta_k) = \frac{\partial p_{00(i,j)}}{\partial \beta_k} \). In the next section, we will provide details for deriving \( p_{00(i,j)}(\beta_k) \), along with \( p_{10(i,j)}(\beta_k) \) and \( p_{01(i,j)}(\beta_k) \) which will come up momentarily.
Similarly, \( \gamma'_{i,10} = \left( \frac{\partial \gamma_{i,10}}{\partial \beta_1}, \frac{\partial \gamma_{i,10}}{\partial \beta_2} \right)^T \), where

\[
\frac{\partial \gamma_{i,10}}{\partial \beta_k} = \sum_{j=1}^{n} (p_{00(ij)} + p_{10(ij)}) \prod_{j' \neq j} (p_{00(ij')} + (p_{10(ij')}))) - \frac{\partial \gamma_{i,00}}{\partial \beta_k},
\]

and \( \gamma'_{i,01} = \left( \frac{\partial \gamma_{i,01}}{\partial \beta_1}, \frac{\partial \gamma_{i,01}}{\partial \beta_2} \right)^T \), where

\[
\frac{\partial \gamma_{i,01}}{\partial \beta_k} = \sum_{j=1}^{n} (p_{00(ij)} + p_{01(ij)}) \prod_{j' \neq j} (p_{00(ij')} + (p_{01(ij')}))) - \frac{\partial \gamma_{i,00}}{\partial \beta_k}.
\]

**Case II:** When \( \mathcal{G}_i \) is tested positively for at least one infection, then

\[
A_i(\theta) = \left( h_{00}^i(y_i) S_{p:1}^{1-z_{i1}} S_{p:2}^{1-z_{i2}} S_{p:2}^{s_{i2}} + h_{10}^i(y_i) S_{e:1}^{1-z_{i1}} S_{e:2}^{1-z_{i2}} S_{e:2}^{s_{i2}} + h_{01}^i(y_i) S_{p:1}^{1-z_{i1}} S_{p:2}^{1-z_{i2}} S_{p:2}^{s_{i2}} + h_{11}^i(y_i) S_{e:1}^{1-z_{i1}} S_{e:2}^{1-z_{i2}} S_{e:2}^{s_{i2}} \right)^{-1},
\]

where

\[
h_{00}^i(y_i) = \prod_{j=1}^{n} \left\{ p_{00(ij)} S_{p:1}^{1-Y_{ij1}} S_{p:2}^{1-Y_{ij2}} S_{p:2}^{Y_{ij2}} - h_{00}^i(y_i) \right\},
\]

\[
h_{10}^i(y_i) = \prod_{j=1}^{n} \left\{ p_{00(ij)} S_{p:1}^{1-Y_{ij1}} S_{p:2}^{1-Y_{ij2}} S_{p:2}^{Y_{ij2}} + p_{10(ij)} S_{e:1}^{1-Y_{ij1}} S_{e:2}^{1-Y_{ij2}} S_{e:2}^{Y_{ij2}} - h_{00}^i(y_i) \right\},
\]

\[
h_{01}^i(y_i) = \prod_{j=1}^{n} \left\{ p_{00(ij)} S_{p:1}^{1-Y_{ij1}} S_{p:2}^{1-Y_{ij2}} S_{p:2}^{Y_{ij2}} + p_{01(ij)} S_{p:1}^{1-Y_{ij1}} S_{p:2}^{1-Y_{ij2}} S_{p:2}^{Y_{ij2}} - h_{00}^i(y_i) \right\},
\]

and

\[
h_{11}^i(y_i) = \prod_{j=1}^{n} \left\{ p_{00(ij)} S_{p:1}^{1-Y_{ij1}} S_{p:2}^{1-Y_{ij2}} S_{p:2}^{Y_{ij2}} + p_{10(ij)} S_{e:1}^{1-Y_{ij1}} S_{e:2}^{1-Y_{ij2}} S_{e:2}^{Y_{ij2}} + p_{01(ij)} S_{p:1}^{1-Y_{ij1}} S_{p:2}^{1-Y_{ij2}} S_{p:2}^{Y_{ij2}} + p_{11(ij)} S_{e:1}^{1-Y_{ij1}} S_{e:2}^{1-Y_{ij2}} S_{e:2}^{Y_{ij2}} - h_{00}^i(y_i) - h_{10}^i(y_i) - h_{01}^i(y_i) \right\}.
\]
In addition,

\[
B_i(\theta) = h''_{00}(y_i)S_{p_1}^{1-z_{i1}}\overline{S}_{p_1}^{z_{i1}}S_{p_2}^{1-z_{i2}}\overline{S}_{p_2}^{z_{i2}} + h''_{10}(y_i)S_{e_1}^{z_{i1}}\overline{S}_{e_1}^{1-z_{i1}}S_{p_2}^{1-z_{i2}}\overline{S}_{e_2}^{z_{i2}}
\]

where \( h''_{ab}(y_i) = \left( \frac{\partial h_a(y_i)}{\partial \beta_1}, \frac{\partial h_a(y_i)}{\partial \beta_2} \right) \) for \( a, b \in \{0, 1\} \), in which

\[
\frac{\partial h_{00}^i(y_i)}{\partial \beta_k} = \sum_{j=1}^{n} \left\{ p_{00(ij)}^{(\beta_k)} S_{p_1}^{1-Y_{ij1}}\overline{S}_{p_1}^{Y_{ij1}}S_{p_2}^{1-Y_{ij2}}\overline{S}_{p_2}^{Y_{ij2}} \right\} \prod_{j' \neq j} \left\{ p_{00(ij')}^{(\beta_k)} S_{p_1}^{1-Y_{ij'1}}\overline{S}_{p_1}^{Y_{ij'1}}S_{p_2}^{1-Y_{ij'2}}\overline{S}_{p_2}^{Y_{ij'2}} \right\} - \frac{\partial h_{00}^i(y_i)}{\partial \beta_k},
\]

\[
\frac{\partial h_{10}^i(y_i)}{\partial \beta_k} = \sum_{j=1}^{n} \left\{ p_{01(ij)}^{(\beta_k)} S_{p_1}^{1-Y_{ij1}}\overline{S}_{p_1}^{Y_{ij1}}S_{p_2}^{1-Y_{ij2}}\overline{S}_{p_2}^{Y_{ij2}} + p_{10(ij)}^{(\beta_k)} S_{e_1}^{Y_{ij1}}\overline{S}_{e_1}^{1-Y_{ij1}}S_{p_2}^{1-Y_{ij2}}\overline{S}_{p_2}^{Y_{ij2}} \right\}
\times \prod_{j' \neq j} \left\{ p_{00(ij')}^{(\beta_k)} S_{p_1}^{1-Y_{ij'1}}\overline{S}_{p_1}^{Y_{ij'1}}S_{p_2}^{1-Y_{ij'2}}\overline{S}_{p_2}^{Y_{ij'2}} + p_{10(ij')}^{(\beta_k)} S_{e_1}^{Y_{ij'1}}\overline{S}_{e_1}^{1-Y_{ij'1}}S_{e_2}^{1-Y_{ij'2}}\overline{S}_{e_2}^{Y_{ij'2}} \right\} - \frac{\partial h_{00}^i(y_i)}{\partial \beta_k},
\]

and

\[
\frac{\partial h_{11}^i(y_i)}{\partial \beta_k} = \sum_{j=1}^{n} \left\{ p_{01(ij)}^{(\beta_k)} S_{p_1}^{1-Y_{ij1}}\overline{S}_{p_1}^{Y_{ij1}}S_{p_2}^{1-Y_{ij2}}\overline{S}_{p_2}^{Y_{ij2}} + p_{10(ij)}^{(\beta_k)} S_{p_1}^{1-Y_{ij1}}\overline{S}_{p_1}^{Y_{ij1}}S_{e_2}^{1-Y_{ij2}}\overline{S}_{e_2}^{Y_{ij2}} \right\}
\times \prod_{j' \neq j} \left\{ p_{00(ij')}^{(\beta_k)} S_{p_1}^{1-Y_{ij'1}}\overline{S}_{p_1}^{Y_{ij'1}}S_{p_2}^{1-Y_{ij'2}}\overline{S}_{p_2}^{Y_{ij'2}} + p_{10(ij')}^{(\beta_k)} S_{e_1}^{Y_{ij'1}}\overline{S}_{e_1}^{1-Y_{ij'1}}S_{e_2}^{1-Y_{ij'2}}\overline{S}_{e_2}^{Y_{ij'2}} \right\} - \frac{\partial h_{00}^i(y_i)}{\partial \beta_k} - \frac{\partial h_{10}^i(y_i)}{\partial \beta_k} - \frac{\partial h_{01}^i(y_i)}{\partial \beta_k}.
\]
Now denote the true parameter as \( \theta_0 \), then
\[
G(\hat{\theta}) - G(\theta_0) = \sum_{i=1}^{I} A_i(\hat{\theta})B_i(\hat{\theta}) - \sum_{i=1}^{I} A_i(\theta_0)B_i(\theta_0)
\]
\[
\approx \sum_{i=1}^{I} (\hat{\theta} - \theta_0)^T \left[ \frac{\partial A_i(\theta)}{\partial \theta} \right] |_{\theta = \hat{\theta}} B_i(\hat{\theta})
\]
\[
= \sum_{i=1}^{I} (\hat{\theta} - \theta_0)^T \left[ -A_i^2(\hat{\theta})B_i(\hat{\theta}) \right] B_i(\hat{\theta})
\]
\[
= - \left[ \sum_{i=1}^{I} A_i^2(\hat{\theta})B_i(\hat{\theta})B_i^T(\hat{\theta}) \right] (\hat{\theta} - \theta_0) = -\Sigma_\theta(\hat{\theta} - \theta_0),
\]
where \( \hat{\theta} \) is a vector between \( \theta_0 \) and \( \hat{\theta} \) and here takes value on \( \hat{\theta} \), therefore \( \Sigma_\theta = \sum_{j=1}^{J} A_j^2(\hat{\theta})B_j(\hat{\theta})B_j^T(\hat{\theta}) \). Clearly,
\[
\hat{\theta} - \theta_0 \approx -\Sigma^{-1}_\theta \left[ G(\hat{\theta}) - G(\theta_0) \right] = \Sigma^{-1}_\theta G(\theta_0),
\]
where \( G(\theta_0) \sim N(0, \Sigma_\theta) \). Hence \( \hat{\theta} - \theta_0 \xrightarrow{d} N \left( 0, \Sigma^{-1}_\theta \right) \).

Estimate \( p^{(\beta_k)}_{00(ij)} \), \( p^{(\beta_k)}_{10(ij)} \), and \( p^{(\beta_k)}_{01(ij)} \)

As shown in previous section, \( \Sigma_\theta \) requires \( p_{00(ij)}, p_{10(ij)}, p_{01(ij)} \), and \( p_{11(ij)} \) evaluated at \( \theta = \hat{\theta} \). In addition, it requires \( \hat{p}^{(\beta_k)}_{00(ij)}, \hat{p}^{(\beta_k)}_{10(ij)} \), and \( \hat{p}^{(\beta_k)}_{01(ij)} \). These partial derivatives can be obtained through estimating \( \hat{\eta}^{(\beta_k)} = \frac{\partial \hat{\eta}(x^T \beta_1, x^T \beta_2)}{\partial \beta_k} \). First of all, it can be shown that
\[
\frac{\partial \hat{\eta}(u^T \beta_1, x^T \beta_2)}{\partial \beta_1} \xrightarrow{p} \eta_u(x^T \beta_1, x^T \beta_2) \times \left[ d(x^T \beta_1, x^T \beta_2) - x \right] \times \int u \kappa_u(u, v) dudv,
\]
\[
\frac{\partial \hat{\eta}(x^T \beta_1, x^T \beta_2)}{\partial \beta_2} \xrightarrow{p} \eta_v(x^T \beta_1, x^T \beta_2) \times \left[ d(x^T \beta_1, x^T \beta_2) - x \right] \times \int v \kappa_v(u, v) dudv,
\]
(4.4)

where \( \eta_u(u, v) \) and \( \kappa_u(u, v) \) are derivatives of \( \eta(u, v) \) and \( \kappa(u, v) \) with respect to \( u \), respectively, and \( \eta_v(u, v) \) and \( \kappa_v(u, v) \) are derivatives of \( \eta(u, v) \) and \( \kappa(u, v) \) with respect to \( v \), respectively. We now evaluate the right hand side of Expression 4.4.

First of all, since \( K(\cdot) \) is Gaussian kernel, it is straightforward to see that
\[
\int u \kappa_u(u, v) dudv = \int v \kappa_v(u, v) dudv = -1.
\]
Next, we estimate $\eta_u(x^T \beta_1, x^T \beta_2)$ and $\eta_v(x^T \beta_1, x^T \beta_2)$ through minimizing

$$
\sum_{i=1}^{I} \sum_{j=1}^{n} \left[ \delta_{ij3} - \eta(x^T \beta_1, x^T \beta_2) - \eta_u(x^T \beta_1, x^T \beta_2) (X_{ij}^T \beta_1 - x^T \beta_1) 
- \eta_v(x^T \beta_1, x^T \beta_2) (X_{ij}^T \beta_2 - x^T \beta_2) \right]^2 \times \kappa_H(X_{ij}^T \beta_1 - x^T \beta_1, X_{ij}^T \beta_2 - x^T \beta_2).
$$

The above expression can be minimized through matrix representation of Weighted Least Square, i.e., let

$$
\delta_w = \begin{pmatrix}
\delta_{113} \\
\delta_{123} \\
\vdots \\
\delta_{1n3} \\
\delta_{213} \\
\vdots \\
\delta_{ln3}
\end{pmatrix}, \quad X_w = \begin{pmatrix}
1, X_{11}^T \beta_1 - x^T \beta_1, X_{11}^T \beta_2 - x^T \beta_2 \\
1, X_{12}^T \beta_1 - x^T \beta_1, X_{12}^T \beta_2 - x^T \beta_2 \\
\vdots \\
1, X_{ln}^T \beta_1 - x^T \beta_1, X_{ln}^T \beta_2 - x^T \beta_2
\end{pmatrix},
$$

and $W = \text{diag}(\kappa_H(X_{11}^T \beta_1 - x^T \beta_1, X_{11}^T \beta_2 - x^T \beta_2), \ldots, \kappa_H(X_{ln}^T \beta_1 - x^T \beta_1, X_{ln}^T \beta_2 - x^T \beta_2))$, then the solution $\hat{\beta}_w$ is

$$
\hat{\beta}_w = \left( X_w^T W X_w \right)^{-1} X_w^T W \delta_w.
$$

Lastly, we estimate $d(x^T \beta_1, x^T \beta_2)$ through minimizing

$$
\sum_{i=1}^{I} \sum_{j=1}^{n} \left\{ X_{ij} - d(x^T \beta_1, x^T \beta_2) \right\}^2 \kappa_H(X_{ij}^T \beta_1 - x^T \beta_1, X_{ij}^T \beta_2 - x^T \beta_2).
$$

Hence, our estimate of $d(x^T \beta_1, x^T \beta_2)$ is

$$
\hat{d}(x^T \beta_1, x^T \beta_2) = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n} X_{ij} \kappa_H(X_{ij}^T \beta_1 - x^T \beta_1, X_{ij}^T \beta_2 - x^T \beta_2)}{\sum_{i=1}^{I} \sum_{j=1}^{n} \kappa_H(X_{ij}^T \beta_1 - x^T \beta_1, X_{ij}^T \beta_2 - x^T \beta_2)}.
$$
This completes the derivation of \( \hat{\eta}(\beta_k) \). Therefore,

\[
P^{(\beta_k)}_{10(ij)} = \left( X^T_{ij} \frac{e^{X^T_{ij}\beta_1}}{(1 + e^{X^T_{ij}\beta_1})^2} 0^T \right)^T \ - \hat{\eta}(\beta_k),
\]

\[
P^{(\beta_k)}_{01(ij)} = \left( 0^T, X^T_{ij} \frac{e^{X^T_{ij}\beta_2}}{(1 + e^{X^T_{ij}\beta_2})^2} \right)^T \ - \hat{\eta}(\beta_k),
\]

and \( p^{(\beta_k)}_{00(ij)} = -p^{(\beta_k)}_{01(ij)} - p^{(\beta_k)}_{10(ij)} - \hat{\eta}(\beta_k) \).

### 4.3 Simulation

In this section, we use simulation study to evaluate the performance of the proposed method in Section 4.2. We first generate \( X \) for \( N = 3000 \) individuals and randomly assign these individuals into \( I \) groups, each with size \( n \). For given \( \beta_1 \) and \( \beta_2 \), we calculate \( \eta_k(x^T_{ij}\beta_k) \) for \( i = 1, 2, \ldots, I, j = 1, 2, \ldots, n, \) and \( k = 1, 2 \). These values are treated as the true marginal probabilities for each individual. Based on pre-specified \( \eta(\cdot, \cdot) \), we generate \( \eta(x^T_{ij}\beta_1, x^T_{ij}\beta_2) \), which represents the probability that \( I_{ij} \) is truly positive for both infections. Then, we sample \( \tilde{Y}_{ij} \) based on multinomial cell probabilities \( \vartheta_{(ij)} = (p_{00(ij)}, p_{10(ij)}, p_{01(ij)}, p_{11(ij)}) \), where \( pr(\tilde{Y}_{ij} = (y_1, y_2)^T) = p_{y_1y_2(ij)} \), and \( p_{00(ij)} = 1 - \eta_1(x^T_{ij}\beta_1) - \eta_2(x^T_{ij}\beta_2) + \eta(x^T_{ij}\beta_1, x^T_{ij}\beta_2) \), \( p_{10(ij)} = \eta_1(x^T_{ij}\beta_1) - \eta(x^T_{ij}\beta_1, x^T_{ij}\beta_2) \), \( p_{01(ij)} = \eta_2(x^T_{ij}\beta_2) - \eta(x^T_{ij}\beta_1, x^T_{ij}\beta_2) \), and \( p_{11(ij)} = \eta(x^T_{ij}\beta_1, x^T_{ij}\beta_2) \).

To generate group responses, we mimic the Iowa IPP algorithm. We first simulate group responses \( Z_{ik} \) from \( Z_{ik} = B^+_k I\{\sum_{j=1}^n \tilde{Y}_{ijk} > 0\} + B^-_k I\{\sum_{j=1}^n \tilde{Y}_{ijk} = 0\} \) for \( k = 1, 2 \), where \( B^+_k \) and \( B^-_k \) are Bernoulli random variables with probability of success \( S_{e,k} \) and \( S_{p,k} \), respectively. If \( Z_{i1} + Z_{i2} > 0 \), we simulate individual test responses as \( D_{ijk} = Y_{ijk} = B^+_k I\{\tilde{Y}_{ijk} > 0\} + B^-_k I\{\tilde{Y}_{ijk} = 0\} \); otherwise, we let \( D_{ijk} = 0 \).

In our investigation, we let \( S_{e,k} = S_{p,k} = 0.95 \). These values are chosen to mimic the sensitivity and specificity of AC2A to screen for chlamydia and gonorrhea. To explore the effect of different group sizes on estimation efficiency, we choose \( n = 2 \) and 5. In our first study, we let \( \beta_1 = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}) = (-3, -1, 2, 0) \) and \( \beta_2 = \ldots \)
Table 4.1: Simulation results for 500 data sets, each with \( N = 3000 \) individuals, when \((X_1, X_2, X_3)^T \sim \text{MVN}_3(0, \Omega)\), where \( \Omega_{ij} = 0.5^{i-j} \). Clayton Copula is used to generate \( \eta(\cdot, \cdot) \), i.e., \( \eta(x_{ij}^T \beta_1, x_{ij}^T \beta_2) = \left[ \max \left\{ (x_{ij}^T \beta_1)^{-0.5} + (x_{ij}^T \beta_2)^{-0.5} - 1; 0 \right\} \right]^{-2} \). The columns labeled “IPP” give the results for group testing, and the columns labeled “Individual” give the results for individual testing. Note that \( S_{e,k} = 0.95 \), \( S_{p,k} = 0.95 \), for \( k = 1, 2 \).

\[
\begin{array}{cccccc|cccccc}
\hline
 & \text{IPP} & & \text{Individual} & & \\
 & n = 2 (T = 2512.5) & & n = 5 (T = 2346.2) & & \\
 & \text{True} & \text{Mean} & \text{SD} & \text{SE} & \text{Cov} & \text{Mean} & \text{SD} & \text{SE} & \text{Cov} & \\
\beta_{10} & -3 & -3.03 & 0.13 & 0.13 & 0.96 & -3.02 & 0.13 & 0.14 & 0.97 & -3.02 & 0.16 \\
\beta_{11} & -1 & -1.00 & 0.10 & 0.10 & 0.95 & -1.00 & 0.11 & 0.11 & 0.95 & -1.01 & 0.11 \\
\beta_{12} & 2 & 2.03 & 0.13 & 0.14 & 0.96 & 2.02 & 0.15 & 0.15 & 0.96 & 2.02 & 0.15 \\
\beta_{13} & 0 & -0.05 & 0.09 & 0.09 & 0.93 & -0.04 & 0.09 & 0.09 & 0.94 & 0.00 & 0.09 \\
\beta_{20} & -4 & -4.06 & 0.21 & 0.22 & 0.95 & -4.07 & 0.23 & 0.24 & 0.95 & -4.05 & 0.28 \\
\beta_{21} & 0 & 0.05 & 0.12 & 0.12 & 0.91 & 0.05 & 0.12 & 0.13 & 0.95 & 0.01 & 0.14 \\
\beta_{22} & -1 & -1.07 & 0.14 & 0.16 & 0.95 & -1.07 & 0.16 & 0.16 & 0.92 & -1.01 & 0.17 \\
\beta_{23} & 2 & 2.00 & 0.17 & 0.18 & 0.95 & 2.01 & 0.18 & 0.19 & 0.96 & 2.02 & 0.21 \\
\hline
\end{array}
\]

\((\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23}) = (-4, 0, -1, 2)\). In addition, we consider a vector of predictors of the form \( X = (1, X_1, X_2, X_3) \), where \((X_1, X_2, X_3)^T \sim \text{MVN}_3(0, \Omega)\), \( \Omega = (\Omega_{ij}) \) and \( \Omega_{ij} = 0.5^{i-j} \). We also choose two different curves for \( \eta(\cdot, \cdot) \). The first curve is determined by Clayton Copula, i.e.,

\[
\eta(x_{ij}^T \beta_1, x_{ij}^T \beta_2) = \left[ \max \left\{ (x_{ij}^T \beta_1)^{-0.5} + (x_{ij}^T \beta_2)^{-0.5} - 1; 0 \right\} \right]^{-2}.
\]

The second curve is determined by

\[
\eta(x_{ij}^T \beta_1, x_{ij}^T \beta_2) = \eta_1(x_{ij}^T \beta_1) \eta_2(x_{ij}^T \beta_2) \sin^2 \left( 2\pi(x_{ij}^T \beta_1 + x_{ij}^T \beta_2) \right).
\]

By using these settings, the prevalence for first infection is around 10% and second is around 5%, which are close to the prevalence for chlamydia and gonorrhea.

The entire process is repeated 500 times. The averaged estimates of \( \beta_k \), the averaged standard errors (SE), and the coverage probability (Cov) are given in Table 4.1 and Table 4.2, for two different \( \eta(\cdot, \cdot) \), along with the sample standard deviations of \( \hat{\beta}_k \) (SD). We also include the average number of tests \( T \). In addition, we compare our method with individual testing estimation results, i.e., we simulate individual testing responses \( Y_{mk} \) based on each \( \tilde{Y}_{mk} \), for \( m = 1, 2, \ldots, nI \), \( k = 1, 2 \), and use \( Y_{mk} \).
Table 4.2: Simulation results for 500 data sets, each with $N = 3000$ individuals, when $S_{c:k} = 0.95$, $S_{p:k} = 0.95$, for $k = 1, 2$, $(X_1, X_2, X_3)^T \sim \text{MVN}_3(0, \Omega)$, $\Omega_{ij} = 0.5^{[i-j]}$. The curve $\eta(\cdot, \cdot)$ is $\eta(x_i^T \beta_1, x_i^T \beta_2) = \eta_1(x_i^T \beta_1)\eta_2(x_i^T \beta_2)\sin^2 \left(2\pi(x_i^T \beta_1 + x_i^T \beta_2)\right)$. The columns labeled “IPP” give the results for group testing, and the columns labeled “Individual” give the results for individual testing.

<table>
<thead>
<tr>
<th>$n=2$ ($T = 2512.5$)</th>
<th>$n=5$ ($T = 2346.2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{10}$</td>
<td>$\beta_{11}$</td>
</tr>
<tr>
<td>True Mean SD SE Cov</td>
<td>Mean SD SE Cov</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>$-3$</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>$-1$</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>$2$</td>
</tr>
<tr>
<td>$\beta_{13}$</td>
<td>$0$</td>
</tr>
<tr>
<td>$\beta_{20}$</td>
<td>$-4$</td>
</tr>
<tr>
<td>$\beta_{21}$</td>
<td>$0$</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>$-1$</td>
</tr>
<tr>
<td>$\beta_{23}$</td>
<td>$2$</td>
</tr>
</tbody>
</table>

to estimate $\beta_k$ via maximizing

$$\sum_{m=1}^{nI} [Y_{mk}\log(p^*_mk) + (1 - Y_{mk})\log(1 - p^*_mk)],$$

for $k = 1, 2$, where $p^*_mk = \overline{S}_{p:k} + (S_{c:k} + S_{p:k} - 1) \times e^{x_i^T \beta_k}/(1 + e^{x_i^T \beta_k})$. Note that the number of tests for individual testing is always $N$.

In Table 4.3 and Table 4.4, we give the results for our second study. We let $\beta_1 = (1.9, 0, 0.7, -1.5)$ and $\beta_2 = (-5, -3, 2.2, 0)$, $X_1 \sim N(0, 1)$, $X_2 \sim \text{Bern}(0.6)$ and $X_3 \sim \text{Bern}(0.2)$. We choose this setting because all the predictors collected by Iowa IPP are binary variables, except for age. The curve $\eta(\cdot, \cdot)$ is the same as in the first study.

The results in Table 4.1 - 4.4 demonstrate that our method performs well. The mean estimates are very close to the true value of the parameters, and the averaged standard errors are in close agreement to the standard deviations. Estimated coverage probabilities for nominal 95% Wald confidence intervals are usually within the margin of Monte Carlo error (0.03 assuming a 99% confidence). In addition, it is worth noting that the standard error increases as the pool size increases while holding the total number of individuals constant. This is intuitive since as the sample size increases,
Table 4.3: Simulation results for 500 data sets, each with $N = 3000$ individuals, when $X_1 \sim N(0,1), X_2 \sim \text{Bern}(0.6)$, and $X_3 \sim \text{Bern}(0.2)$. Clayton Copula is used to generate $\eta(\cdot, \cdot)$, i.e., $\eta(x_{ij}^T \beta_1, x_{ij}^T \beta_2) = \left[ \max \{ (x_{ij}^T \beta_1)^{-0.5} + (x_{ij}^T \beta_2)^{-0.5} - 1 \} \right]^{-2}$. The columns labeled “IPP” give the results for group testing, and the columns labeled “Individual” give the results for individual testing. Note that $S_{c:k} = 0.95$, $S_{p:k} = 0.95$, for $k = 1, 2$.

<table>
<thead>
<tr>
<th></th>
<th>IPP</th>
<th>Individual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 2 (T = 2512.5)$</td>
<td>$n = 5 (T = 2346.2)$</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>-1.9</td>
<td>-1.90 0.12</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>0</td>
<td>0.00 0.06</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>0.7</td>
<td>0.02 0.06</td>
</tr>
<tr>
<td>$\beta_{13}$</td>
<td>-1.5</td>
<td>0.12 0.06</td>
</tr>
<tr>
<td>$\beta_{20}$</td>
<td>-5</td>
<td>-5.07 0.43</td>
</tr>
<tr>
<td>$\beta_{21}$</td>
<td>-3</td>
<td>-3.02 0.31</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>-2.2</td>
<td>-2.29 0.34</td>
</tr>
<tr>
<td>$\beta_{23}$</td>
<td>0</td>
<td>0.03 0.38</td>
</tr>
</tbody>
</table>

Table 4.4: Simulation results for 500 data sets, each with $N = 3000$ individuals, when $S_{c:k} = 0.95$, $S_{p:k} = 0.95$, for $k = 1, 2$, $X_1 \sim N(0,1), X_2 \sim \text{Bern}(0.6)$, and $X_3 \sim \text{Bern}(0.2)$. The curve $\eta(\cdot, \cdot)$ is $\eta(x_{ij}^T \beta_1, x_{ij}^T \beta_2) = \eta_1(x_{ij}^T \beta_1)\eta_2(x_{ij}^T \beta_2)\sin^2(2\pi(x_{ij}^T \beta_1 + x_{ij}^T \beta_2))$. The columns labeled “IPP” give the results for group testing, and the columns labeled “Individual” give the results for individual testing.

<table>
<thead>
<tr>
<th></th>
<th>IPP</th>
<th>Individual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 2 (T = 2512.5)$</td>
<td>$n = 5 (T = 2346.2)$</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>-1.9</td>
<td>-1.91 0.11</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>0</td>
<td>0.00 0.06</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>0.7</td>
<td>0.01 0.05</td>
</tr>
<tr>
<td>$\beta_{13}$</td>
<td>-1.5</td>
<td>0.06 0.07</td>
</tr>
<tr>
<td>$\beta_{20}$</td>
<td>-5</td>
<td>-5.07 0.39</td>
</tr>
<tr>
<td>$\beta_{21}$</td>
<td>-3</td>
<td>-3.05 0.27</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>-2.2</td>
<td>-2.24 0.32</td>
</tr>
<tr>
<td>$\beta_{23}$</td>
<td>0</td>
<td>0.00 0.35</td>
</tr>
</tbody>
</table>

the number of pools decreases. Further, it is interesting to see that the standard errors obtained through our method by using group testing data are smaller than those obtained from individual testing. Similar phenomenon has been observed in group testing estimation literature for a single infection.
4.4 Application

Tebbs et al. (2013) used individual testing data from the IPP to estimate population prevalence when screening for chlamydia and gonorrhea in Nebraska. Zhang et al. (2013) also used this data to estimate individual-level probability of each infection, assuming no further retesting information. In this project, we incorporate the retesting information in our framework and provide interpretable inference for each infection on an individual basis.

Our data set consists of 14,530 swab testing results for all female individuals screened in Nebraska during 2009. During this time, all tests were performed at the Nebraska Public Health Laboratory (NPHL) in Omaha. To evaluate our method on the group testing data, we first assign the 2009 individuals to master pools of size five chronologically based on the specimen’s arrival date at the NPHL. Similar or identical group sizes are used in Morre et al. (2001), Rours et al. (2005), Clark et al. (2001), and Zhang et al. (2013) for chlamydia and gonorrhea screening. We treat the 2009 individuals’ responses as the “true” statuses; we then test and decode pools ourselves by simulating test outcomes using the assay sensitivities and specificities reported for the AC2A; see Appendix A.6 for more details.

In addition to the individual testing responses, the NPHL also collected several covariates on each individual. In this section, we consider the following four covariates in our model: age, whether individual had contacted someone with sexually transmitted disease, whether individual had contracted cervicitis, and pelvic inflammatory disease (PID). Except for age, all other covariates are binary variables. These covariates are also considered in Zhang et al. (2013). To average over the effect of simulation, we repeat this procedure 500 times. The results are displayed in Table 4.5, along with the results obtained by individual testing estimation method explained in Section 4.3.

Table 4.5 shows that our method works well. The averaged parameter estimates
Table 4.5: 2009 Nebraska female swab estimation results. The columns labeled “IPP” give the results for group testing, and the columns labeled “Individual” give the results for individual testing. The mean estimates from “IPP” and “Individual” and standard error estimates from “IPP” are averaged over 500 data sets.

<table>
<thead>
<tr>
<th>Infection</th>
<th>Covariate</th>
<th>IPP (n = 5)</th>
<th>Individual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
<td>Mean</td>
</tr>
<tr>
<td>Chlamydia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.90</td>
<td>0.24</td>
<td>-0.91</td>
</tr>
<tr>
<td>Age</td>
<td>-0.08</td>
<td>0.01</td>
<td>-0.08</td>
</tr>
<tr>
<td>Contact to STD</td>
<td>1.10</td>
<td>0.14</td>
<td>1.12</td>
</tr>
<tr>
<td>Cervicitis</td>
<td>0.71</td>
<td>0.14</td>
<td>0.70</td>
</tr>
<tr>
<td>PID</td>
<td>0.61</td>
<td>0.58</td>
<td>0.57</td>
</tr>
<tr>
<td>Gonorrhea</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-4.36</td>
<td>0.82</td>
<td>-4.02</td>
</tr>
<tr>
<td>Age</td>
<td>-0.02</td>
<td>0.03</td>
<td>-0.03</td>
</tr>
<tr>
<td>Contact to STD</td>
<td>1.78</td>
<td>0.42</td>
<td>1.68</td>
</tr>
<tr>
<td>Cervicitis</td>
<td>1.05</td>
<td>0.40</td>
<td>1.00</td>
</tr>
<tr>
<td>PID</td>
<td>1.77</td>
<td>0.78</td>
<td>1.66</td>
</tr>
</tbody>
</table>

are reasonably close to those from individual testing. The averaged standard errors obtained from applying the method in Section 4.2 to group testing data are in close agreement to the standard deviations from individual testing. From the results in Table 4.5, we can conclude that when individual gets one year older, the odds of being infected with chlamydia is approximately 8% less, while the odds of being infected with gonorrhea does not change. However, contacted with sexually transmitted disease increases the odds of being infected with chlamydia by roughly 3 times, and gonorrhea 6 times. Contracted with cervicitis increases the odds of chlamydia by roughly 2 times, and gonorrhea 2.9 times. Lastly, contracted with PID does not affect the odds of being infected with chlamydia, but increases the odds of gonorrhea by roughly 5.9 times.

4.5 Future work

In this project, motivated by the recent development of multiplex assays, we developed a semi-parametric group testing estimation framework to estimate individual-specific probability of infections. By incorporating a flexible correlation structure in our joint modeling approach, our method can produce interpretable inference for each
infection. Simulations and real data application show that the proposed method works reasonably well.

In Section 4.1, we make several assumptions regarding the testing process. Some discussion on relaxing them are given in Section 2.6 and Section 3.6. In addition, our method can be extended in several ways. First, we assume a parametric form on marginal probability of infection in Section 4.2. It is common to assume a logistic relationship in group testing estimation literature since the responses are usually binary when screening for infectious diseases. Doing so also significantly simplifies our derivations. A possible future work is to relax this assumption and treat the marginal relationship as unknown. One could simply puts a nonparametric form on the marginal probability, \( \eta_k(x^T_j \beta_k) \). Instead of just minimizing Equation 4.2 to get \( \hat{\eta}(\cdot, \cdot) \), one also minimize

\[
\sum_{i=1}^{I} \sum_{j=1}^{n} \left[ \delta_{ijk}^n - \eta(x^T_i \beta_k) \right]^2 K_{h_k}(X_{ij}^T \beta_k - x^T \beta_k)
\]

simultaneously, for \( k = 1, 2 \), to get \( \hat{\eta}_k(\cdot) \). This generalization is even more flexible than the proposed method, since it can incorporate any type of marginal relationship, as well as the correlation structure between infections.

Second, our method is designed specifically for the Iowa IPP algorithm, a two-stage hierarchical algorithm with the presence of two infections. In Chapter 2 and Chapter 3, we proposed alternative group testing algorithms and we demonstrated that the proposed algorithms are better than the current employed two-stage algorithm in terms of cost savings. Therefore, another possible extension of our estimation framework is to incorporate those testing algorithms in Chapter 2 and Chapter 3. The extension for \( S \)-stage hierarchical algorithms is straightforward since all the required components for the estimation framework can be easily obtained. For example, we have already developed an EM algorithm to estimate the population prevalence in an \( S \)-stage algorithm, and matrix \( W \) in Section 4.2 can be calculated through the
Markov Chain idea proposed in Chapter 2. The extension for array testing algorithms can require more work but conceptually easy.


C virus, and human immunodeficiency virus type-1 by nucleic acid amplification testing with specific and sensitive multiplex reagent in Japan”. In: *Journal of Virological Methods* 112, pp. 145–151.


Appendix A

Chapter 2 Supplementary Materials

A.1 Additional derivations from Sections 2.2 and 2.3

In Section 2.2, we let $\theta_{n,\tilde{z}_1\tilde{z}_2}$ denote the probability that a pool of size $n$ has true statuses $\tilde{z}_1 \in \{0, 1\}$ and $\tilde{z}_2 \in \{0, 1\}$. Recall that $\tilde{Y}_l = (\tilde{Y}_{l1}, \tilde{Y}_{l2})'$ are independent and identically distributed with probability mass function

$$
\text{pr}(\tilde{Y}_{l1} = \tilde{y}_1, \tilde{Y}_{l2} = \tilde{y}_2) = p_{00}(1-\tilde{y}_1)(1-\tilde{y}_2) + p_{10}(1-\tilde{y}_1)\tilde{y}_2 + p_{01}\tilde{y}_1(1-\tilde{y}_2) + p_{11}\tilde{y}_1\tilde{y}_2,
$$

for $\tilde{y}_1, \tilde{y}_2 \in \{0, 1\}$, where $p_{00} + p_{10} + p_{01} + p_{11} = 1$. Clearly,

$$
\theta_{n,00} = \prod_{l=1}^{n} \text{pr}(\tilde{Y}_{l1} = 0, \tilde{Y}_{l2} = 0) = p_{00}^n.
$$

From the Law of Total Probability, we have

$$
\theta_{n,10} = \prod_{l=1}^{n} \text{pr}(\tilde{Y}_{l2} = 0) - \prod_{l=1}^{n} \text{pr}(\tilde{Y}_{l1} = 0, \tilde{Y}_{l2} = 0) = (p_{00} + p_{10})^n - p_{00}^n.
$$

Similarly, $\theta_{n,01} = (p_{00} + p_{01})^n - p_{00}^n$. With the three probabilities determined above, $\theta_{n,11} = 1 - \theta_{n,10} - \theta_{n,01} - \theta_{n,00}$.

We now establish that $E(T_{s+1}) = (n_1/n_{s+1})\text{pr}(Z_{s,i1}+Z_{s,i2} > 0)$, for $s = 1, 2, ..., S-1$, as stated in the beginning of Section 2.3. Clearly, $T_{s+1}$, the number of tests expended in stage $s+1$, is a discrete random variable. It takes the value 0 if all pools in stage $s$ either test negatively or were not created to begin with. It takes the value $n_s/n_{s+1}$ if exactly one pool in stage $s$ tests positively. It takes the value $2n_s/n_{s+1}$ if exactly two pools in stage $s$ test positively, and so on. It takes the value $(n_1/n_s)n_s/n_{s+1}$ if every pool in stage $s$ tests positively. For notational simplicity, let
\[ \delta = \Pr(Z_{s,1} + Z_{s,2} > 0) \], the probability a pool in stage \( s \) tests positively. Therefore, \( 1 - \delta \) is the probability a pool in stage \( s \) tests negatively or was never created. The probability mass function of \( T_{s+1} \) is depicted in the table below:

<table>
<thead>
<tr>
<th>( t )</th>
<th>0</th>
<th>( \frac{n_1}{n_s} )</th>
<th>( 2 \frac{n_1}{n_s} )</th>
<th>( \cdots )</th>
<th>( \frac{n_1}{n_s} \times \frac{n_s}{n_{s+1}} )</th>
</tr>
</thead>
</table>
| \( \Pr(T_{s+1} = t) \) | \( (1 - \delta)^{n_s} \) | \( \frac{n_1}{n_s} \delta (1 - \delta)^{n_s-1} \) | \( \frac{n_1}{n_s} \delta^2 (1 - \delta)^{n_s-2} \) | \( \cdots \) | \( \frac{n_1}{n_s} \delta^{n_s-1} \)

Therefore, the expected value of \( T_{s+1} \) is given by

\[
E(T_{s+1}) = \sum_{\text{all } t} t \times \Pr(T_{s+1} = t),
\]

that is,

\[
E(T_{s+1}) = \frac{n_s}{n_{s+1}} \left[ \frac{n_1}{1} \delta (1 - \delta)^{n_s-1} + \frac{n_1}{2} \delta^2 (1 - \delta)^{n_s-2} + \cdots + \frac{n_1}{n_s} \delta^{n_s-1} \right]
\]

\[
= \frac{n_s}{n_{s+1}} \frac{n_1}{n_s} \times \delta \left[ (1 - \delta)^{n_s-1} + \frac{n_1}{1} \delta (1 - \delta)^{n_s-2} + \cdots + \frac{n_1}{n_s} \delta^{n_s-1} \right].
\]

The quantity within the square brackets above is the binomial expansion of \( [(1 - \delta) + \delta]^{n_1/n_s-1} \). Hence, \( E(T_{s+1}) = (n_1/n_{s+1})\delta = (n_1/n_{s+1})\Pr(Z_{s,1} + Z_{s,2} > 0) \), as claimed. Note that this argument holds for all \( s = 1, 2, \ldots, S - 1 \).

We now show how to derive the transition probabilities \( \pi_{A \rightarrow B}^{(t)} \) described in Section 2.3. Recall that \( \pi_{A \rightarrow B}^{(t)} \) gives the probability that the parent pool \( G_{t+1,i}^{(t)} \) in stage \( t \) transitions from state \( A \) to state \( B \) for its subpool \( G_{t+1,i} \) in stage \( t + 1 \). Note that a subpool cannot contain truly positive individuals if its parent pool does not. Therefore, all elements above the diagonal in

\[
\pi^{(t)} = \begin{pmatrix}
\pi_{00 \rightarrow 00}^{(t)} & \pi_{00 \rightarrow 10}^{(t)} & \pi_{00 \rightarrow 01}^{(t)} & \pi_{00 \rightarrow 11}^{(t)} \\
\pi_{10 \rightarrow 00}^{(t)} & \pi_{10 \rightarrow 10}^{(t)} & \pi_{10 \rightarrow 01}^{(t)} & \pi_{10 \rightarrow 11}^{(t)} \\
\pi_{01 \rightarrow 00}^{(t)} & \pi_{01 \rightarrow 10}^{(t)} & \pi_{01 \rightarrow 01}^{(t)} & \pi_{01 \rightarrow 11}^{(t)} \\
\pi_{11 \rightarrow 00}^{(t)} & \pi_{11 \rightarrow 10}^{(t)} & \pi_{11 \rightarrow 01}^{(t)} & \pi_{11 \rightarrow 11}^{(t)}
\end{pmatrix}
\]

83
are equal to 0; similarly, \( \pi_{01 \rightarrow 10}^{(t)} = 0 \) and \( \pi_{00 \rightarrow 00}^{(t)} = 1 \). The remaining probabilities are all derived in the same way, so we illustrate with the derivation for \( \pi_{10 \rightarrow 00}^{(t)} \). We have

\[
\pi_{10 \rightarrow 00}^{(t)} = \frac{\Pr(Z_{t+1,i}^{(t)} = (0, 0)', Z_{t+1,i}^{(t)} = (1, 0)')}{\Pr(Z_{t+1,i}^{(t)} = (1, 0)')}.
\]

(A.1)

The denominator of the right-hand side of Equation (A.1) is the probability a pool of size \( n_t \) has true status “10,” which is \( \theta_{n_t,10} \). To calculate the probability in the numerator, note that if \( G_{t+1,i}^{(t)} \) has true status “10” and its subpool \( G_{t+1,i}^{(t)} \) has status “00,” then the true status of the set of \( n_t - n_t+1 \) individuals that are in \( G_{t+1,i}^{(t)} \) but not in \( G_{t+1,i}^{(t)} \) must be “10”. Because individual true statuses are mutually independent, the numerator is \( \theta_{n_t+1,00} \theta_{n_t-n_t+1,10} \). Therefore,

\[
\pi_{10 \rightarrow 00}^{(t)} = \frac{\theta_{n_t+1,00} \theta_{n_t-n_t+1,10}}{\theta_{n_t,10}},
\]

as stated in Section 2.3. Applying the same argument, the remaining non-zero transition probabilities in \( \pi^{(t)} \) are

\[
\begin{align*}
\pi_{10 \rightarrow 10}^{(t)} &= \frac{\theta_{n_t+1,10} (\theta_{n_t-n_t+1,10} + \theta_{n_t-n_t+1,00})}{\theta_{n_t,10}} \\
\pi_{01 \rightarrow 01}^{(t)} &= \frac{\theta_{n_t+1,01} (\theta_{n_t-n_t+1,01} + \theta_{n_t-n_t+1,00})}{\theta_{n_t,01}} \\
\pi_{11 \rightarrow 10}^{(t)} &= \frac{\theta_{n_t+1,10} (\theta_{n_t-n_t+1,01} + \theta_{n_t-n_t+1,11})}{\theta_{n_t,11}} \\
\pi_{01 \rightarrow 00}^{(t)} &= \frac{\theta_{n_t+1,00} (\theta_{n_t-n_t+1,01} + \theta_{n_t-n_t+1,00})}{\theta_{n_t,01}} \\
\pi_{11 \rightarrow 00}^{(t)} &= \frac{\theta_{n_t+1,00} (\theta_{n_t-n_t+1,11} + \theta_{n_t-n_t+1,00})}{\theta_{n_t,11}} \\
\pi_{11 \rightarrow 10}^{(t)} &= \frac{\theta_{n_t+1,10} (\theta_{n_t-n_t+1,10} + \theta_{n_t-n_t+1,11})}{\theta_{n_t,11}} \\
\pi_{11 \rightarrow 11}^{(t)} &= \frac{\theta_{n_t+1,11} (\theta_{n_t-n_t+1,10} + \theta_{n_t-n_t+1,11})}{\theta_{n_t,11}} \\
\end{align*}
\]

and \( \pi_{11 \rightarrow 11}^{(t)} = \theta_{n_t+1,11} / \theta_{n_t,11} \).
A.2 Calculating $E(T^{(S)})$ for $J = 3$ infections.

It is easy to see how one could generalize our hierarchical algorithms for use with three or more infections. For example, the infectious disease literature is replete with situations where individuals are simultaneously tested for HIV, HBV, and HCV; see, e.g., Stramer et al. (2011), O’Brien et al. (2012), and Schmidt et al. (2010). These references are cited at the end of Appendix A.2. We illustrate how to generalize our framework in Section 2.3 to $J = 3$ infections specifically; extensions to $J > 3$ would also be obvious (but more tedious).

The derivation when $J = 3$ is a natural extension of the derivation when there are $J = 2$ infections. Each pool now has $2^3 = 8$ possible true status. Let $\theta_{n_s,z_1z_2z_3}$ denote the probability that a pool of size $n_s$ has true statuses $z_1, z_2, z_3 \in \{0, 1\}$. The argument in Appendix A.1 can be extended to show

$$
\begin{align*}
\theta_{n_s,000} &= (p_{100} + p_{000})^n_s - p_{000}^n_s, \\
\theta_{n_s,010} &= (p_{010} + p_{000})^n_s - p_{000}^n_s, \\
\theta_{n_s,001} &= (p_{001} + p_{000})^n_s - p_{000}^n_s, \\
\theta_{n_s,110} &= (p_{000} + p_{100} + p_{010} + p_{110})^n_s - \theta_{n_s,000} - \theta_{n_s,100} - \theta_{n_s,010}, \\
\theta_{n_s,101} &= (p_{000} + p_{100} + p_{001} + p_{101})^n_s - \theta_{n_s,000} - \theta_{n_s,100} - \theta_{n_s,001}, \\
\theta_{n_s,011} &= (p_{000} + p_{010} + p_{001} + p_{011})^n_s - \theta_{n_s,000} - \theta_{n_s,010} - \theta_{n_s,001}, \\
\theta_{n_s,111} &= 1 - \theta_{n_s,000} - \theta_{n_s,100} - \theta_{n_s,010} - \theta_{n_s,001} - \theta_{n_s,110} - \theta_{n_s,101} - \theta_{n_s,011},
\end{align*}
$$

where $p_{y_1y_2y_3}$ is the joint probability for the true individual statuses. Our goal is to calculate

$$
E(T^{(S)}) = 1 + \sum_{s=1}^{S-1} \left( \frac{n_1}{n_{s+1}} \right) \text{pr}(Z_{s,i1} + Z_{s,i2} + Z_{s,i3} > 0), \quad (A.2)
$$

where $Z_{s,ij}$ is the testing response for the $j$th infection ($j = 1, 2, 3$) in the $i$th pool tested at stage $s$. The true status path of $G_{s,i}^{(1)}, G_{s,i}^{(2)}, ..., G_{s,i}$ can also be viewed as a Markov chain but with possible states in $\Omega = \{000, 100, 010, 001, 110, 101, 011, 111\}$. 
The rest of the argument is analogous to the $J = 2$ case presented in the paper, except that we have to alter $\pi^{(t)}$, $M$, and $P^{(s)}$. The transition matrix $\pi^{(t)}$ is now $8 \times 8$ and has the following form:

$$
\begin{pmatrix}
\pi^{(t)}_{000\to000} & \pi^{(t)}_{000\to100} & \pi^{(t)}_{000\to101} & \pi^{(t)}_{000\to110} & \pi^{(t)}_{000\to111} \\
\pi^{(t)}_{100\to000} & \pi^{(t)}_{100\to100} & \pi^{(t)}_{100\to101} & \pi^{(t)}_{100\to110} & \pi^{(t)}_{100\to111} \\
\pi^{(t)}_{010\to000} & \pi^{(t)}_{010\to100} & \pi^{(t)}_{010\to101} & \pi^{(t)}_{010\to110} & \pi^{(t)}_{010\to111} \\
\pi^{(t)}_{001\to000} & \pi^{(t)}_{001\to100} & \pi^{(t)}_{001\to101} & \pi^{(t)}_{001\to110} & \pi^{(t)}_{001\to111} \\
\pi^{(t)}_{110\to000} & \pi^{(t)}_{110\to100} & \pi^{(t)}_{110\to101} & \pi^{(t)}_{110\to110} & \pi^{(t)}_{110\to111} \\
\pi^{(t)}_{101\to000} & \pi^{(t)}_{101\to100} & \pi^{(t)}_{101\to101} & \pi^{(t)}_{101\to110} & \pi^{(t)}_{101\to111} \\
\pi^{(t)}_{011\to000} & \pi^{(t)}_{011\to100} & \pi^{(t)}_{011\to101} & \pi^{(t)}_{011\to110} & \pi^{(t)}_{011\to111} \\
\pi^{(t)}_{111\to000} & \pi^{(t)}_{111\to100} & \pi^{(t)}_{111\to101} & \pi^{(t)}_{111\to110} & \pi^{(t)}_{111\to111}
\end{pmatrix}
$$

Transition probabilities of the form $\pi^{(t)}_{A\to B}$ are calculated in the same way as in the $J = 2$ case described in Appendix A.1; for example, the entries in the last row of $\pi^{(t)}$ are

$$
\begin{align*}
\pi^{(t)}_{111\to000} &= \frac{\theta_{n_{t+1},000}\theta_{n_t-n_{t+1},111}}{\theta_{n_t,111}} \\
\pi^{(t)}_{111\to010} &= \frac{\theta_{n_{t+1},010}\sum_{\tilde{z}_2=0}^{1}\theta_{n_t-n_{t+1},1\tilde{z}_21}}{\theta_{n_t,111}} \\
\pi^{(t)}_{111\to110} &= \frac{\theta_{n_{t+1},110}\sum_{\tilde{z}_2,\tilde{z}_3=0}^{1}\theta_{n_t-n_{t+1},1\tilde{z}_2\tilde{z}_3}}{\theta_{n_t,111}} \\
\pi^{(t)}_{111\to011} &= \frac{\theta_{n_{t+1},011}\sum_{\tilde{z}_2,\tilde{z}_3=0}^{1}\theta_{n_t-n_{t+1},1\tilde{z}_2\tilde{z}_3}}{\theta_{n_t,111}}
\end{align*}
$$

and other entries are calculated similarly. Note that $\pi^{(t)}$ remains lower triangular. The initial state matrix is

$$
M = \text{diag}(\theta_{n_1,000}, \theta_{n_1,100}, \theta_{n_1,010}, \theta_{n_1,001}, \theta_{n_1,110}, \theta_{n_1,101}, \theta_{n_1,011}, \theta_{n_1,111}).
$$

The matrix operator $P^{(s)}$ that diagnoses both truly positive and truly negative
pools as positive for at least one infection in stage $s$ is

$$
P^{(s)} = \text{diag}(1 - S_{p:1}^{(s)} S_{p:2}^{(s)}), 1 - S_{e:1}^{(s)} S_{e:2}^{(s)}), 1 - S_{p:1}^{(s)} S_{p:2}^{(s)} S_{p:3}^{(s)}, 1 - S_{e:1}^{(s)} S_{e:2}^{(s)} S_{e:3}^{(s)}, 1 - S_{p:1}^{(s)} S_{p:2}^{(s)} S_{p:3}^{(s)} S_{p:4}^{(s)}),$$

for $s = 1, 2, ..., S - 1$. Joint probabilities for all paths where $G_{s,i}$ tests positively are collected in the entries of $D = MP^{(1)} \pi^{(1)} P^{(2)} \pi^{(2)} \cdots \pi^{(s-1)} P^{(s)}$. Equation (A.2) becomes

$$E(T^{(S)}) = 1 + \sum_{s=1}^{S-1} \binom{n_1}{n_{s+1}} 1_8^t MP^{(1)} \prod_{t=0}^{s-1} (\pi^{(t)} P^{(t+1)}) 1_8,$$

where we set $\pi^{(0)} = (P^{(1)})^{-1}$ and $1_8 = (1, 1, 1, 1, 1, 1, 1, 1)'.$

**References cited in Appendix A.2**


A.3 Classification accuracy probabilities for $J = 2$ infections.

In Section 2.3 of the paper, we derived $P_{s:1} = \text{pr}(Z_{S,i1} = 1|\tilde{Z}_{S,i1} = 1)$ for $S > 2$ and $J = 2$ infections. Derivations for $P_{s:2}$, $P_{p:1}$ and $P_{p:2}$ also exploit the Markov chain structure of $G_{S,i}^{(1)}$, $G_{S,i}^{(2)}$, ..., $G_{S,i}^{(S-1)}$ after removing one individual. For $P_{s:2}$, the individual removed ($G_{S,i}$) is truly positive for the second infection. The same argument presented in the paper applies and

$$P_{s:2} = \left(\frac{p_{01}}{p_{01} + p_{11}}\right) 1'_{4}M_{-1}P_{-+}^{(1)} + \sum_{t=1}^{S-2} \left(\frac{p_{11}}{p_{01} + p_{11}}\right) 1'_{4}M_{-1}P_{++}^{(1)} \prod_{t=1}^{S-2} (\pi_{-+}^{(t)}P_{++}^{(t+1)}) 1_{4}S_{e:2}^{(S)}$$

where $P_{-+}^{(s)} = \text{diag}(1-S_{e:2}^{(s)}, 1-S_{e:2}^{(s)}, 1-S_{e:2}^{(s)}, 1-S_{e:2}^{(s)}), (s = 1, 2, ..., S-1)$. The matrices $M_{-1,1}, \pi_{-1}^{(t)}$, and $P_{++}$ are defined in Section 2.3. To derive $P_{p:1} = 1 - \text{pr}(Z_{S,i1} = 1|\tilde{Z}_{S,i1} = 0)$, we use the same Markov chain argument to first calculate $\text{pr}(Z_{S,i1} = 1|\tilde{Z}_{S,i1} = 0)$. Note that if $\tilde{Z}_{S,i} = (0,0)'$, then the true status of $G_{S,i}^{(t)}$ is determined by $G_{S,i}^{(t)}$. Therefore, we use the matrix $P^{(s)}$ to augment $C_{-1} = M_{-1,1} \pi_{-1}^{(1)} \pi_{-1}^{(2)} \ldots \pi_{-1}^{(S-2)}$, where $P^{(s)}$ is defined in Section 2.3. Expressions for $P_{p:1}$ and $P_{p:2}$ are found by solving

$$1 - P_{p:1} = \left(\frac{p_{00}}{p_{01} + p_{00}}\right) 1'_{4}M_{-1}P_{-+}^{(1)} \prod_{t=1}^{S-2} (\pi_{-+}^{(t)}P_{++}^{(t+1)}) 1_{4}S_{e:1}^{(S)}$$

and

$$1 - P_{p:2} = \left(\frac{p_{00}}{p_{01} + p_{00}}\right) 1'_{4}M_{-1}P_{-+}^{(1)} \prod_{t=1}^{S-2} (\pi_{-+}^{(t)}P_{++}^{(t+1)}) 1_{4}S_{e:2}^{(S)}$$

respectively, where $P_{-+}^{(s)}$ is defined in Section 2.3 and where $P_{++}^{(s)}$ is defined above. Our formulas also apply in the $S = 2$ case; the only difference is that matrix products like $\prod_{t=1}^{S-2} (\pi_{-+}^{(t)}P_{++}^{(t+1)})$ are identity matrices.
A.4 Classification accuracy probabilities for \( J = 3 \) infections.

We now extend the derivation in Section 2.3 to derive \( \text{PS}_{e_1} \) when \( S > 2 \) and there are \( J = 3 \) infections. Recall that \( \bar{Z}_{-S,i}^{(1)}, \bar{Z}_{-S,i}^{(2)}, \ldots, \bar{Z}_{-S,i}^{(S-1)} \) denotes the true status path of \( G_{S,i}^{(1)}, G_{S,i}^{(2)}, \ldots, G_{S,i}^{(S-1)} \) after individual \( G_{S,i} \) is removed. The joint probability of the true status path of \( G_{S,i}^{(1)}, G_{S,i}^{(2)}, \ldots, G_{S,i} \), conditional on the event \( \{ \bar{Z}_{S,i} = 1 \} \), is found by calculating

\[
\text{pr}(\bar{Z}_{-S,i}^{(1)} = \bar{z}_1, \bar{Z}_{-S,i}^{(2)} = \bar{z}_2, \ldots, \bar{Z}_{-S,i}^{(S-1)} = \bar{z}_{S-1}, \bar{Z}_{S,i} = (1, \bar{z}_2, \bar{z}_3)|\bar{Z}_{S,i} = 1) = \text{pr}(\bar{Z}_{S,i} = (1, \bar{z}_2, \bar{z}_3)|\bar{Z}_{S,i} = 1) \text{ pr}(\bar{Z}_{-S,i}^{(1)} = \bar{z}_1, \bar{Z}_{-S,i}^{(2)} = \bar{z}_2, \ldots, \bar{Z}_{-S,i}^{(S-1)} = \bar{z}_{S-1}),
\]

(A.3)

where

\( \bar{z}_1, \bar{z}_2, \ldots, \bar{z}_{S-1} \in \{(0,0,0), (1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1), (1,1,1)\} \)

and \( \bar{z}_2, \bar{z}_3 \in \{0,1\} \). The first probability on the right-hand side of Equation (A.3) is equal to \( p_{1\bar{z}_2\bar{z}_3}/(p_{100} + p_{110} + p_{101} + p_{111}) \). The second probability is collected in the entries of \( C_{-1} = M_{-1}\pi_{-1}^{(1)}\pi_{-1}^{(2)} \cdots \pi_{-1}^{(S-2)} \). The matrices \( M_{-1} \) and \( \pi^{(t)} \) take the same form as \( M \) and \( \pi^{(t)} \) (defined in Appendix A.2), respectively, except that pool sizes are reduced by one. To incorporate misclassification, note that all the pools in the sequence \( G_{S,i}^{(1)}, G_{S,i}^{(2)}, \ldots, G_{S,i} \) are truly positive for first infection because of the given event \( \{ \bar{Z}_{S,i} = 1 \} \). The values of \( \bar{z}_2 \) and \( \bar{z}_3 \in \{0,1\} \) must be treated separately. If \( \bar{z}_j = 1 \), for \( j = 2,3 \), so are \( \bar{Z}_{S,i}^{(t)} \), for \( t = 1,2,\ldots,S-1 \); otherwise, the values of \( \bar{Z}_{S,i}^{(t)} \) are determined by \( \bar{Z}_{-S,i}^{(t)} \). To cover all possible combinations of \( (\bar{z}_2, \bar{z}_3) \in \{(0,0),(1,0),(0,1),(1,1)\} \), we define the four matrix operators:

\[
P^{(s)}_{++} = \text{diag}(1 - S_{e_1}S_{p_2}S_{p_3}, 1 - S_{e_1}S_{p_2}S_{e_3}, 1 - S_{e_1}S_{p_2}S_{p_3}, 1 - S_{e_1}S_{e_2}S_{p_3}, 1 - S_{e_1}S_{e_2}S_{e_3}, 1 - S_{e_1}S_{e_2}S_{e_3}),
\]
\[
P^{(s)}_{+-} = \text{diag}(1 - S_{e_1}S_{p_2}S_{p_3}, 1 - S_{e_1}S_{p_2}S_{e_3}, 1 - S_{e_1}S_{p_2}S_{p_3}, 1 - S_{e_1}S_{e_2}S_{p_3}, 1 - S_{e_1}S_{e_2}S_{e_3}, 1 - S_{e_1}S_{e_2}S_{e_3}),
\]
\[
P^{(s)}_{-+} = \text{diag}(1 - S_{e_1}S_{e_2}S_{p_3}, 1 - S_{e_1}S_{e_2}S_{e_3}, 1 - S_{e_1}S_{e_2}S_{p_3}, 1 - S_{e_1}S_{e_2}S_{e_3}, 1 - S_{e_1}S_{e_2}S_{e_3}, 1 - S_{e_1}S_{e_2}S_{e_3}),
\]
\[
P^{(s)}_{--} = \text{diag}(1 - S_{e_1}S_{e_2}S_{p_3}, 1 - S_{e_1}S_{e_2}S_{e_3}, 1 - S_{e_1}S_{e_2}S_{p_3}, 1 - S_{e_1}S_{e_2}S_{e_3}, 1 - S_{e_1}S_{e_2}S_{e_3}, 1 - S_{e_1}S_{e_2}S_{e_3}),
\]

89
\[ P^{(s)}_{++} = \text{diag}(1 - S^{(s)}_{e1}S^{(s)}_{e2}S^{(s)}_{e3}) 1 - S^{(s)}_{e1}S^{(s)}_{e2}S^{(s)}_{e3} \]

Similarly, for the second infection,

\[ P^{(s)}_{++} = (1 - S^{(s)}_{e1}S^{(s)}_{e2}S^{(s)}_{e3}) I_8, \]

for \( s = 1, 2, \ldots, S - 1 \), where \( I_8 \) is the \( 8 \times 8 \) identity matrix. Following a similar argument to the \( J = 2 \) infection case in Section 2.3, we obtain

\[
\begin{align*}
PS_{e1} &= \left( \frac{p_{100}}{p_{100} + p_{110} + p_{010} + p_{111}} \right) 1_8' M_{-1} P^{(1)}_{++} \prod_{t=1}^{S-2} (\pi^{(t)} P^{(t+1)}_{-+}) I_8 S^{(S)}_{e1} \\
&+ \left( \frac{p_{110}}{p_{100} + p_{110} + p_{010} + p_{111}} \right) 1_8' M_{-1} P^{(1)}_{++} \prod_{t=1}^{S-2} (\pi^{(t)} P^{(t+1)}_{-+}) I_8 S^{(S)}_{e1} \\
&+ \left( \frac{p_{101}}{p_{100} + p_{110} + p_{010} + p_{111}} \right) 1_8' M_{-1} P^{(1)}_{++} \prod_{t=1}^{S-2} (\pi^{(t)} P^{(t+1)}_{-+}) I_8 S^{(S)}_{e1} \\
&+ \left( \frac{p_{111}}{p_{100} + p_{110} + p_{010} + p_{111}} \right) 1_8' M_{-1} P^{(1)}_{++} \prod_{t=1}^{S-2} (\pi^{(t)} P^{(t+1)}_{-+}) I_8 S^{(S)}_{e1}.
\end{align*}
\]

Similarly, for the second infection,

\[
\begin{align*}
PS_{e2} &= \left( \frac{p_{010}}{p_{010} + p_{110} + p_{011} + p_{111}} \right) 1_8' M_{-1} P^{(1)}_{++} \prod_{t=1}^{S-2} (\pi^{(t)} P^{(t+1)}_{-+}) I_8 S^{(S)}_{e2} \\
&+ \left( \frac{p_{110}}{p_{010} + p_{110} + p_{011} + p_{111}} \right) 1_8' M_{-1} P^{(1)}_{++} \prod_{t=1}^{S-2} (\pi^{(t)} P^{(t+1)}_{-+}) I_8 S^{(S)}_{e2} \\
&+ \left( \frac{p_{101}}{p_{010} + p_{110} + p_{011} + p_{111}} \right) 1_8' M_{-1} P^{(1)}_{++} \prod_{t=1}^{S-2} (\pi^{(t)} P^{(t+1)}_{-+}) I_8 S^{(S)}_{e2} \\
&+ \left( \frac{p_{111}}{p_{010} + p_{110} + p_{011} + p_{111}} \right) 1_8' M_{-1} P^{(1)}_{++} \prod_{t=1}^{S-2} (\pi^{(t)} P^{(t+1)}_{-+}) I_8 S^{(S)}_{e2},
\end{align*}
\]

where \( P^{(s)}_{++} \) and \( P^{(s)}_{++} \) are defined above, and, for \( s = 1, 2, \ldots, S - 1 \), the matrix operators

\[
P^{(s)}_{++} = \text{diag}(1 - S^{(s)}_{e1}S^{(s)}_{e2}S^{(s)}_{e3}),
\]

and

\[
P^{(s)}_{+++} = (1 - S^{(s)}_{e1}S^{(s)}_{e2}S^{(s)}_{e3}) I_8.
\]

90
\[ \mathbf{P}^{(s)}_{--} = \text{diag}(1 - S_{p:1}^{(s)} S_{e:2}^{(s)} S_{e:3}^{(s)}, 1 - S_{e:1}^{(s)} S_{e:2}^{(s)} S_{e:3}^{(s)}, 1 - S_{p:1}^{(s)} S_{e:2}^{(s)} S_{e:3}^{(s)}, 1 - S_{e:1}^{(s)} S_{e:2}^{(s)} S_{e:3}^{(s)}). \]

For the third infection,
\[
\begin{align*}
\text{PS}_{e:3} &= \left( \frac{p_{001}}{p_{001} + p_{101} + p_{011} + p_{111}} \right) 1_{8}^{t} \mathbf{M}_{-1}^{(1)} 1_{8}^{(t+1)} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} \mathbf{P}^{(t+1)}_{-+-} \mathbf{1}_{8} \mathbf{S}_{e:3}^{(S)}) \\
&+ \left( \frac{p_{101}}{p_{001} + p_{101} + p_{011} + p_{111}} \right) 1_{8}^{t} \mathbf{M}_{-1}^{(1)} 1_{8}^{(t+1)} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} \mathbf{P}^{(t+1)}_{+-+} \mathbf{1}_{8} \mathbf{S}_{e:3}^{(S)}) \\
&+ \left( \frac{p_{011}}{p_{001} + p_{101} + p_{011} + p_{111}} \right) 1_{8}^{t} \mathbf{M}_{-1}^{(1)} 1_{8}^{(t+1)} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} \mathbf{P}^{(t+1)}_{++-} \mathbf{1}_{8} \mathbf{S}_{e:3}^{(S)}),
\end{align*}
\]

where \( \mathbf{P}^{(s)}_{++-}, \mathbf{P}^{(s)}_{-+-}, \) and \( \mathbf{P}^{(s)}_{+++} \) are defined above, and where
\[
\begin{align*}
\mathbf{P}^{(s)}_{---} &= \text{diag}(1 - S_{p:1}^{(s)} S_{p:2}^{(s)} S_{e:3}^{(s)}, 1 - S_{e:1}^{(s)} S_{e:2}^{(s)} S_{e:3}^{(s)}, 1 - S_{p:1}^{(s)} S_{e:2}^{(s)} S_{e:3}^{(s)}, 1 - S_{e:1}^{(s)} S_{e:2}^{(s)} S_{e:3}^{(s)}),
\end{align*}
\]

for \( s = 1, 2, ..., S-1 \). Derivations for \( \text{PS}_{p:1}, \text{PS}_{p:2} \), and \( \text{PS}_{p:3} \) follow the same arguments as those for the \( J = 2 \) infection case. They are found by solving
\[
\begin{align*}
1 - \text{PS}_{p:1} &= \left( \frac{p_{000}}{p_{000} + p_{100} + p_{001} + p_{111}} \right) 1_{8}^{t} \mathbf{M}_{-1}^{(1)} 1_{8}^{(t+1)} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} \mathbf{P}^{(t+1)}_{-+-} \mathbf{1}_{8} \mathbf{S}_{p:1}^{(S)}) \\
&+ \left( \frac{p_{100}}{p_{000} + p_{100} + p_{001} + p_{111}} \right) 1_{8}^{t} \mathbf{M}_{-1}^{(1)} 1_{8}^{(t+1)} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} \mathbf{P}^{(t+1)}_{+-+} \mathbf{1}_{8} \mathbf{S}_{p:1}^{(S)}) \\
&+ \left( \frac{p_{001}}{p_{000} + p_{100} + p_{001} + p_{111}} \right) 1_{8}^{t} \mathbf{M}_{-1}^{(1)} 1_{8}^{(t+1)} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} \mathbf{P}^{(t+1)}_{++-} \mathbf{1}_{8} \mathbf{S}_{p:1}^{(S)}),
\end{align*}
\]

as those for the \( J = 2 \) infection case. They are found by solving
1 − PS_{p:2} = \left( \frac{p_{000}}{p_{000} + p_{100} + p_{001} + p_{101}} \right) 1_{8}^{\prime} M_{-1} P^{(1)} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} P^{(t+1)}) 1_{8} S_{p:2}^{(S)} + \left( \frac{p_{100}}{p_{000} + p_{100} + p_{001} + p_{101}} \right) 1_{8}^{\prime} M_{-1} P^{(1)}_{++-} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} P^{(t+1)}_{+-}) 1_{8} S_{p:2}^{(S)} + \left( \frac{p_{001}}{p_{000} + p_{100} + p_{001} + p_{101}} \right) 1_{8}^{\prime} M_{-1} P^{(1)}_{-+-} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} P^{(t+1)}_{-+-}) 1_{8} S_{p:2}^{(S)} + \left( \frac{p_{101}}{p_{000} + p_{100} + p_{001} + p_{101}} \right) 1_{8}^{\prime} M_{-1} P^{(1)}_{--+} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} P^{(t+1)}_{+-+}) 1_{8} S_{p:2}^{(S)},

and

1 − PS_{p:3} = \left( \frac{p_{000}}{p_{000} + p_{100} + p_{010} + p_{110}} \right) 1_{8}^{\prime} M_{-1} P^{(1)} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} P^{(t+1)}) 1_{8} S_{p:3}^{(S)} + \left( \frac{p_{100}}{p_{000} + p_{100} + p_{010} + p_{110}} \right) 1_{8}^{\prime} M_{-1} P^{(1)}_{+-+} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} P^{(t+1)}_{+++}) 1_{8} S_{p:3}^{(S)} + \left( \frac{p_{010}}{p_{000} + p_{100} + p_{010} + p_{110}} \right) 1_{8}^{\prime} M_{-1} P^{(1)}_{---} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} P^{(t+1)}_{---}) 1_{8} S_{p:3}^{(S)} + \left( \frac{p_{110}}{p_{000} + p_{100} + p_{010} + p_{110}} \right) 1_{8}^{\prime} M_{-1} P^{(1)}_{-++} \prod_{t=1}^{S-2} (\pi_{-1}^{(t)} P^{(t+1)}_{-++}) 1_{8} S_{p:3}^{(S)}.

The matrix $P^{(s)}$ is defined in Appendix A.2 and the other matrix operators are defined above.
A.5 Additional results from Section 2.4.

Here is an outline of the material in this Appendix:

**Page 94:** Table A.1. Expected number of tests per individual and optimal configurations when $S_{e_j}^{(s)}$ is a decreasing function of $n_s$. This is a modified version of Table 2.1 and Table 2.2.

**Page 95:** Figure A.1. Simulation study for the first and second cases in Table 2.1 and Table 2.2. See the discussion in Section 2.4.

**Page 96:** Figure A.2. Simulation study for the fourth and fifth cases in Table 2.1. See the discussion in Section 2.4 and Table 2.2.

**Page 97:** Table A.2. Classification accuracy measures when $S_{e_j}^{(s)} = 0.95$ and $S_{p_j}^{(s)} = 0.99$. These are the classification accuracy measures associated with the configurations in Table 2.1.

**Page 98:** Table A.3. Classification accuracy measures when $S_{e_j}^{(s)} = 0.95$ and $S_{p_j}^{(s)} = 0.99$. These are the classification accuracy measures associated with the configurations in Table 2.2.

**Pages 99-100:** Derivation of $E(C^{(S)})$, the expected number of individuals correctly classified in a master pool tested in $S$ stages.

**Page 101:** Table A.4. Expected number of tests per individual when $S_{e_j}^{(s)} = 0.95$, $S_{p_j}^{(s)} = 0.99$, optimal configurations determined from maximizing $E(C^{(S)})/E(T^{(S)})$, the objective function recommended by Malinovsky, Albert, and Roy (2016). This is a modified version of Table 2.1 and Table 2.2. See the remark below.

**Page 102:** Figure A.3. Optimal number of stages $S$ when $S_{e_j}^{(s)} = 0.95$, $S_{p_j}^{(s)} = 0.99$, optimal configurations determined from maximizing $E(C^{(S)})/E(T^{(S)})$. This is a modified version of Figure 2.2.
Table A.1: Expected number of tests per individual $n_1^{-1}E(T(S))$ when $S_{i,j}^{(s)} = (0.95)^{(i+0.5)(s-j)}$, $s_{p,j}^{(s)} = 0.99$, and $S \in \{2, 3, 4, 5, 6\}$. The column labeled “Optimal” gives the configuration of $n_1, n_2, ..., n_S$ that minimizes $n_1^{-1}E(T(S))$. The column labeled “Halving” gives the master pool size for the optimal halving algorithm. The percent reduction in $n_1^{-1}E(T(S))$ when compared to $n_1^{-1}E(T(2))$ is provided. The expected proportion of correct classifications $n_1^{-1}E(C(S))$ is also shown; see the discussion at the end of Section 2.4. The maximum allowable master pool size is 100.

<table>
<thead>
<tr>
<th>S</th>
<th>Optimal</th>
<th>$n_1^{-1}E(T(S))$</th>
<th>% Reduction</th>
<th>$n_1^{-1}E(C(S))$</th>
<th>Halving</th>
<th>$n_1^{-1}E(T(S))$</th>
<th>% Reduction</th>
<th>$n_1^{-1}E(C(S))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4 : 1</td>
<td>0.586</td>
<td>--</td>
<td>0.983</td>
<td>$n_1 = 4$</td>
<td>0.586</td>
<td>--</td>
<td>0.983</td>
</tr>
<tr>
<td>3</td>
<td>9 : 3 : 1</td>
<td>0.548</td>
<td>6.5</td>
<td>0.980</td>
<td>$n_1 = 6$</td>
<td>0.554</td>
<td>5.5</td>
<td>0.979</td>
</tr>
<tr>
<td>4</td>
<td>99 : 9 : 3 : 1</td>
<td>0.550</td>
<td>6.1</td>
<td>0.979</td>
<td>$n_1 = 12$</td>
<td>0.553</td>
<td>5.6</td>
<td>0.974</td>
</tr>
<tr>
<td>5</td>
<td>96 : 12 : 6 : 3 : 1</td>
<td>0.552</td>
<td>5.8</td>
<td>0.972</td>
<td>$n_1 = 24$</td>
<td>0.559</td>
<td>4.6</td>
<td>0.970</td>
</tr>
<tr>
<td>6</td>
<td>96 : 48 : 12 : 6 : 3 : 1</td>
<td>0.552</td>
<td>5.8</td>
<td>0.970</td>
<td>$n_1 = 48$</td>
<td>0.559</td>
<td>4.6</td>
<td>0.968</td>
</tr>
<tr>
<td>2</td>
<td>5 : 1</td>
<td>0.429</td>
<td>--</td>
<td>0.990</td>
<td>$n_1 = 5$</td>
<td>0.429</td>
<td>--</td>
<td>0.990</td>
</tr>
<tr>
<td>3</td>
<td>9 : 3 : 1</td>
<td>0.359</td>
<td>16.3</td>
<td>0.989</td>
<td>$n_1 = 8$</td>
<td>0.372</td>
<td>13.3</td>
<td>0.989</td>
</tr>
<tr>
<td>4</td>
<td>18 : 6 : 3 : 1</td>
<td>0.346</td>
<td>19.3</td>
<td>0.987</td>
<td>$n_1 = 12$</td>
<td>0.349</td>
<td>18.6</td>
<td>0.986</td>
</tr>
<tr>
<td>5</td>
<td>24 : 12 : 6 : 3 : 1</td>
<td>0.340</td>
<td>20.7</td>
<td>0.984</td>
<td>$n_1 = 24$</td>
<td>0.340</td>
<td>20.7</td>
<td>0.984</td>
</tr>
<tr>
<td>6</td>
<td>32 : 16 : 8 : 4 : 2 : 1</td>
<td>0.336</td>
<td>21.7</td>
<td>0.981</td>
<td>$n_1 = 32$</td>
<td>0.336</td>
<td>21.7</td>
<td>0.981</td>
</tr>
<tr>
<td>2</td>
<td>7 : 1</td>
<td>0.342</td>
<td>--</td>
<td>0.993</td>
<td>$n_1 = 7$</td>
<td>0.342</td>
<td>--</td>
<td>0.993</td>
</tr>
<tr>
<td>3</td>
<td>16 : 4 : 1</td>
<td>0.265</td>
<td>22.5</td>
<td>0.993</td>
<td>$n_1 = 10$</td>
<td>0.281</td>
<td>17.8</td>
<td>0.993</td>
</tr>
<tr>
<td>4</td>
<td>27 : 9 : 3 : 1</td>
<td>0.245</td>
<td>28.4</td>
<td>0.993</td>
<td>$n_1 = 16$</td>
<td>0.254</td>
<td>25.7</td>
<td>0.992</td>
</tr>
<tr>
<td>5</td>
<td>36 : 12 : 6 : 3 : 1</td>
<td>0.241</td>
<td>29.5</td>
<td>0.991</td>
<td>$n_1 = 24$</td>
<td>0.241</td>
<td>29.5</td>
<td>0.991</td>
</tr>
<tr>
<td>6</td>
<td>32 : 16 : 8 : 4 : 2 : 1</td>
<td>0.238</td>
<td>30.4</td>
<td>0.991</td>
<td>$n_1 = 32$</td>
<td>0.238</td>
<td>30.4</td>
<td>0.989</td>
</tr>
<tr>
<td>2</td>
<td>11 : 1</td>
<td>0.207</td>
<td>--</td>
<td>0.997</td>
<td>$n_1 = 11$</td>
<td>0.207</td>
<td>--</td>
<td>0.997</td>
</tr>
<tr>
<td>3</td>
<td>25 : 5 : 1</td>
<td>0.131</td>
<td>36.7</td>
<td>0.997</td>
<td>$n_1 = 16$</td>
<td>0.151</td>
<td>27.1</td>
<td>0.997</td>
</tr>
<tr>
<td>4</td>
<td>48 : 12 : 4 : 1</td>
<td>0.110</td>
<td>46.9</td>
<td>0.997</td>
<td>$n_1 = 24$</td>
<td>0.123</td>
<td>40.6</td>
<td>0.997</td>
</tr>
<tr>
<td>5</td>
<td>81 : 27 : 9 : 3 : 1</td>
<td>0.100</td>
<td>51.7</td>
<td>0.997</td>
<td>$n_1 = 40$</td>
<td>0.107</td>
<td>48.3</td>
<td>0.996</td>
</tr>
<tr>
<td>6</td>
<td>72 : 36 : 18 : 9 : 3 : 1</td>
<td>0.095</td>
<td>54.1</td>
<td>0.996</td>
<td>$n_1 = 64$</td>
<td>0.096</td>
<td>53.6</td>
<td>0.996</td>
</tr>
<tr>
<td>2</td>
<td>33 : 1</td>
<td>0.080</td>
<td>--</td>
<td>0.999</td>
<td>$n_1 = 33$</td>
<td>0.080</td>
<td>--</td>
<td>0.999</td>
</tr>
<tr>
<td>3</td>
<td>99 : 11 : 1</td>
<td>0.031</td>
<td>61.3</td>
<td>1.000</td>
<td>$n_1 = 50$</td>
<td>0.045</td>
<td>43.8</td>
<td>0.999</td>
</tr>
<tr>
<td>4</td>
<td>96 : 24 : 6 : 1</td>
<td>0.023</td>
<td>71.3</td>
<td>1.000</td>
<td>$n_1 = 72$</td>
<td>0.032</td>
<td>60.0</td>
<td>0.999</td>
</tr>
<tr>
<td>5</td>
<td>96 : 48 : 16 : 4 : 1</td>
<td>0.021</td>
<td>73.8</td>
<td>1.000</td>
<td>$n_1 = 96$</td>
<td>0.024</td>
<td>70.0</td>
<td>0.999</td>
</tr>
<tr>
<td>6</td>
<td>96 : 48 : 24 : 12 : 4 : 1</td>
<td>0.020</td>
<td>75.0</td>
<td>1.000</td>
<td>$n_1 = 96$</td>
<td>0.021</td>
<td>73.8</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Figure A.1: Simulation study for the first and second cases in Table 2.1 and Table 2.2 with $S_{i;j}^{(s)} = 0.95$ and $S_{p;j}^{(s)} = 0.99$. Boxplots of the number of tests per individual are constructed from $B = 5000$ replications under the optimal and halving group configurations shown in Table 2.1 and Table 2.2. Top left: First case in Table 2.1 with $p_{00} = 0.90$. Top right: First case in Table 2.2 with $p_{00} = 0.90$. Bottom left: Second case in Table 2.1 with $p_{00} = 0.95$. Bottom right: Second case in Table 2.2 with $p_{00} = 0.95$. 
Figure A.2: Simulation study for the fourth and fifth cases in Table 2.1 and Table 2.2. Boxplots of the number of tests per individual are constructed from $B = 5000$ replications under the optimal and halving group configurations shown in Table 2.1 and Table 2.2. Top left: Fourth case in Table 2.1 with $p_{00} = 0.99$. Top right: Fourth case in Table 2.2 with $p_{00} = 0.99$. Bottom left: Fifth case in Table 2.1 with $p_{00} = 0.999$. Bottom right: Fifth case in Table 2.2 with $p_{00} = 0.999$. 

96
Table A.2: Classification accuracy measures when $S_{e,j} = 0.95$, $S_{p,j} = 0.99$, and number of stages $S \in \{2, 3, 4, 5, 6\}$. These are calculated from our expressions in Section 2.3 and Appendix A.3. The optimal algorithm is the same as that identified in Table 2.1.

<table>
<thead>
<tr>
<th>$p_{00}$</th>
<th>$S$</th>
<th>Optimal</th>
<th>$PS_{e1}$</th>
<th>$PS_{e2}$</th>
<th>$PS_{p1}$</th>
<th>$PS_{p2}$</th>
<th>$PPV_1$</th>
<th>$PPV_2$</th>
<th>$NPV_1$</th>
<th>$NPV_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>2</td>
<td>4 : 1</td>
<td>0.916</td>
<td>0.918</td>
<td>0.997</td>
<td>0.997</td>
<td>0.951</td>
<td>0.940</td>
<td>0.995</td>
<td>0.996</td>
</tr>
<tr>
<td>0.05</td>
<td>3</td>
<td>9 : 3 : 1</td>
<td>0.888</td>
<td>0.893</td>
<td>0.998</td>
<td>0.998</td>
<td>0.963</td>
<td>0.955</td>
<td>0.993</td>
<td>0.994</td>
</tr>
<tr>
<td>0.04</td>
<td>4</td>
<td>99 : 9 : 3 : 1</td>
<td>0.886</td>
<td>0.891</td>
<td>0.998</td>
<td>0.998</td>
<td>0.963</td>
<td>0.955</td>
<td>0.993</td>
<td>0.994</td>
</tr>
<tr>
<td>0.01</td>
<td>5</td>
<td>90 : 45 : 9 : 3 : 1</td>
<td>0.880</td>
<td>0.886</td>
<td>0.998</td>
<td>0.998</td>
<td>0.963</td>
<td>0.955</td>
<td>0.992</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>96 : 48 : 24 : 6 : 3 : 1</td>
<td>0.864</td>
<td>0.873</td>
<td>0.998</td>
<td>0.998</td>
<td>0.964</td>
<td>0.956</td>
<td>0.991</td>
<td>0.993</td>
</tr>
<tr>
<td>0.95</td>
<td>2</td>
<td>5 : 1</td>
<td>0.917</td>
<td>0.929</td>
<td>0.998</td>
<td>0.998</td>
<td>0.950</td>
<td>0.897</td>
<td>0.997</td>
<td>0.999</td>
</tr>
<tr>
<td>0.03</td>
<td>3</td>
<td>9 : 3 : 1</td>
<td>0.886</td>
<td>0.909</td>
<td>0.999</td>
<td>0.999</td>
<td>0.973</td>
<td>0.939</td>
<td>0.995</td>
<td>0.998</td>
</tr>
<tr>
<td>0.01</td>
<td>4</td>
<td>18 : 6 : 3 : 1</td>
<td>0.861</td>
<td>0.895</td>
<td>0.999</td>
<td>0.999</td>
<td>0.973</td>
<td>0.941</td>
<td>0.994</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>90 : 18 : 6 : 3 : 1</td>
<td>0.854</td>
<td>0.892</td>
<td>0.999</td>
<td>0.999</td>
<td>0.973</td>
<td>0.941</td>
<td>0.994</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>96 : 48 : 24 : 6 : 3 : 1</td>
<td>0.839</td>
<td>0.885</td>
<td>0.999</td>
<td>0.999</td>
<td>0.973</td>
<td>0.942</td>
<td>0.993</td>
<td>0.998</td>
</tr>
<tr>
<td>0.97</td>
<td>2</td>
<td>7 : 1</td>
<td>0.928</td>
<td>0.928</td>
<td>0.998</td>
<td>0.998</td>
<td>0.910</td>
<td>0.910</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>0.01</td>
<td>3</td>
<td>16 : 4 : 1</td>
<td>0.909</td>
<td>0.909</td>
<td>0.999</td>
<td>0.999</td>
<td>0.951</td>
<td>0.951</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>0.01</td>
<td>4</td>
<td>27 : 9 : 3 : 1</td>
<td>0.892</td>
<td>0.892</td>
<td>0.999</td>
<td>0.999</td>
<td>0.965</td>
<td>0.965</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>36 : 12 : 6 : 3 : 1</td>
<td>0.875</td>
<td>0.875</td>
<td>0.999</td>
<td>0.999</td>
<td>0.966</td>
<td>0.966</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>96 : 24 : 12 : 6 : 3 : 1</td>
<td>0.867</td>
<td>0.867</td>
<td>0.999</td>
<td>0.999</td>
<td>0.966</td>
<td>0.966</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td>0.99</td>
<td>2</td>
<td>11 : 1</td>
<td>0.920</td>
<td>0.920</td>
<td>0.999</td>
<td>0.999</td>
<td>0.831</td>
<td>0.831</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.004</td>
<td>3</td>
<td>25 : 5 : 1</td>
<td>0.892</td>
<td>0.892</td>
<td>1.000</td>
<td>1.000</td>
<td>0.925</td>
<td>0.925</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>0.004</td>
<td>4</td>
<td>48 : 12 : 4 : 1</td>
<td>0.867</td>
<td>0.867</td>
<td>1.000</td>
<td>1.000</td>
<td>0.943</td>
<td>0.943</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>0.002</td>
<td>5</td>
<td>81 : 27 : 9 : 3 : 1</td>
<td>0.846</td>
<td>0.846</td>
<td>1.000</td>
<td>1.000</td>
<td>0.959</td>
<td>0.959</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>72 : 24 : 12 : 6 : 3 : 1</td>
<td>0.819</td>
<td>0.819</td>
<td>1.000</td>
<td>1.000</td>
<td>0.961</td>
<td>0.961</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>0.9990</td>
<td>2</td>
<td>33 : 1</td>
<td>0.918</td>
<td>0.918</td>
<td>1.000</td>
<td>1.000</td>
<td>0.525</td>
<td>0.525</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.0004</td>
<td>3</td>
<td>99 : 11 : 1</td>
<td>0.889</td>
<td>0.889</td>
<td>1.000</td>
<td>1.000</td>
<td>0.823</td>
<td>0.823</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.0004</td>
<td>4</td>
<td>96 : 24 : 6 : 1</td>
<td>0.860</td>
<td>0.860</td>
<td>1.000</td>
<td>1.000</td>
<td>0.910</td>
<td>0.910</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.0002</td>
<td>5</td>
<td>96 : 48 : 16 : 4 : 1</td>
<td>0.833</td>
<td>0.833</td>
<td>1.000</td>
<td>1.000</td>
<td>0.941</td>
<td>0.941</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>96 : 48 : 24 : 12 : 4 : 1</td>
<td>0.807</td>
<td>0.807</td>
<td>1.000</td>
<td>1.000</td>
<td>0.943</td>
<td>0.943</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Table A.3: Classification accuracy measures when $S_{e,j} = 0.95$, $S_{p,j} = 0.99$, and number of stages $S \in \{2, 3, 4, 5, 6\}$. These are calculated from our expressions in Section 2.3 and Appendix A.3. The optimal halving algorithm is the same as that identified in Table 2.2. The master pool size is $n_1$.

<table>
<thead>
<tr>
<th>$S$</th>
<th>$n_1$</th>
<th>$P_{S_{e,1}}$</th>
<th>$P_{S_{e,2}}$</th>
<th>$P_{S_{p,1}}$</th>
<th>$P_{S_{p,2}}$</th>
<th>$PPV_1$</th>
<th>$PPV_2$</th>
<th>$NPV_1$</th>
<th>$NPV_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{00} = 0.90$</td>
<td>2</td>
<td>4</td>
<td>0.916</td>
<td>0.918</td>
<td>0.997</td>
<td>0.997</td>
<td>0.951</td>
<td>0.940</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6</td>
<td>0.884</td>
<td>0.889</td>
<td>0.998</td>
<td>0.998</td>
<td>0.964</td>
<td>0.955</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>12</td>
<td>0.863</td>
<td>0.870</td>
<td>0.998</td>
<td>0.998</td>
<td>0.964</td>
<td>0.956</td>
<td>0.991</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>24</td>
<td>0.851</td>
<td>0.861</td>
<td>0.998</td>
<td>0.998</td>
<td>0.964</td>
<td>0.956</td>
<td>0.991</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>48</td>
<td>0.846</td>
<td>0.857</td>
<td>0.998</td>
<td>0.998</td>
<td>0.964</td>
<td>0.956</td>
<td>0.990</td>
</tr>
<tr>
<td>$p_{01} = 0.01$</td>
<td>2</td>
<td>5</td>
<td>0.917</td>
<td>0.929</td>
<td>0.998</td>
<td>0.998</td>
<td>0.950</td>
<td>0.897</td>
<td>0.997</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>8</td>
<td>0.886</td>
<td>0.910</td>
<td>0.999</td>
<td>0.999</td>
<td>0.963</td>
<td>0.921</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>12</td>
<td>0.858</td>
<td>0.892</td>
<td>0.999</td>
<td>0.999</td>
<td>0.973</td>
<td>0.941</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>24</td>
<td>0.838</td>
<td>0.883</td>
<td>0.999</td>
<td>0.999</td>
<td>0.973</td>
<td>0.942</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>48</td>
<td>0.825</td>
<td>0.878</td>
<td>0.999</td>
<td>0.999</td>
<td>0.973</td>
<td>0.943</td>
<td>0.993</td>
</tr>
<tr>
<td>$p_{10} = 0.03$</td>
<td>2</td>
<td>7</td>
<td>0.928</td>
<td>0.928</td>
<td>0.998</td>
<td>0.998</td>
<td>0.910</td>
<td>0.910</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>10</td>
<td>0.907</td>
<td>0.907</td>
<td>0.999</td>
<td>0.999</td>
<td>0.940</td>
<td>0.940</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>16</td>
<td>0.889</td>
<td>0.889</td>
<td>0.999</td>
<td>0.999</td>
<td>0.953</td>
<td>0.953</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>24</td>
<td>0.872</td>
<td>0.872</td>
<td>0.999</td>
<td>0.999</td>
<td>0.966</td>
<td>0.966</td>
<td>0.997</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>32</td>
<td>0.855</td>
<td>0.855</td>
<td>1.000</td>
<td>1.000</td>
<td>0.980</td>
<td>0.980</td>
<td>0.997</td>
</tr>
<tr>
<td>$p_{01} = 0.97$</td>
<td>2</td>
<td>11</td>
<td>0.920</td>
<td>0.920</td>
<td>0.999</td>
<td>0.999</td>
<td>0.831</td>
<td>0.831</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>16</td>
<td>0.891</td>
<td>0.891</td>
<td>0.999</td>
<td>0.999</td>
<td>0.888</td>
<td>0.888</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>24</td>
<td>0.864</td>
<td>0.864</td>
<td>1.000</td>
<td>1.000</td>
<td>0.916</td>
<td>0.916</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>32</td>
<td>0.839</td>
<td>0.839</td>
<td>1.000</td>
<td>1.000</td>
<td>0.945</td>
<td>0.945</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>48</td>
<td>0.816</td>
<td>0.816</td>
<td>1.000</td>
<td>1.000</td>
<td>0.961</td>
<td>0.961</td>
<td>0.999</td>
</tr>
<tr>
<td>$p_{11} = 0.002$</td>
<td>2</td>
<td>33</td>
<td>0.918</td>
<td>0.918</td>
<td>1.000</td>
<td>1.000</td>
<td>0.525</td>
<td>0.525</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>48</td>
<td>0.888</td>
<td>0.888</td>
<td>1.000</td>
<td>1.000</td>
<td>0.706</td>
<td>0.706</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>68</td>
<td>0.860</td>
<td>0.860</td>
<td>1.000</td>
<td>1.000</td>
<td>0.778</td>
<td>0.778</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>96</td>
<td>0.833</td>
<td>0.833</td>
<td>1.000</td>
<td>1.000</td>
<td>0.835</td>
<td>0.835</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>96</td>
<td>0.807</td>
<td>0.807</td>
<td>1.000</td>
<td>1.000</td>
<td>0.915</td>
<td>0.915</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Derivation of $E(C(S))$:

Let $C(S)$ denote the number of individuals correctly classified in a master pool that is to be resolved in $S$ stages. In the two-infection case, a “correct” classification is understood to mean that an individual is correctly classified for both infections. This derivation assumes the two-infection case; generalizing it for more than two infections is straightforward.

It is possible to derive $E(C(S))$ in closed form by using our Markov chain conceptualization described in Section 2.3 of the manuscript. From the definition of $C(S)$, it is easy to see that

$$E(C(S)) = n_1 \sum_{z_1=0}^{1} \sum_{z_2=0}^{1} \text{pr}(Z'_{S,i} = (z_1, z_2), \bar{Z}'_{S,i} = (z_1, z_2))$$

$$= n_1 \sum_{z_1=0}^{1} \sum_{z_2=0}^{1} \text{pr}(Z'_{S,i} = (z_1, z_2) \mid \bar{Z}'_{S,i} = (z_1, z_2)) \text{pr}(\bar{Z}'_{S,i} = (z_1, z_2)), \quad \text{(A.4)}$$

where $\bar{Z}_{S,i} = (\bar{Z}_{S,i1}, \bar{Z}_{S,i2})'$ is the vector of true infection statuses for the $i$th pool (i.e., individual) at stage $S$ and where $Z_{S,i} = (Z_{S,i1}, Z_{S,i2})'$ is the vector of (potentially incorrect) testing outcomes for this same individual. The probability $\text{pr}(\bar{Z}'_{S,i} = (z_1, z_2))$ on the right-hand side of Equation (A.4) is $p_{z_1z_2}$, where $z_1, z_2 \in \{0, 1\}$. We use our Markov chain framework to calculate $\text{pr}(Z'_{S,i} = (z_1, z_2) \mid \bar{Z}'_{S,i} = (z_1, z_2))$ for $z_1, z_2 \in \{0, 1\}$. These four calculations are similar in spirit, so we illustrate with $z_1 = 1$ and $z_2 = 0$.

The probability $\text{pr}(Z'_{S,i} = (1,0) \mid \bar{Z}'_{S,i} = (1,0))$ is derived in the same manner as $P_{S_{x,j}}$ and $P_{S_{p,j}}$, namely, by exploiting the Markov structure of $G^{(1)}_{S,i}, G^{(2)}_{S,i}, \ldots, G^{(S-1)}_{S,i}$ emerging after the individual $G_{S,i}$ is removed. The joint probability of the true status path of $G^{(1)}_{S,i}, G^{(2)}_{S,i}, \ldots, G_{S,i}$, conditional on the event $\{\bar{Z}'_{S,i} = (1,0)\}$ can be found by calculating

$$\text{pr}(Z^{(1)}_{S,i} = \bar{z}_1, Z^{(2)}_{S,i} = \bar{z}_2, \ldots, Z^{(S-1)}_{S,i} = \bar{z}_{S-1}, \bar{Z}_{S,i} = (1, 0)' \mid \bar{Z}'_{S,i} = (1, 0)')$$

$$= \text{pr}(Z^{(1)}_{S,i} = \bar{z}_1, Z^{(2)}_{S,i} = \bar{z}_2, \ldots, Z^{(S-1)}_{S,i} = \bar{z}_{S-1}, \tilde{z}_{S-1}) \quad \text{(A.5)}$$
where $z'_1, z'_2, \ldots, z'_{S-1} \in \{(0, 0), (1, 0), (0, 1), (1, 1)\}$. In Section 2.3 of the manuscript, we have shown that the probability on the right-hand side of Equation (A.5) is collected in $C_{-1} = M_{-1} \pi^{(1)} \pi^{(2)} \cdots \pi^{(S-2)}$, where $M_{-1}$ and $\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(S-2)}$ are defined therein.

The next step is to incorporate the effect of assay error so that $G_{S,i}^{(1)}, G_{S,i}^{(2)}, \ldots, G_{S,i}^{(S-1)}$ are all diagnosed positively for at least one infection. Because of the conditional event $\{z'_{S,i} = (1, 0)\}$, the true status of the first infection in each of $G_{S,i}^{(1)}, G_{S,i}^{(2)}, \ldots, G_{S,i}^{(S-1)}$ must be positive, and the true statuses for the second infection are determined by $z'_{S,i}$, $t = 1, 2, \ldots, S - 1$. Therefore, we use the matrix operator $P^{(s)}_{+-}$, defined in Section 2.3 of the manuscript, to augment $C_{-1}$ and obtain

$$\Pr(z'_{S,i} = (1, 0)|z'_{S,i} = (1, 0)) = 1'_{1}M_{-1}P^{(1)}_{+-} \prod_{t=1}^{S-2} (\pi^{(t)}_{-1}P^{(t+1)}_{+-}) 1_{4}S^{(S)}_{e:1}S^{(S)}_{p:2}. \quad (A.6)$$

The additional $S^{(S)}_{e:1}S^{(S)}_{p:2}$ in Equation (A.6) accounts for the correct diagnosis at stage $S$ where individual testing occurs.

The preceding argument applies for the other choices of $z_1$ and $z_2$ in Equation (A.4); i.e., for $(z_1, z_2) \in \{(0, 0), (0, 1), (1, 1)\}$. The double sum in Equation (A.4) equals

$$p_{00} \left[ 1 - 1'_{1}M_{-1}P^{(1)}_{+-} \prod_{t=1}^{S-2} (\pi^{(t)}_{-1}P^{(t+1)}_{+-}) 1_{4}(1 - S^{(S)}_{e:1}S^{(S)}_{p:2}) \right]$$

$$+ p_{10} \left[ 1'_{1}M_{-1}P^{(1)}_{+-} \prod_{t=1}^{S-2} (\pi^{(t)}_{-1}P^{(t+1)}_{+-}) 1_{4}S^{(S)}_{e:1}S^{(S)}_{p:2} \right]$$

$$+ p_{01} \left[ 1'_{1}M_{-1}P^{(1)}_{+-} \prod_{t=1}^{S-2} (\pi^{(t)}_{-1}P^{(t+1)}_{+-}) 1_{4}S^{(S)}_{e:2}S^{(S)}_{p:2} \right]$$

$$+ p_{11} \left[ 1'_{1}M_{-1}P^{(1)}_{+-} \prod_{t=1}^{S-2} (\pi^{(t)}_{-1}P^{(t+1)}_{+-}) 1_{4}S^{(S)}_{e:1}S^{(S)}_{e:2} \right], \quad (A.7)$$

where the matrix operators $P^{(s)}_{-+}, P^{(s)}_{+-},$ and $P^{(s)}_{++}$ are defined in Section 2.3 of the manuscript and in Appendix A.3. This derivation applies for $S > 2$ stages. When $S = 2$, simply replace each of the matrix products like $\prod_{t=1}^{S-2}(\pi^{(t)}_{-1}P^{(t+1)}_{+-})$ in Equation (A.7) with the $4 \times 4$ identity matrix $I_4$.  

100
Table A.4: Expected number of tests per individual $n_1^{-1}E(T^{(S)})$ when $p_{e_{ij}}^{(S)} = 0.95$, $S_{p_{ij}}^{(S)} = 0.99$, and number of stages $S \in \{2, 3, 4, 5, 6\}$. The column labeled “Optimal” gives the configuration of $n_1, n_2, ..., n_5$ that maximizes $E(C^{(S)})/E(T^{(S)})$; see Malinovsky et al. (2016). The column labeled “Halving” gives the master pool size for the optimal halving algorithm. The percent reduction in $n_1^{-1}E(T^{(S)})$ when compared to $n_1^{-1}E(T^{(2)})$ is provided. The expected proportion of correct classifications $n_1^{-1}E(C^{(S)})$ is also shown. The maximum allowable master pool size is 100.

<table>
<thead>
<tr>
<th>$S$</th>
<th>Optimal</th>
<th>$n_1^{-1}E(T^{(S)})$</th>
<th>% Reduction</th>
<th>$n_1^{-1}E(C^{(S)})$</th>
<th>Halving</th>
<th>$n_1^{-1}E(T^{(S)})$</th>
<th>% Reduction</th>
<th>$n_1^{-1}E(C^{(S)})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{00} = 0.90$</td>
<td>2</td>
<td>4 : 1</td>
<td>0.593</td>
<td>--</td>
<td>0.985</td>
<td>$n_1 = 4$</td>
<td>0.593</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>9 : 3 : 1</td>
<td>0.569</td>
<td>4.0</td>
<td>0.984</td>
<td>$n_1 = 6$</td>
<td>0.574</td>
<td>3.2</td>
<td>0.984</td>
</tr>
<tr>
<td>4</td>
<td>99 : 9 : 3 : 1</td>
<td>0.577</td>
<td>2.7</td>
<td>0.984</td>
<td>$n_1 = 12$</td>
<td>0.595</td>
<td>-0.3</td>
<td>0.982</td>
</tr>
<tr>
<td>5</td>
<td>90 : 45 : 9 : 3 : 1</td>
<td>0.595</td>
<td>-0.3</td>
<td>0.983</td>
<td>$n_1 = 24$</td>
<td>0.620</td>
<td>-4.6</td>
<td>0.980</td>
</tr>
<tr>
<td>6</td>
<td>96 : 48 : 24 : 6 : 3 : 1</td>
<td>0.619</td>
<td>-4.4</td>
<td>0.982</td>
<td>$n_1 = 48$</td>
<td>0.637</td>
<td>-7.4</td>
<td>0.980</td>
</tr>
<tr>
<td>$p_{00} = 0.95$</td>
<td>2</td>
<td>5 : 1</td>
<td>0.433</td>
<td>--</td>
<td>0.991</td>
<td>$n_1 = 5$</td>
<td>0.433</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>9 : 3 : 1</td>
<td>0.371</td>
<td>14.3</td>
<td>0.992</td>
<td>$n_1 = 8$</td>
<td>0.385</td>
<td>11.1</td>
<td>0.991</td>
</tr>
<tr>
<td>4</td>
<td>18 : 6 : 3 : 1</td>
<td>0.370</td>
<td>14.5</td>
<td>0.990</td>
<td>$n_1 = 12$</td>
<td>0.373</td>
<td>13.9</td>
<td>0.990</td>
</tr>
<tr>
<td>5</td>
<td>90 : 18 : 6 : 3 : 1</td>
<td>0.377</td>
<td>12.9</td>
<td>0.990</td>
<td>$n_1 = 24$</td>
<td>0.381</td>
<td>12.0</td>
<td>0.989</td>
</tr>
<tr>
<td>6</td>
<td>96 : 48 : 12 : 6 : 3 : 1</td>
<td>0.388</td>
<td>10.4</td>
<td>0.989</td>
<td>$n_1 = 48$</td>
<td>0.392</td>
<td>9.5</td>
<td>0.989</td>
</tr>
<tr>
<td>$p_{00} = 0.97$</td>
<td>2</td>
<td>6 : 1</td>
<td>0.345</td>
<td>--</td>
<td>0.994</td>
<td>$n_1 = 6$</td>
<td>0.345</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>16 : 4 : 1</td>
<td>0.273</td>
<td>20.9</td>
<td>0.995</td>
<td>$n_1 = 10$</td>
<td>0.289</td>
<td>16.2</td>
<td>0.994</td>
</tr>
<tr>
<td>4</td>
<td>27 : 9 : 3 : 1</td>
<td>0.260</td>
<td>24.6</td>
<td>0.994</td>
<td>$n_1 = 12$</td>
<td>0.270</td>
<td>21.7</td>
<td>0.994</td>
</tr>
<tr>
<td>5</td>
<td>36 : 12 : 6 : 3 : 1</td>
<td>0.264</td>
<td>23.5</td>
<td>0.994</td>
<td>$n_1 = 24$</td>
<td>0.265</td>
<td>23.2</td>
<td>0.994</td>
</tr>
<tr>
<td>6</td>
<td>96 : 48 : 12 : 6 : 3 : 1</td>
<td>0.271</td>
<td>21.4</td>
<td>0.994</td>
<td>$n_1 = 48$</td>
<td>0.272</td>
<td>21.2</td>
<td>0.994</td>
</tr>
<tr>
<td>$p_{00} = 0.99$</td>
<td>2</td>
<td>11 : 1</td>
<td>0.209</td>
<td>--</td>
<td>0.997</td>
<td>$n_1 = 11$</td>
<td>0.209</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>25 : 5 : 1</td>
<td>0.135</td>
<td>35.4</td>
<td>0.998</td>
<td>$n_1 = 16$</td>
<td>0.156</td>
<td>25.4</td>
<td>0.997</td>
</tr>
<tr>
<td>4</td>
<td>48 : 12 : 4 : 1</td>
<td>0.117</td>
<td>44.0</td>
<td>0.998</td>
<td>$n_1 = 24$</td>
<td>0.131</td>
<td>37.3</td>
<td>0.997</td>
</tr>
<tr>
<td>5</td>
<td>81 : 27 : 9 : 3 : 1</td>
<td>0.112</td>
<td>46.4</td>
<td>0.998</td>
<td>$n_1 = 32$</td>
<td>0.118</td>
<td>43.5</td>
<td>0.998</td>
</tr>
<tr>
<td>6</td>
<td>72 : 24 : 12 : 6 : 3 : 1</td>
<td>0.112</td>
<td>46.4</td>
<td>0.997</td>
<td>$n_1 = 48$</td>
<td>0.113</td>
<td>45.9</td>
<td>0.997</td>
</tr>
<tr>
<td>$p_{00} = 0.999$</td>
<td>2</td>
<td>33 : 1</td>
<td>0.081</td>
<td>--</td>
<td>0.999</td>
<td>$n_1 = 33$</td>
<td>0.081</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>99 : 11 : 1</td>
<td>0.032</td>
<td>60.5</td>
<td>1.000</td>
<td>$n_1 = 48$</td>
<td>0.046</td>
<td>43.2</td>
<td>0.999</td>
</tr>
<tr>
<td>4</td>
<td>96 : 24 : 6 : 1</td>
<td>0.024</td>
<td>70.4</td>
<td>1.000</td>
<td>$n_1 = 68$</td>
<td>0.034</td>
<td>58.0</td>
<td>1.000</td>
</tr>
<tr>
<td>5</td>
<td>96 : 48 : 16 : 4 : 1</td>
<td>0.023</td>
<td>71.6</td>
<td>1.000</td>
<td>$n_1 = 96$</td>
<td>0.027</td>
<td>66.7</td>
<td>1.000</td>
</tr>
<tr>
<td>6</td>
<td>96 : 48 : 24 : 12 : 4 : 1</td>
<td>0.022</td>
<td>72.8</td>
<td>1.000</td>
<td>$n_1 = 96$</td>
<td>0.023</td>
<td>71.6</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Figure A.3: Optimal number of stages $S$ when $S_{e:j}^{(s)} = 0.95$, $S_{p:j}^{(s)} = 0.99$, and optimal configurations are determined from maximizing $E(C^{(S)})/E(T^{(S)})$; see Malinovsky et al. (2016). The maximum allowable master pool size is 100. In the lower left corner of each subfigure, we did not show values of $S$ larger than 6 to avoid crowding. Values of $\eta_1$ and $\eta_2$ in the white regions (barely detectable in the $\rho = 0.10$ subfigure) are not possible because correlations for binary random variables are restricted. Note that “$S = 1$” corresponds to individual testing.
A.6 **Additional results from Section 2.5.**

Below are the values of $S_{e:j}$ and $S_{p:j}$ for the Aptima Combo 2 Assay for female swab and urine specimens ($j = 1$: chlamydia; $j = 2$: gonorrhea). These values come from the most recent product literature, available at [http://www.hologic.com/](http://www.hologic.com/).

**Swab:** $S_{e:1} = 0.942$, $S_{e:2} = 0.992$, $S_{p:1} = 0.976$, $S_{p:2} = 0.987$

**Urine:** $S_{e:1} = 0.947$, $S_{e:2} = 0.913$, $S_{p:1} = 0.989$, $S_{p:2} = 0.993$

Here is an outline of additional material in this Appendix:

**Page 104:** Table A.5. Summary information for females tested in 2010 and 2011 in the four Region X states: Alaska, Idaho, Oregon, and Washington. The estimates are calculated from the individual diagnoses made during these years. The stratum sample sizes $N$ are also given for each year.

**Page 105:** Figure A.4. Boxplots of the 5000 simulated values of $T^{(S)}$, shown cross-classified by specimen type (swab/urine) and state (AK, ID, OR, WA). The average number of tests (and sample standard deviation) in each of these 8 strata are shown in Table 2.3 and Table 2.4.

**Page 106:** Figure A.5. Boxplots for the number of tests in Region X overall during 2011. These boxplots are formed by aggregating over specimen type and state; i.e., over the results displayed in Figure A.4.

**Page 107:** Table A.6. Estimates of $PS_{e:j}$, $PS_{p:j}$, $PPV_j$, and $NPV_j$ ($j = 1$: chlamydia; $j = 2$: gonorrhea) for each specimen type/state combination. These estimates are averages over $B = 5000$ implementations (as described in Section 2.5).
Table A.5: Region X IPP data summary (females only). Estimates are calculated from the individual test results during 2010-2011 assuming no testing error. The stratum sample sizes $N$ are also given. The 2010 prevalence estimates are used to determine the optimal hierarchical configuration for 2011 analysis shown in Table 2.3 and Table 2.4.

<table>
<thead>
<tr>
<th>State</th>
<th>Specimen</th>
<th>C</th>
<th>G</th>
<th>2010</th>
<th>2011</th>
<th># Individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>Swab</td>
<td>−</td>
<td>−</td>
<td>$\hat{p}_{00} = 0.941$</td>
<td>$\hat{p}_{00} = 0.929$</td>
<td>2010: $N = 2104$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>−</td>
<td>$\hat{p}_{10} = 0.057$</td>
<td>$\hat{p}_{10} = 0.066$</td>
<td>2011: $N = 2910$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>$\hat{p}_{01} = 0.002$</td>
<td>$\hat{p}_{01} = 0.003$</td>
<td>2010: $N = 6430$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>$\hat{p}_{11} = 0.000$</td>
<td>$\hat{p}_{11} = 0.002$</td>
<td>2011: $N = 4558$</td>
</tr>
<tr>
<td>Urine</td>
<td></td>
<td>−</td>
<td>−</td>
<td>$\hat{p}_{00} = 0.905$</td>
<td>$\hat{p}_{00} = 0.905$</td>
<td>2010: $N = 2104$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>−</td>
<td>$\hat{p}_{10} = 0.087$</td>
<td>$\hat{p}_{10} = 0.089$</td>
<td>2011: $N = 4558$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>$\hat{p}_{01} = 0.006$</td>
<td>$\hat{p}_{01} = 0.004$</td>
<td>2010: $N = 4558$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>$\hat{p}_{11} = 0.002$</td>
<td>$\hat{p}_{11} = 0.002$</td>
<td>2011: $N = 4558$</td>
</tr>
<tr>
<td>Idaho</td>
<td>Swab</td>
<td>−</td>
<td>−</td>
<td>$\hat{p}_{00} = 0.932$</td>
<td>$\hat{p}_{00} = 0.926$</td>
<td>2010: $N = 7718$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>−</td>
<td>$\hat{p}_{10} = 0.066$</td>
<td>$\hat{p}_{10} = 0.072$</td>
<td>2011: $N = 7470$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>$\hat{p}_{01} = 0.001$</td>
<td>$\hat{p}_{01} = 0.001$</td>
<td>2010: $N = 4420$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>$\hat{p}_{11} = 0.001$</td>
<td>$\hat{p}_{11} = 0.001$</td>
<td>2011: $N = 4420$</td>
</tr>
<tr>
<td>Urine</td>
<td></td>
<td>−</td>
<td>−</td>
<td>$\hat{p}_{00} = 0.930$</td>
<td>$\hat{p}_{00} = 0.918$</td>
<td>2010: $N = 7718$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>−</td>
<td>$\hat{p}_{10} = 0.068$</td>
<td>$\hat{p}_{10} = 0.079$</td>
<td>2011: $N = 7470$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>$\hat{p}_{01} = 0.001$</td>
<td>$\hat{p}_{01} = 0.002$</td>
<td>2011: $N = 4420$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>$\hat{p}_{11} = 0.001$</td>
<td>$\hat{p}_{11} = 0.001$</td>
<td>2011: $N = 4420$</td>
</tr>
<tr>
<td>Oregon</td>
<td>Swab</td>
<td>−</td>
<td>−</td>
<td>$\hat{p}_{00} = 0.938$</td>
<td>$\hat{p}_{00} = 0.932$</td>
<td>2010: $N = 35652$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>−</td>
<td>$\hat{p}_{10} = 0.060$</td>
<td>$\hat{p}_{10} = 0.065$</td>
<td>2011: $N = 38654$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>$\hat{p}_{01} = 0.001$</td>
<td>$\hat{p}_{01} = 0.002$</td>
<td>2010: $N = 8848$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>$\hat{p}_{11} = 0.001$</td>
<td>$\hat{p}_{11} = 0.001$</td>
<td>2011: $N = 8381$</td>
</tr>
<tr>
<td>Urine</td>
<td></td>
<td>−</td>
<td>−</td>
<td>$\hat{p}_{00} = 0.919$</td>
<td>$\hat{p}_{00} = 0.920$</td>
<td>2010: $N = 35652$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>−</td>
<td>$\hat{p}_{10} = 0.077$</td>
<td>$\hat{p}_{10} = 0.076$</td>
<td>2011: $N = 38654$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>$\hat{p}_{01} = 0.003$</td>
<td>$\hat{p}_{01} = 0.002$</td>
<td>2010: $N = 8848$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>$\hat{p}_{11} = 0.001$</td>
<td>$\hat{p}_{11} = 0.002$</td>
<td>2011: $N = 8381$</td>
</tr>
<tr>
<td>Washington</td>
<td>Swab</td>
<td>−</td>
<td>−</td>
<td>$\hat{p}_{00} = 0.942$</td>
<td>$\hat{p}_{00} = 0.935$</td>
<td>2010: $N = 23761$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>−</td>
<td>$\hat{p}_{10} = 0.055$</td>
<td>$\hat{p}_{10} = 0.061$</td>
<td>2011: $N = 21134$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>$\hat{p}_{01} = 0.002$</td>
<td>$\hat{p}_{01} = 0.003$</td>
<td>2010: $N = 18530$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>$\hat{p}_{11} = 0.001$</td>
<td>$\hat{p}_{11} = 0.001$</td>
<td>2011: $N = 16415$</td>
</tr>
<tr>
<td>Urine</td>
<td></td>
<td>−</td>
<td>−</td>
<td>$\hat{p}_{00} = 0.938$</td>
<td>$\hat{p}_{00} = 0.928$</td>
<td>2010: $N = 23761$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>−</td>
<td>$\hat{p}_{10} = 0.060$</td>
<td>$\hat{p}_{10} = 0.069$</td>
<td>2011: $N = 21134$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>$\hat{p}_{01} = 0.001$</td>
<td>$\hat{p}_{01} = 0.002$</td>
<td>2010: $N = 18530$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>$\hat{p}_{11} = 0.001$</td>
<td>$\hat{p}_{11} = 0.001$</td>
<td>2011: $N = 16415$</td>
</tr>
</tbody>
</table>
Figure A.4: Boxplots of the 5000 simulated values of $T^{(S)}$, shown cross-classified by specimen type (swab/urine) and state (AK, ID, OR, WA). The average number of tests (and sample standard deviation) in each of these 8 strata are shown in Table 2.3 and Table 2.4.
Figure A.5: Boxplots for the number of tests in Region X overall during 2011. These boxplots are formed by aggregating the number of tests expended over specimen type and state.
**Table A.6:** Region X IPP 2011 chlamydia and gonorrhea data. Classification accuracy (averaged over $B = 5000$ sets of pools) for 2-, 3-, and 4-stage hierarchical algorithms under the optimal configuration; see Table 2.3 and Table 2.4.

<table>
<thead>
<tr>
<th>State</th>
<th>Specimen</th>
<th>$S$</th>
<th>$PS_{e:1}$</th>
<th>$PS_{e:2}$</th>
<th>$PS_{p:1}$</th>
<th>$PS_{p:2}$</th>
<th>$PPV_1$</th>
<th>$PPV_2$</th>
<th>$NPV_1$</th>
<th>$NPV_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>Swab</td>
<td>$S = 2$</td>
<td>0.891</td>
<td>0.988</td>
<td>0.993</td>
<td>0.996</td>
<td>0.910</td>
<td>0.566</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 3$</td>
<td>0.842</td>
<td>0.985</td>
<td>0.997</td>
<td>0.998</td>
<td>0.949</td>
<td>0.688</td>
<td>0.989</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 4$</td>
<td>0.801</td>
<td>0.983</td>
<td>0.997</td>
<td>0.998</td>
<td>0.949</td>
<td>0.701</td>
<td>0.986</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Urine</td>
<td>$S = 2$</td>
<td>0.899</td>
<td>0.872</td>
<td>0.997</td>
<td>0.998</td>
<td>0.969</td>
<td>0.690</td>
<td>0.990</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 3$</td>
<td>0.854</td>
<td>0.841</td>
<td>0.998</td>
<td>0.998</td>
<td>0.978</td>
<td>0.744</td>
<td>0.986</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 4$</td>
<td>0.813</td>
<td>0.825</td>
<td>0.998</td>
<td>0.998</td>
<td>0.979</td>
<td>0.751</td>
<td>0.982</td>
<td>0.999</td>
</tr>
<tr>
<td>Idaho</td>
<td>Swab</td>
<td>$S = 2$</td>
<td>0.890</td>
<td>0.990</td>
<td>0.993</td>
<td>0.996</td>
<td>0.913</td>
<td>0.378</td>
<td>0.991</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 3$</td>
<td>0.840</td>
<td>0.987</td>
<td>0.997</td>
<td>0.997</td>
<td>0.952</td>
<td>0.508</td>
<td>0.998</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 4$</td>
<td>0.795</td>
<td>0.985</td>
<td>0.997</td>
<td>0.998</td>
<td>0.953</td>
<td>0.527</td>
<td>0.984</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Urine</td>
<td>$S = 2$</td>
<td>0.896</td>
<td>0.879</td>
<td>0.997</td>
<td>0.998</td>
<td>0.961</td>
<td>0.529</td>
<td>0.991</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 3$</td>
<td>0.852</td>
<td>0.848</td>
<td>0.998</td>
<td>0.999</td>
<td>0.978</td>
<td>0.641</td>
<td>0.987</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 4$</td>
<td>0.811</td>
<td>0.829</td>
<td>0.998</td>
<td>0.999</td>
<td>0.979</td>
<td>0.648</td>
<td>0.984</td>
<td>1.000</td>
</tr>
<tr>
<td>Oregon</td>
<td>Swab</td>
<td>$S = 2$</td>
<td>0.890</td>
<td>0.988</td>
<td>0.994</td>
<td>0.996</td>
<td>0.910</td>
<td>0.420</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 3$</td>
<td>0.840</td>
<td>0.984</td>
<td>0.997</td>
<td>0.998</td>
<td>0.950</td>
<td>0.552</td>
<td>0.989</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 4$</td>
<td>0.795</td>
<td>0.981</td>
<td>0.997</td>
<td>0.998</td>
<td>0.952</td>
<td>0.571</td>
<td>0.986</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Urine</td>
<td>$S = 2$</td>
<td>0.899</td>
<td>0.877</td>
<td>0.998</td>
<td>0.998</td>
<td>0.968</td>
<td>0.644</td>
<td>0.991</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 3$</td>
<td>0.854</td>
<td>0.850</td>
<td>0.998</td>
<td>0.999</td>
<td>0.978</td>
<td>0.709</td>
<td>0.988</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 4$</td>
<td>0.812</td>
<td>0.835</td>
<td>0.998</td>
<td>0.999</td>
<td>0.978</td>
<td>0.715</td>
<td>0.984</td>
<td>0.999</td>
</tr>
<tr>
<td>Washington</td>
<td>Swab</td>
<td>$S = 2$</td>
<td>0.890</td>
<td>0.987</td>
<td>0.994</td>
<td>0.996</td>
<td>0.906</td>
<td>0.524</td>
<td>0.993</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 3$</td>
<td>0.841</td>
<td>0.983</td>
<td>0.997</td>
<td>0.998</td>
<td>0.948</td>
<td>0.655</td>
<td>0.990</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 4$</td>
<td>0.797</td>
<td>0.980</td>
<td>0.997</td>
<td>0.998</td>
<td>0.950</td>
<td>0.671</td>
<td>0.987</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Urine</td>
<td>$S = 2$</td>
<td>0.899</td>
<td>0.879</td>
<td>0.997</td>
<td>0.998</td>
<td>0.960</td>
<td>0.544</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 3$</td>
<td>0.853</td>
<td>0.851</td>
<td>0.999</td>
<td>0.999</td>
<td>0.978</td>
<td>0.660</td>
<td>0.989</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 4$</td>
<td>0.811</td>
<td>0.831</td>
<td>0.999</td>
<td>0.999</td>
<td>0.978</td>
<td>0.667</td>
<td>0.986</td>
<td>1.000</td>
</tr>
</tbody>
</table>
A.7 Description of online resources.

This appendix summarizes the R program we have written to perform all of the calculations. The R program and the accompanying C++ code is available at the Biometrics website on Wiley Online Library and at www.chrisbilder.com/grouptesting. The R program contains five major functions:

1. **EFF()**: Calculates the expected number of tests per individual

2. **ACCU()**: Calculates the classification accuracy measures

3. **OPT.CONFIG()**: Finds the optimal configuration and the corresponding operating characteristics based on the user’s choice of objective function and pool splitting criterion; this function is written to allow for $S = 2, 3, 4, 5, 6$ stages

4. **OPT.STAGE()**: Creates contour plots like those in Figure 2.2

5. **SIM()**: Performs Monte Carlo simulations similar to those summarized in Sections 2.4 and 2.5; note that Region X data are not available for public distribution so exact replication of results in Section 2.5 is not possible.

The program also includes examples to illustrate how each function is used. Many of these functions rely on C++ code to speed up the calculations. This C++ code is called from within the R program by using components of the `Rcpp` and `RcppArmadillo` packages.
Appendix B

Chapter 3 Supplementary Materials

B.1 Derivations for operating characteristics of AT

Efficiency

Recall that AT entails testing all the row and column pools in the first stage, and testing individuals in class \( \mathcal{M}_+ \) separately in the second stage, where

\[
\mathcal{M}_+ = \left\{ I_{ij} : T_{ij1}^{(\text{AT})} + T_{ij2}^{(\text{AT})} \geq 1 \right\},
\]

and

\[
T_{ijk}^{(\text{AT})} = \begin{cases} 
1, & \text{if } R_{ik} = 1, C_{jk} = 1; \\
1, & \text{if } R_{ik} = 1, \sum_{j=1}^n C_{jk} = 0; \\
1, & \text{if } \sum_{i=1}^n R_{ik} = 0, C_{jk} = 1; \\
0, & \text{otherwise},
\end{cases}
\]

for \( k = 1, 2 \). Therefore, the efficiency of AT, \( \text{EFF}(\text{AT}) \), i.e., the expected number of tests per specimen to classify all individuals in an array with \( n \) rows and \( n \) columns, is

\[
\text{EFF}(\text{AT}) = \frac{1}{n^2} \left[ 2n + n^2 \text{pr} \left( T_{ij1}^{(\text{AT})} + T_{ij2}^{(\text{AT})} \geq 1 \right) \right]. \tag{B.1}
\]

We now derive \( \text{pr}(T_{ij1}^{(\text{AT})} + T_{ij2}^{(\text{AT})} \geq 1) \) on the right-hand side (RHS) of Equation (B.1). By addition rule for probability,

\[
\text{pr} \left( T_{ij1}^{(\text{AT})} + T_{ij2}^{(\text{AT})} \geq 1 \right) = \sum_{k=1}^2 \text{pr} \left( T_{ijk}^{(\text{AT})} = 1 \right) - \text{pr} \left( T_{ij1}^{(\text{AT})} = 1, T_{ij2}^{(\text{AT})} = 1 \right). \tag{B.2}
\]
On the RHS of Equation (B.2), the first term \( \text{pr}(T_{ijk}^{(AT)} = 1) \) gives the probability that individual \((i, j)\) being retested because row and column test responses suggest it is positive for the \( k \)th infection, regardless of the other infection, which we denote it by \( k' \)th infection, where \( k' \in \{1, 2\} \setminus k \). In other words, the event \( \{T_{ijk}^{(AT)} = 1\} \) can be treated as if there were only one infection, and it occurs if either the \( i \)th row and \( j \)th column tests positive for the \( k \)th infection, or the \( i \)th row (\( j \)th column) tests positive for the \( k \)th infection while all the columns (rows) test negative for the \( k \)th infection. Therefore, one can simply follow the derivation of Equation (9) in Kim et al. (2007), with \( q \) replaced by \( \pi_k = 1 - \pi_k \), \( S_e \) and \( S_p \) replaced by \( S_{e:k} \) and \( S_{p:k} \), respectively, to obtain \( \text{pr}(T_{ijk}^{(AT)} = 1) \). We skip the details here.

Now the key problem is to derive the second term on the RHS of Equation (B.2), i.e., \( \text{pr}(T_{ij1}^{(AT)} = 1, T_{ij2}^{(AT)} = 1) \). By the symmetry between rows and columns of the array, and the definition of \( T_{ijk}^{(AT)} \), the target probability can be expressed as

\[
\text{pr} \left( T_{ij1}^{(AT)} = 1, T_{ij2}^{(AT)} = 1 \right) = \text{pr} \left( R_i = (1, 1)', C_j = (1, 1)' \right) \\
+ 2 \sum_{k=1}^{2} \text{pr} \left( R_i = (1, 1)', C_{jk} = 1, \sum_{j'=1}^{n} C_{j'k'} = 0 \right) \\
+ 2 \text{pr} \left( R_i = (1, 1)', \sum_{j'=1}^{n} C_{j'1} = 0, \sum_{j'=1}^{n} C_{j'2} = 0 \right) \\
+ 2 \text{pr} \left( R_{i1} = 1, \sum_{j'=1}^{n} C_{j'1} = 0, C_{j2} = 1, \sum_{i'=1}^{n} R_{i'2} = 0 \right). \tag{B.3}
\]

To provide a closed-form expression for the probabilities on the RHS of Equation (B.3), we first define some functions to mitigate the complexity of notation.

**Function 1:** Let \( f_k(s) \) denote the probability of a pool with size \( s \) being diagnosed as positive for the \( k \)th infection, for \( k = 1, 2 \). Under the assumptions 1-4 in Section 3.3, it can be shown that

\[
f_k(s) = S_{e:k} + \pi_k^s(1 - S_{e:k} - S_{p:k}),
\]
for \( k = 1, 2 \).

**Function 2:** Let \( \eta_0^{(s)}, \eta_1^{(s)}, \eta_2^{(s)}, \) and \( \eta_3^{(s)} \) denote the probability that the true status of a pool with size \( s \) is “00”, “10”, “01”, and “11”, respectively. Then by Law of Total Probability (LOTP), we obtain

\[
\eta_0^{(s)} = p_{00}^s, \quad \eta_1^{(s)} = (p_{00} + p_{10})^s - p_{00}^s, \\
\eta_2^{(s)} = (p_{00} + p_{01})^s - p_{00}^s, \quad \text{and} \\
\eta_3^{(s)} = 1 - \eta_0^{(s)} - \eta_1^{(s)} - \eta_2^{(s)}.
\]

Let \( f_b(s) \) denote the probability of a pool with size \( s \) being diagnosed as positive for both infections. Then

\[
f_b(s) = \eta_0^{(s)}S_{p:1}S_{p:2} + \eta_1^{(s)}S_{e:1}S_{p:2} + \eta_2^{(s)}S_{p:1}S_{e:2} + \eta_3^{(s)}S_{e:1}S_{e:2},
\]

where \( S_{e:k} = 1 - S_{e:k}, \quad S_{p:k} = 1 - S_{p:k} \).

**Function 3:** For an array with \( s_1 \) columns and \( s_2 \) rows, let \( g_k(c; s_2, s_1) \) denote the probability that \( c \) columns, \( c = 0, 1, \ldots, s_1 \), are truly positive for \( k \)th infection, for \( k = 1, 2 \). We have

\[
g_k(c; s_2, s_1) = \binom{s_1}{c} \left( \eta_0^{(s_2)} + \eta_3^{(s_2)} \right)^c \left( 1 - \eta_0^{(s_2)} - \eta_3^{(s_2)} \right)^{s_1-c}.
\]

**Function 4:** For an array with \( n \) rows and \( s \) columns, suppose we know that each column is truly “00” with probability \( \theta_{00} \), truly “10” with probability \( \theta_{10} \), truly “01” with probability \( \theta_{01} \), and truly “11” with probability \( \theta_{11} \), we can calculate the probability that \( c_1 \) columns and \( c_2 \) columns are truly positive for the first and second infection, respectively, by

\[
g_b(s, c_1, c_2; \theta_{00}, \theta_{10}, \theta_{01}, \theta_{11}) = \sum_{w=0}^{\min(c_1, c_2)} \binom{s}{w} \binom{s-w}{c_1-w} \binom{s-c_1-c_2+w}{c_2-w} \theta_{11}^{c_1-w} \theta_{10}^{c_2-w} \theta_{01}^{c_1-c_2+w} \theta_{00}^{c_1-c_2+w},
\]

for \( c_1, c_2 = 0, 1, \ldots, s \).

**Function 5:** For an array with \( s_1 \) columns and \( s_2 \) rows, let \( B(s_1, c_1; s_2, r_2) \) denote the probability that \( c_1 \) columns and \( r_2 \) rows are truly positive for the first and second
infection, respectively. For $c_1 = 0, 1, \ldots, s_1$ and $r_2 = 0, 1, \ldots, s_2$, we have

$$B(s_1, c_1; s_2, r_2) = \frac{s_1}{c_1} \frac{s_2}{r_2} \pi_2^{s_2-r_2} c_1 \pi_1^{s_1-c_1} r_2 \left[ 1 - \left( 1 - \pi_1^{r_2} \left( \frac{p_{00}}{\pi_1} \right)^{s_2-r_2} \right)^{c_1} \right] - \left[ 1 - \left( 1 - \pi_2^{r_2} \left( \frac{p_{00}}{\pi_2} \right)^{s_1-c_1} \right)^{r_2} \right] + \sum_{l=1}^{c_1} \sum_{m=1}^{r_2} (-1)^{m+l} \left( \frac{c_1}{l} \right) \left( \frac{r_2}{m} \right) \frac{p_{00}^m \pi_1^{l(r_2-m)} \pi_2^{(c_1-l)m}}{\pi_2} \times \left( \frac{p_{00}}{\pi_2} \right)^{(s_2-r_2)} \left( \frac{p_{00}}{\pi_1} \right)^{m(s_1-c_1)} (s_1-c_1)(s_2-r_2).$$

With Function 1-5 defined, we now provide the closed-form expressions for the four terms on the RHS of Equation (B.3) one-by-one.

- The first term in Equation (B.3) can be calculated by

$$\text{pr} \left( R_i = (1, 1)', C_j = (1, 1)' \right) = \sum_{\bar{y}_1=0}^{1} \sum_{\bar{y}_2=0}^{1} \text{pr} \left( R_i = (1, 1)', C_j = (1, 1)' \mid \bar{Y}_{ij} = (\bar{y}_1, \bar{y}_2)' \right) \text{pr} \left( \bar{Y}_{ij} = (\bar{y}_1, \bar{y}_2)' \right) = p_{00} [f_2(n-1)]^2 + p_{10} S_{e:1} [f_2(n-1)]^2 + p_{01} S_{e:2} [f_1(n-1)]^2 + p_{11} S_{e:1}^2 S_{e:2}^2,$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are defined in Function 1, and $f_k(\cdot)$ is defined in Function 2.

- The second term in Equation (B.3) can be decomposed by

$$\text{pr} \left( R_i = (1, 1)', C_{jk} = 1, \sum_{j'=1}^{n} C_{j'k'} = 0 \right) = \sum_{\bar{r}_k=0}^{1} \sum_{\bar{c}_k=0}^{1} \sum_{\bar{r}_{k'}=0}^{1} \sum_{\bar{c}_{k'}=0}^{1} \text{pr} \left( \bar{R}_{ik} = r_k, \bar{C}_{jk} = c_k, \bar{R}_{ik'} = r_{k'}, \sum_{j'=1}^{n} \bar{C}_{j'k'} = c_{k'} \right) \times \text{pr} \left( R_i = (1, 1)', C_{jk} = 1, \sum_{j'=1}^{n} C_{j'k'} = 0 \mid \bar{R}_{ik} = r_k, \bar{C}_{jk} = c_k, \bar{R}_{ik'} = r_{k'}, \sum_{j'=1}^{n} \bar{C}_{j'k'} = c_{k'} \right)$$

$$= \sum_{\bar{r}_k=0}^{1} \sum_{\bar{c}_k=0}^{1} \sum_{\bar{r}_{k'}=0}^{1} \sum_{\bar{c}_{k'}=0}^{1} \text{pr} \left( \bar{R}_{ik} = r_k, \bar{C}_{jk} = c_k, \bar{R}_{ik'} = r_{k'}, \sum_{j'=1}^{n} \bar{C}_{j'k'} = c_{k'} \right) \times S_{e:1}^{r_k+c_k} S_{p:k}^{2-r_k-c_k} S_{e:k'}^{r_{k'}-c_{k'}} S_{p:k'}^{1-r_{k'}-c_{k'}} S_{e:k'}^{r_{k'}} S_{p:k'}^{n-c_{k'}}.$$

Denote the probability $\text{pr}(\bar{R}_{ik} = r_k, \bar{C}_{jk} = c_k, \bar{R}_{ik'} = r_{k'}, \sum_{j'=1}^{n} \bar{C}_{j'k'} = c_{k'})$ in above equation as $h_{r_k c_k r_{k'}}^{(1)}(n; k, c_{k'})$. In the following, we calculate $h_{r_k c_k r_{k'}}^{(1)}(n; k, c_{k'})$ case by
case for $k = 1, 2$, and each value of $r_k, c_k, r_{k'} \in \{0, 1\}$ and $c_{k'} \in \{r_{k'}, r_{k'} + 1, \ldots n\}$. By LOTP, it is straightforward to see that when $r_k, c_k, r_{k'} = 0$ and $c_{k'} \in \{0, 1, \ldots n\}$, we have

$$h_{000}^{(1)}(n; k, c_{k'}) = \frac{1}{c=0} \text{pr} \left( \tilde{R}_{ik} = 0, \tilde{C}_{jk} = 0, \tilde{R}_{ik'} = 0, \tilde{C}_{jk'} = c, \sum_{j \neq j'} \tilde{C}_{j'k'} = c_{k'} - c \right)$$

$$= \frac{1}{c=0} \eta_0^{(n)} \eta_{C \times k'}^{(n-1)} g_{k'}(c_{k'} - c; n - 1, n - 1),$$

for $k = 1, 2$, and $g_{k'}(\cdot)$ in above equation is defined in Function 3. Follow similar arguments, we present $h_{rk,crk'}^{(1)}(n; k, c_{k'})$ for the other values of $r_k, c_k, r_{k'}$ and $c_{k'} \in \{r_{k'}, r_{k'} + 1, \ldots n\}$ below.

- When $r_k = 0$, $c_k = 0$, and $r_{k'} = 1$,

$$h_{001}^{(1)}(n; k, c_{k'}) = \frac{1}{c=0} \eta_{C \times k}^{(n)} \eta_{C \times k'}^{(n-1)} \sum_{w=0}^{n-c} \binom{n-1}{w} \left[ \eta_{C}^{(1)} \right]_{n-w}^{w} \eta_{k'}^{(1)} \eta_{k}^{(1)} p_{00}^{n-w-1}$$

$$\times g_{k'}(c_{k'} - w - c; n - 1, n - w - 1) - h_{000}^{(1)}(n; k, c_{k'});$$

- When $r_k = 0$, $c_k = 1$, and $r_{k'} = 0$,

$$h_{010}^{(1)}(n; k, c_{k'}) = \sum_{c=0}^{1} \eta_{0}^{(n)} \eta_{C \times k}^{(n-1)} g_{k'}(c_{k'} - c; n - 1, n - 1);$$

- When $r_k = 1$, $c_k = 0$, and $r_{k'} = 0$,

$$h_{100}^{(1)}(n; k, c_{k'}) = \eta_{k}^{(n-1)} \sum_{c=0}^{1} \eta_{0}^{(n)} \eta_{C \times k}^{(n-c)} g_{k'}(c_{k'} - c; n - 1, n - 1);$$

- When $r_k = 0$, $c_k = 1$, and $r_{k'} = 1$,

$$h_{011}^{(1)}(n; k, c_{k'}) = \sum_{c=0}^{1} \eta_{C \times k}^{(n)} \eta_{C \times k'}^{(n-c)} g_{k'}(c_{k'} - c; n, n - 1, n - w)$$

$$- h_{000}^{(1)}(n; k, c_{k'}) - h_{001}^{(1)}(n; k, c_{k'}) - h_{010}^{(1)}(n; k, c_{k'});$$

- When $r_k = 1$, $c_k = 0$, and $r_{k'} = 1$,

$$h_{101}^{(1)}(n; k, c_{k'}) = \sum_{c=0}^{1} \eta_{C \times k}^{(n)} g_{k'}(c_{k'} - c; n, n - 1)$$

$$- h_{000}^{(1)}(n; k, c_{k'}) - h_{001}^{(1)}(n; k, c_{k'}) - h_{100}^{(1)}(n; k, c_{k'});$$

113
- When $r_k = 1$, $c_k = 1$, and $r_{k'} = 0$,

$$h_{110}^{(1)}(n; k, c_{k'}) = \left[ \eta_0^{(n)} + \eta_k^{(n)} \right] g_{k'}(c_{k'}; n - 1, n) - h_{000}^{(1)}(n; k, c_{k'}) - h_{010}^{(1)}(n; k, c_{k'}) - h_{100}^{(1)}(n; k, c_{k'});$$

- When $r_k = 1$, $c_k = 1$, and $r_{k'} = 1$,

$$h_{111}^{(1)}(n; k, c_{k'}) = g_{k'}(c_{k'}; n, n) - \sum_{r_k=0}^{1} \sum_{c_k=0}^{1} \sum_{r_{k'}=0}^{1} h_{r_k c_k r_{k'}}^{(1)}(n; k, c_{k'}).$$

This completes the derivation of $h_{r_k c_k r_{k'}}^{(1)}(n; k, c_{k'})$.

- Similarly, the third term in Equation (B.3) can be expressed by

$$\text{pr} \left( \mathbf{R}_i = (1, 1), \sum_{j'=1}^{n} C_{j'1} = 0, \sum_{j'=1}^{n} C_{j'2} = 0 \right) = \sum_{r_1=0}^{1} \sum_{r_2=0}^{1} \sum_{c_1=r_1}^{n} \sum_{c_2=r_2}^{n} \text{pr} \left( \mathbf{R}_i = (r_1, r_2)', \sum_{j'=1}^{n} \tilde{C}_{j'1} = c_1, \sum_{j'=1}^{n} \tilde{C}_{j'2} = c_2 \right) \times S_{e_1}^{r_1} S_{e_2}^{1-r_1} S_{e_2}^{1-r_2} S_{e_2}^{\tilde{C}_{1}-c_1} S_{e_2}^{\tilde{C}_{2}-c_2}.$$

Denote $\text{pr}(\mathbf{R}_i = (r_1, r_2)', \sum_{j'=1}^{n} \tilde{C}_{j'1} = c_1, \sum_{j'=1}^{n} \tilde{C}_{j'2} = c_2)$ on the RHS of the equation above as $h_{r_1 r_2}^{(2)}(n; c_1, c_2)$. Then for any $c_1 \in \{r_1, r_1 + 1, \ldots, n\}$ and $c_2 \in \{r_2, r_2 + 1, \ldots, n\}$, we have the following table.

- When $r_1 = 0$ and $r_2 = 0$,

$$h_{00}^{(2)}(n; c_1, c_2) = g_b \left( n, c_1, c_2; \eta_0^{(n)}, \eta_0^{(n-1)}, \eta_0^{(n-1)}, \eta_0^{(n-1)} \right);$$

- When $r_1 = 0$ and $r_2 = 1$,

$$h_{01}^{(2)}(n; c_1, c_2) = g_b \left( n, c_1, c_2; \eta_0^{(n)}, \eta_1^{(n-1)}, \eta_2^{(n)}, \eta_0^{(n-1)} \right) + p_{01} \left[ \eta_1^{(n-1)} + \eta_3^{(n-1)} \right] - h_{00}^{(2)}(n; c_1, c_2);$$
- When \( r_1 = 1 \) and \( r_2 = 0 \),

\[
h^{(2)}_{10}(n; c_1, c_2) = g_b\left(n, c_1, c_2; \eta_0^{(m)}, \eta_1^{(n)}, p_{00}h_2^{(n-1)} + p_{10}h_3^{(n-1)}\right) - h^{(2)}_{00}(n; c_1, c_2);
\]

- When \( r_1 = 1 \) and \( r_2 = 1 \),

\[
h^{(2)}_{11}(n; c_1, c_2) = g_b\left(n, c_1, c_2; \eta_0^{(m)}, \eta_1^{(n)}, \eta_2^{(n)}, \eta_3^{(n)}\right) - \sum_{r_1=0}^{1} \sum_{r_2=0}^{1} h^{(2)}_{r_1r_2}(n; c_1, c_2).
\]

This completes the derivation of \( h^{(2)}_{r_1r_2}(n; c_1, c_2) \). Note that Function \( g_b(\cdot) \) is defined in Function 4.

- Lastly, the fourth term in Equation (B.3) is

\[
\text{pr}\left( R_{i1} = 1, \sum_{j' = 1}^{n} C_{j'1} = 0, C_{j2} = 1, \sum_{r' = 1}^{n} R_{i'r2} = 0 \right) = \sum_{r_1=0}^{1} \sum_{c_1=1}^{n} \sum_{c_2=0}^{1} \sum_{r_2=0}^{1} \text{pr}\left( \tilde{R}_{i1} = r_1, \sum_{j' = 1}^{n} \tilde{C}_{j'1} = c_1, \tilde{C}_{j2} = c_2, \sum_{r' = 1}^{n} \tilde{R}_{i'r2} = r_2 \right) \times S_{e_1}^{1-r_1} S_{e_2}^{1-c_1} S_{p_1}^{1-c_1} S_{p_2}^{1-c_2} S_{r_1}^{1-r_1} S_{r_2}^{1-c_2} S_{r_2}^{n-r_2}.
\]

Denote \( \text{pr}(\tilde{R}_{i1} = r_1, \sum_{j' = 1}^{n} \tilde{C}_{j'1} = c_1, \tilde{C}_{j2} = c_2, \sum_{r' = 1}^{n} \tilde{R}_{i'r2} = r_2) \) on the RHS of the equation above as \( h^{(3)}_{r_1c_2}(n; c_1, r_2) \). For \( r_1, c_2 \in \{0,1\} \), \( c_1 \in \{r_1, r_1 + 1, \ldots, n\} \), and \( r_2 \in \{c_2, c_2 + 1, \ldots, n\} \), we summarize \( h^{(3)}_{r_1c_2}(n; c_1, r_2) \) below.

- When \( r_1 = 0 \) and \( c_2 = 0 \),

\[
h^{(3)}_{00}(n; c_1, r_2) = \eta_0^{(2n-1)} B(n - 1, c_1; n - 1, r_2)
\]

\[
+ \eta_0^{(n)} \eta_1^{(n-1)} B(n - 1, c_1 - 1; n - 1, r_2)
\]

\[
+ \eta_0^{(n)} \eta_2^{(n-1)} B(n - 1, c_1; n - 1, r_2 - 1)
\]

\[
+ p_{00} \eta_1^{(n-1)} \eta_2^{(n-1)} B(n - 1, c_1 - 1; n - 1, r_2 - 1);
\]

- When \( r_1 = 0 \) and \( c_2 = 1 \),

\[
h^{(3)}_{01}(n; c_1, r_2) = \eta_0^{(n)} B(n, c_1; n - 1, r_2) + \eta_2^{(n)} B(n, c_1; n - 1, r_2 - 1) - h^{(3)}_{00}(n; c_1, r_2);
\]
- When $r_1 = 1$ and $c_2 = 0$,

$$h_{10}^{(3)}(n; c_1, r_2) = \eta_0^{(n)} B(n-1, c_1; n, r_2) + \eta_1^{(n)} B(n-1, c_1-1; n, r_2) - h_{00}^{(3)}(n; c_1, r_2);$$

- When $r_1 = 1$ and $c_2 = 1$,

$$h_{11}^{(3)}(n; c_1, r_2) = B(n, c_1; n, r_2) - h_{00}^{(3)}(n; c_1, r_2) - h_{10}^{(3)}(n; c_1, r_2) - h_{01}^{(3)}(n; c_1, r_2).$$

Note that Function $B(\cdot)$ is defined in Function 5. This completes the derivation of $h_{r_1, c_2}^{(3)}(n; c_1, r_2)$, and hence Equation (B.3). Substitute Equation (B.3) in Equation (B.2), we can obtain the efficiency of AT given by Equation (B.1).

**Pooling Sensitivity $PS_{e,k}$**

In this section, we derive the pooling sensitivity for the $k$th infection of AT, denoted by $PS_{e,k}^{(AT)}$. Recall that the pooling sensitivity for the $k$th infection gives the probability an individual is classified as positive for the $k$th infection given that it is truly positive, i.e., $PS_{e,k}^{(AT)} = \Pr(Y_{ijk} = 1, T_{ijk}^{(AT)} + T_{ijk}^{(AT)} \geq 1 | \widetilde{Y}_{ijk} = 1)$, where $Y_{ijk} = 1$ denotes that individual $(i, j)$ is diagnosed as positive for the $k$th infection while being individually diagnosed in the last stage; and 0 otherwise. Again, by addition rule, we have

$$PS_{e,k}^{(AT)} = \Pr \left( Y_{ijk} = 1, T_{ijk}^{(AT)} = 1 \mid \widetilde{Y}_{ijk} = 1 \right)$$

$$+ \Pr \left( Y_{ijk} = 1, T_{ijk}^{(AT)} = 1 \mid \widetilde{Y}_{ijk} = 1 \right)$$

$$- \Pr \left( Y_{ijk} = 1, T_{ijk}^{(AT)} = 1, T_{ijk}^{(AT)} = 1 \mid \widetilde{Y}_{ijk} = 1 \right). \quad (B.4)$$

We now derive three probabilities on the RHS of Equation (B.4). The first term
\[ \text{pr} \left( Y_{ijk} = 1, T_{ijk}^{(AT)} = 1 \left| \bar{Y}_{ijk} = 1 \right. \right) = \text{pr} \left( Y_{ijk} = 1, R_{ik} = 1, C_{jk} = 1 \left| \bar{Y}_{ijk} = 1 \right. \right) + 2 \text{pr} \left( Y_{ijk} = 1, R_{ik} = 1, \sum_{j'=1}^{n} C_{j'k} = 0 \left| \bar{Y}_{ijk} = 1 \right. \right) = S_{e:k}^{3} + 2S_{e:k}^{2}S_{e:k} \left[ 1 - f_{k}(n) \right]^{n-1}, \]

where \( f_{k}(\cdot) \) is defined in Function 1.

The second term in Equation (B.4) is

\[ \text{pr} \left( Y_{ijk} = 1, T_{ijk'}^{(AT)} = 1 \left| \bar{Y}_{ijk} = 1 \right. \right) = S_{e:k}^{-1} \left[ \text{pr} \left( R_{ik'} = 1, C_{jk'} = 1, \bar{Y}_{ijk} = 1 \right) + 2 \text{pr} \left( R_{ik'} = 1, \sum_{j'=1}^{n} C_{j'k'} = 0, \bar{Y}_{ijk} = 1 \right) \right] \]

\[ = S_{e:k}^{-1} \left\{ \eta_{k}^{(1)} [f_{k'}(n-1)]^{2} + p_{11}S_{e:k'}^{2} + 2p_{11}S_{e:k'}S_{e:k'} \left[ 1 - f_{k'}(n) \right]^{n-1} + \right. \]

\[ \left. + \text{pr} \left( R_{ik'} = 1, \sum_{j'=1}^{n} C_{j'k'} = 0, \bar{Y}_{ijk} = 1, \bar{Y}_{ijk'} = 0 \right) \right\}, \]

where

\[ \text{pr} \left( R_{ik'} = 1, \sum_{j'=1}^{n} C_{j'k'} = 0, \bar{Y}_{ijk} = 1, \bar{Y}_{ijk'} = 0 \right) = \sum_{r_{k'}=0}^{1} \sum_{c_{k'}=r_{k'}}^{n} \text{pr} \left( \bar{R}_{ik'} = r_{k'}, \sum_{j'=1}^{n} \bar{C}_{j'k'} = c_{k'}, \bar{Y}_{ijk} = 1, \bar{Y}_{ijk'} = 0 \right) S_{e:2}^{r_{k'}}S_{p:2}^{1-r_{k'}}S_{e:2}^{-1}S_{e:2}^{c_{k'}}. \]

Define \( \text{pr}(\bar{R}_{ik'} = r_{k'}, \sum_{j'=1}^{n} \bar{C}_{j'k'} = c_{k'}, \bar{Y}_{ijk} = 1, \bar{Y}_{ijk'} = 0) \) as \( h_{r_{k'}}^{(4)}(n; k, c_{k'}) \). Then for each \( k = 1, 2, \) and \( r_{k'}, c_{k'} \in \{0, 1\}, \) we get

\[ h_{0}^{(4)}(n; k, c_{k'}) = \eta_{k}^{(1)} \left[ \eta_{0}^{(n-1)} + \eta_{k}^{(n-1)} \right] g_{k'}(c_{k'}; n-1, n), \]

and

\[ h_{1}^{(4)}(n; k, c_{k'}) = \eta_{k}^{(1)} \left[ \eta_{0}^{(n-1)} + \eta_{k}^{(n-1)} \right] g_{k'}(c_{k'}; n, n-1) + \eta_{k}^{(1)} \left[ \eta_{3}^{(n-1)} + \eta_{k}^{(n-1)} \right] g_{k'}(c_{k'} - 1; n, n-1) - h_{0}^{(4)}(n; k, c_{k'}), \]
where \( g_{k'}(\cdot) \) is defined in Function 3.

Lastly, we derive the third term on the RHS of Equation (B.4), which can be decomposed as

\[
\Pr(Y_{ijk} = 1, T_{ijk}^{(AT)} = 1, T_{ijk'}^{(AT)} = 1 \mid Y_{ijk} = 1) = \Pr(Y_{ijk} = 1, R_{ik} = 1, C_{jk} = 1, T_{ijk'}^{(AT)} = 1 \mid \tilde{Y}_{ijk} = 1) \\
+ 2\Pr(Y_{ijk} = 1, R_i = (1, 1)', C_{jk} = 1, \sum_{j'=1}^n C_{j'k} = 0 \mid \tilde{Y}_{ijk} = 1) \\
+ 2\Pr(Y_{ijk} = 1, R_i = (1, 1)', \sum_{j'=1}^n C_{j'1} = 0, \sum_{j'=1}^n C_{j'2} = 0 \mid \tilde{Y}_{ijk} = 1) \\
+ 2\Pr(Y_{ijk} = 1, R_{i1} = 1, \sum_{j'=1}^n C_{j'1} = 0, \sum_{j'=1}^n R_{i'2} = 0, C_{j2} = 1 \mid \tilde{Y}_{ijk} = 1),
\]

(B.5)

Thus, it suffices to derive the four probabilities on the RHS of Equation (B.5).

We do this one-by-one:

- The first term in Equation (B.5) can be easily calculated by

\[
\Pr(Y_{ijk} = 1, R_{ik} = 1, C_{jk} = 1, T_{ijk'}^{(AT)} = 1 \mid \tilde{Y}_{ijk} = 1) = S_{e_k}^2 \Pr(Y_{ijk} = 1, T_{ijk'}^{(AT)} = 1 \mid \tilde{Y}_{ijk} = 1),
\]

where \( \Pr(Y_{ijk} = 1, T_{ijk'}^{(AT)} = 1 \mid \tilde{Y}_{ijk} = 1) \) is the second term in Equation (B.4) which has been derived above.

- The second term in Equation (B.5) is

\[
\Pr(Y_{ijk} = 1, R_i = (1, 1)', C_{jk} = 1, \sum_{j'=1}^n C_{j'k} = 0 \mid \tilde{Y}_{ijk} = 1) \\
= S_{e_k}^2 S_{e_k} \pi^{-1}_k \left\{ p_{11} S_{e_k}^2 [1 - f_k(n)]^{n-1} + f_{k'}(n-1) \Pr(R_{ik'} = 1, \sum_{j'\neq j}^n C_{j'k} = 0, \tilde{Y}_{ijk} = 1, \tilde{Y}_{ijk'} = 0) \right\},
\]

118
where \( f_k(\cdot) \) is defined in Function 1, and

\[
\Pr \left( R_{ik'} = 1, \sum_{j' \neq j} C_{j'k} = 0, \tilde{Y}_{ijk} = 1, \tilde{Y}_{ijk'} = 0 \right)
\]

\[
= \frac{1}{n} \sum_{r_{ik'}=0}^{n-1} \sum_{c_k=0}^{n} \Pr \left( R_{ik'} = r_{ik'}, \sum_{j' \neq j} \tilde{C}_{j'k} = c_k, \tilde{Y}_{ijk} = 1, \tilde{Y}_{ijk'} = 0 \right) \times S^r_{e,k} S^{1-r_{ik'}}_{p,k} S^{n-c_k-1}_{p,k} \varepsilon^c_{e,k}.
\]

Denote \( \Pr(\tilde{R}_{ik'} = r_{ik'}, \sum_{j' \neq j} \tilde{C}_{j'k} = c_k, \tilde{Y}_{ijk} = 1, \tilde{Y}_{ijk'} = 0) \) as \( h^{(5)}_{c_k}(n; k, c_k) \), then for \( k = 1, 2 \), and \( r_{ik'} \in \{0, 1\} \), \( c_k \in \{0, 1, \ldots, n - 1\} \), we have

\[
h^{(5)}_0(n; k, c_k) = \eta^{(1)}_k \sum_{w=0}^{c_k} \binom{n-1}{w} \left( \frac{1}{\eta^{(1)}_k} \right)^w p_0^{n-1-w} g_k(c_k - w; n - 1, n - w - 1),
\]

and

\[
h^{(5)}_1(n; k, c_k) = \eta^{(1)}_k g_k(c_k; n, n - 1) - h^{(5)}_0(n; k, c_k),
\]

where \( g_k(\cdot) \) is defined in Function 3.

- The third term in Equation (B.5) can be written as

\[
\Pr \left( Y_{ijk} = 1, R_i = (1, 1)', \sum_{j'=1}^{n} C_{j'1} = 0, \sum_{j'=1}^{n} C_{j'2} = 0 \right| \tilde{Y}_{ijk} = 1)
\]

\[
= S^2_{e,k} S^{p,k} S^{-1}_{e,k} \sum_{c_k=0}^{n} \sum_{r_{ik'}=0}^{n} \sum_{c_k'=0}^{n} \Pr \left( \sum_{j' \neq j} \tilde{C}_{j'k} = c_k, \tilde{R}_{ik'} = r_{ik'}, \sum_{j'=1}^{n} \tilde{C}_{j'k'} = c_k, \tilde{Y}_{ijk} = 1 \right) \times S^c_{e,k} S^{n-c_k-1}_{p,k} S^{1-r_{ik'}}_{p,k} S^{c_k'}_{e,k} S^{n-c_k'}_{p,k'}.
\]

Denote \( \Pr(\sum_{j' \neq j} \tilde{C}_{j'k} = c_k, \tilde{R}_{ik'} = r_{ik'}, \sum_{j'=1}^{n} \tilde{C}_{j'k'} = c_k', \tilde{Y}_{ijk} = 1) \) in above expression as \( h^{(6)}_{ik'}(n; k, c_k, c_k') \). For \( r_{ik'} \in \{0, 1\} \), \( c_k \in \{0, 1, \ldots, n - 1\} \) and \( c_k' \in \{r_{ik'}, r_{ik'}+1, \ldots, n\} \), we now provide closed-form expressions for \( h^{(6)}_{ik'}(n; k, c_k, c_k') \), for each \( k = 1, 2 \).
When $k = 1$,

\[
\hat{h}_0^{(6)}(n; 1, c_1, c_2) = \eta_1^{(1)} \left\{ \pi_2^{n-1} g_b \left( n - 1, c_1, c_2; \eta_0^{(n)}, \eta_1^{(n)}, \eta_2^{(n-1)} + \eta_1^{(1)} \eta_2^{(n-1)} \right) + \left[ 1 - \pi_2^{n-1} \right] g_b \left( n - 1, c_1, c_2 - 1; \eta_0^{(n)}, \eta_1^{(n)}, \eta_2^{(n-1)} + \eta_1^{(1)} \eta_2^{(n-1)} \right) \right\},
\]

and

\[
\hat{h}_1^{(6)}(n; 1, c_1, c_2) = \left\{ p_{11} + p_{10} \left[ 1 - \pi_2^{n-1} \right] \right\} g_b \left( n - 1, c_1, c_2; \eta_0^{(n)}, \eta_1^{(n)}, \eta_2^{(n)} \right) + p_{10} \pi_2^{n-1} g_b \left( n - 1, c_1, c_2; \eta_0^{(n)}, \eta_1^{(n)}, \eta_2^{(n)} \right) - \hat{h}_0^{(6)}(n; 1, c_1, c_2);
\]

When $k = 2$,

\[
\hat{h}_0^{(6)}(n; 2, c_2, c_1) = \eta_2^{(1)} \left\{ \pi_1^{n-1} g_b \left( n - 1, c_1, c_2; \eta_0^{(n)}, \eta_1^{(n-1)}, \eta_2^{(n)} \right) + \left[ 1 - \pi_1^{n-1} \right] g_b \left( n - 1, c_1 - 1, c_2; \eta_0^{(n)}, \eta_1^{(n-1)}, \eta_2^{(n)} \right) \right\},
\]

and

\[
\hat{h}_1^{(6)}(n; 2, c_2, c_1) = \left\{ p_{11} + p_{01} \left[ 1 - \pi_1^{n-1} \right] \right\} g_b \left( n - 1, c_1 - 1, c_2; \eta_0^{(n)}, \eta_1^{(n)}, \eta_2^{(n)} \right) + p_{01} \pi_1^{n-1} g_b \left( n - 1, c_1, c_2; \eta_0^{(n)}, \eta_1^{(n)}, \eta_2^{(n)} \right) - \hat{h}_0^{(6)}(n; 2, c_2, c_1),
\]

where $g_b(\cdot)$ is given in Function 4.
The last term in Equation (B.5) is
\[
\Pr \left( Y_{ijk} = 1, R_{i1} = 1, \sum_{j' = 1}^{n} C_{j'1} = 0, \sum_{i' = 1}^{n} R_{i'2} = 0, C_{j2} = 1 \mid \bar{Y}_{ijk} = 1 \right) = \Sigma_{c \in k} \Sigma_{j' \in k} (\pi_k)^{-1} \left[ \Pr \left( \sum_{j' \neq j} C_{j'k} = 0, \sum_{i' = 1}^{n} R_{i'k'} = 0, C_{jk'} = 1, \bar{Y}_{ijk} = 1, \bar{Y}_{ijk'} = 1 \right) + \Pr \left( \sum_{j' \neq j} C_{j'k} = 0, \sum_{i' = 1}^{n} R_{i'k'} = 0, C_{jk'} = 1, \bar{Y}_{ijk} = 1, \bar{Y}_{ijk'} = 0 \right) \right]. \tag{B.6}
\]

The first probability on the RHS of Equation (B.6) is
\[
\Pr \left( \sum_{j' \neq j} C_{j'k} = 0, \sum_{i' = 1}^{n} R_{i'k'} = 0, C_{jk'} = 1, \bar{Y}_{ijk} = 1, \bar{Y}_{ijk'} = 1 \right) = S_{c \in k} \Sigma_{c \in k} p_{11} \sum_{c_1 = 0}^{n-1} \sum_{w_2 = 0}^{n-1} \Pr \left( \sum_{j' \neq j} \tilde{C}_{j'1} = c_1, \sum_{i' \neq i} \tilde{R}_{i'2} = r_2 \right) \Sigma_{c \in k} \Sigma_{c \in k} n_{1}^{n-1-c_1-1} \Sigma_{r_2}^{n-2} \Sigma_{r_2}^{n-2-1} \times B(n-1, w_1; n-1, w_2),
\]
for any \( c_1, r_2 \in \{0, 1, \ldots, n-1\} \). And the second probability on the RHS of Equation (B.6) is
\[
\Pr \left( \sum_{j' \neq j} C_{j'k} = 0, \sum_{i' = 1}^{n} R_{i'k'} = 0, C_{jk'} = 1, \bar{Y}_{ijk} = 1, \bar{Y}_{ijk'} = 0 \right) = \sum_{c_k = 0}^{n} \sum_{c_k' = 0}^{r_k'} \sum_{r_k' = 0}^{c_k'} \Pr \left( \sum_{j' \neq j} \tilde{C}_{j'k} = c_k, \sum_{i' = 1}^{n} \tilde{R}_{i'k'} = r_k', \tilde{C}_{jk'} = c_k', \bar{Y}_{ijk} = 1, \bar{Y}_{ijk'} = 0 \right) \times \Sigma_{c \in k} \Sigma_{c \in k} n_{1}^{n-c_k-1} \Sigma_{c_k}^{n-r_k'} \Sigma_{c_k'}^{n-r_k'} \Sigma_{c_k}^{1-c_k'}.
\]

Denote \( \Pr(\sum_{j' \neq j} \tilde{C}_{j'k} = c_k, \sum_{i' = 1}^{c_k'} \tilde{R}_{i'k'} = r_k', \tilde{C}_{jk'} = c_k', \bar{Y}_{ijk} = 1, \bar{Y}_{ijk'} = 0) \) in the equation above as \( h_{c_k'}^{(7)}(n; k, c_k, r_k') \). For each \( k = 1, 2 \), we now provide
expressions for $h_{c_k'}(n; k, c_k, r_k')$ for any $c_k' \in \{0, 1\}$, $c_k \in \{0, 1, \ldots, n - 1\}$ and $r_k' \in \{c_k', c_k' + 1, \ldots, n\}$.

When $k = 1$,

$$h_0^{(7)}(n; 1, c_1, r_2) = p_{10} \frac{n-1}{n} B(n-1, c_1; n, r_2),$$

and

$$h_1^{(7)}(n; 1, c_1, r_2) = p_{10} \sum_{w_1=0}^{c_1} \sum_{w_2=1}^{\text{min}(n-1, r_2)} \frac{(n-1)\binom{n-w_1-1}{w_2} \binom{n-c_1-w_1}{w_2}}{\pi_2^{w_2} \pi_2^{n-w_2-1}} \times (1 - \pi_1^{w_2})^{c_1-w_1} \left(\frac{\pi_1^{w_2}}{\pi_1^{n-1}}\right)^{n-1-c_1} B(n-1, w_1; n-w_2, r_2 - w_2).$$

When $k = 2$,

$$h_0^{(7)}(n; 2, c_2, r_1) = p_{01} \frac{n-1}{n} B(n, r_1; n-1, c_2),$$

and

$$h_1^{(7)}(n; 2, c_2, r_1) = p_{01} \sum_{w_1=0}^{c_2} \sum_{w_2=1}^{\text{min}(n-1, r_1)} \frac{(n-1)\binom{n-w_1-1}{w_2} \binom{n-c_2-w_1}{w_2}}{\pi_1^{w_1} \pi_1^{n-w_1-1}} \times (1 - \pi_1^{w_1})^{c_2-w_2} \left(\frac{\pi_1^{w_1}}{\pi_1^{n-1}}\right)^{n-c_2-1} B(n-w_1, r_1 - w_1; n-1, w_2),$$

where $B(\cdot)$ is defined in Function 5.

This completes the derivation of $\text{PS}_{e,k}^{(AT)}$.

**Pooling Specificity $\text{PS}_{p,k}$**

Recall that pooling specificity for the $k$th infection, $\text{PS}_{p,k}$, is the probability that an individual is diagnosed as negative for the $k$th infection given it is truly negative. Under AT, denote the pooling specificity as $\text{PS}_{p,k}^{(AT)}$, it follows that

$$1 - \text{PS}_{p,k}^{(AT)} = \text{pr} \left( Y_{ijk} = 1, T_{ij1}^{(AT)} + T_{ij2}^{(AT)} \geq 1 \mid \tilde{Y}_{ijk} = 0 \right),$$

where

$$\text{pr} \left( Y_{ijk} = 1, T_{ij1}^{(AT)} + T_{ij2}^{(AT)} \geq 1 \mid \tilde{Y}_{ijk} = 0 \right) = \sum_{p,k} \pi_k^{-1} \left[ \text{EFF}(AT) - \frac{2}{n} - \pi_k S_{e,k}^{-1} \text{PS}_{e,k}^{(AT)} \right],$$

for $k = 1, 2$. 

122
### B.2 Derivations for operating characteristics of ATM

#### Efficiency

Recall that ATM starts with testing the master pool which mixes specimens of all individuals in the array. If the master pool test result is positive for at least one infection, then it proceeds as in AT; if the test result is negative for both infections, then all individuals are declared as negative. Let $M_k = 1$ denote the master pool tests positive for the $k$th infection; and 0 otherwise, for $k = 1, 2$. Then the class which contains individuals who need to be retested in the last stage can be mathematically expressed as

$$M_+ = \{ I_{ij} : T_{ij1}^{(ATM)} + T_{ij2}^{(ATM)} \geq 1 \},$$

where for $k = 1, 2$,

$$T_{ijk}^{(ATM)} = \begin{cases} 
1, & \text{if } M_1 + M_2 \geq 1, R_{ik} = 1, C_{jk} = 1; \\
1, & \text{if } M_1 + M_2 \geq 1, R_{ik} = 1, \sum_{j=1}^{n} C_{jk} = 0; \\
1, & \text{if } M_1 + M_2 \geq 1, C_{jk} = 1, \sum_{i=1}^{n} R_{ik} = 0; \\
0, & \text{otherwise.} 
\end{cases}$$

In other words, $T_{ijk}^{(ATM)}$ equals one if $M_1 + M_2 \geq 1$ and also $T_{ijk}^{(AT)} = 1$. For the ATM, the efficiency is

$$\text{EFF(ATM)} = \frac{1}{n^2} \left[ 1 + 2npr (M_1 + M_2 \geq 1) + n^2 pr \left( T_{ij1}^{(ATM)} + T_{ij2}^{(ATM)} \geq 1 \right) \right]. \quad (B.7)$$

To derive Equation (B.7), we again define some functions for notation simplicity.

**Function 6:** Let $f^-_b (s)$ denote the probability that a pool with size $s$ tests positively for at least one infection, without any knowledge of the true status. Then

$$f^-_b (s) = 1 - \eta_0^{(s)} S_p:1 S_p:2 - \eta_1^{(s)} \overline{S}_e:1 S_p:2 - \eta_2^{(s)} S_p:1 \overline{S}_e:2 - \eta_3^{(s)} \overline{S}_e:1 \overline{S}_e:2.$$
\textbf{Function 7:} For $k = 1, 2$, define
\[
\beta_k(n) = \Pr \left( R_{ik'} = 1, C_{jk'} = 1, \tilde{M}_k = 0 \right)
\]
\[
= \eta_k \cdot S_{e,k'}^2 \left( \pi_k \right)^{n^2 - 1} + p_{00} \left[ \eta_0^{(n-1)} \tilde{S}_{p,k'} + \eta_k^{(n-1)} S_{e,k'} \right]^2 \left( \pi_k \right)^{n^2 - 2n + 1},
\]
where $\tilde{M}_k = 1$ indicates the master pool contains at least one individual who is truly positive for the $k$th infection; and 0 otherwise.

\textbf{Function 8:} For $k = 1, 2$, define
\[
\gamma_k(n) = \Pr \left( R_{ik'} = 1, \sum_{j=1}^{n} C_{jk'} = 0, \tilde{M}_k = 0 \right)
\]
\[
= \sum_{r_{ik'}=0}^{1} \sum_{c_{ik'}=r_{ik'}}^{n} \Pr \left( \tilde{R}_{ik'} = r_{ik'}, \sum_{j=1}^{n} C_{jk'} = c_{ik'}, \tilde{M}_k = 0 \right) \tilde{S}_{e,k'}^{r_{ik'}-r_{ik'}} \tilde{S}_{p,k'}^{r_{ik'}} \tilde{S}_{e,k'}^{c_{ik'}-c_{ik'}},
\]
where for $c_{ik'} = \{r_{ik'}, r_{ik'} + 1, \ldots, n\}$,
\[
\Pr \left( \tilde{R}_{ik'} = 0, \sum_{j=1}^{n} C_{jk'} = c_{ik'}, \tilde{M}_k = 0 \right) = \eta_0^{(n)} \left( \eta_k^{(n-1)} \right)^{c_{ik'}} \left( \eta_0^{(n-1)} \right)^{n-c_{ik'}} \equiv \gamma_0(n; c_{ik'}),
\]
and
\[
\Pr \left( \tilde{R}_{ik'} = 1, \sum_{j=1}^{n} C_{jk'} = c_{ik'}, \tilde{M}_k = 0 \right)
\]
\[
= \binom{n}{c_{ik'}} \left( \eta_k^{(n)} \right)^{c_{ik'}} \left( \eta_0^{(n)} \right)^{n-c_{ik'}} - \gamma_0(n; c_{ik'}) \equiv \gamma_1(n; c_{ik'}).
\]

We now derive Equation (B.7). It is straightforward to see that the term $\Pr(M_1 + M_2 \geq 1)$ on the RHS of Equation (B.7) equals $f'_b(n^2)$, where $f'_b(\cdot)$ is given in Function 6. For the other term on the RHS of Equation (B.7), we get
\[
\Pr \left( T_{ij1}^{(ATM)} + T_{ij2}^{(ATM)} \geq 1 \right) = \sum_{m_1=0}^{1} \sum_{m_2=0}^{1} \Pr \left( M_1 + M_2 \geq 1 \big| \tilde{M}_1 = m_1, \tilde{M}_2 = m_2 \right)
\]
\[
\times \Pr \left( T_{ij1}^{(AT)} + T_{ij2}^{(AT)} \geq 1, \tilde{M}_1 = m_1, \tilde{M}_2 = m_2 \right),
\]
where for $m_1, m_2 \in \{0, 1\}$,
\[
\Pr \left( M_1 + M_2 \geq 1 \big| \tilde{M}_1 = m_1, \tilde{M}_2 = m_2 \right) = 1 - \sum_{s_{e,1}:p_{1}}^{m_1} \sum_{s_{e,2}:p_{2}}^{m_2} S_{e,1}^{-m_1} S_{e,2}^{-m_2} S_{e,2}^{-m_2}.\]
Denote \( \text{pr}(T^{(AT)}_{ij1} + T^{(AT)}_{ij2} \geq 1, \tilde{M}_1 = m_1, \tilde{M}_2 = m_2) \) in Equation (B.8) as \( h^{(8)}_{m_1m_2}(n) \).

For \( m_1, m_2 \in \{0, 1\} \), following similar arguments as in Appendix B.1, we have

\[
h^{(8)}_{00}(n) = \eta_0^{(n^2)} \left( \overline{S}_{p:1}^2 + 2 \overline{S}_{p:1} S_{p:1}^0 + \overline{S}_{p:2}^2 + 2 \overline{S}_{p:2} S_{p:2}^0 - \overline{S}_{p:1}^2 \overline{S}_{p:2} - 2 \overline{S}_{p:1} \overline{S}_{p:2} S_{p:2}^n - 2 \overline{S}_{p:2}^2 \overline{S}_{p:1}^n - 4 \overline{S}_{p:1} \overline{S}_{p:1} \overline{S}_{p:2} S_{p:2}^n \right),
\]

\[
h^{(8)}_{10}(n) = \beta_2(n) + 2 \gamma_2(n) + \eta_2 \left( \overline{S}_{p:2}^2 + 2 \overline{S}_{p:2} S_{p:2}^n \right) - \left( \overline{S}_{p:2}^2 \beta_2(n) + 2 \overline{S}_{p:2} S_{p:2}^n \beta_2(n) + 2 \overline{S}_{p:2}^2 \gamma_2(n) + 4 \overline{S}_{p:2} S_{p:2} \gamma_2(n) \right),
\]

\[
h^{(8)}_{01}(n) = \beta_1(n) + 2 \gamma_1(n) + \eta_1 \left( \overline{S}_{p:1}^2 + 2 \overline{S}_{p:1} S_{p:1}^n \right) - \left( \overline{S}_{p:1}^2 \beta_1(n) + 2 \overline{S}_{p:1} S_{p:1}^n \beta_1(n) + 2 \overline{S}_{p:1}^2 \gamma_1(n) + 4 \overline{S}_{p:1} S_{p:1} \gamma_1(n) \right),
\]

and

\[
h^{(8)}_{11}(n) = \text{pr}(T^{(AT)}_{ij1} + T^{(AT)}_{ij2} \geq 1) - h^{(8)}_{00}(n) - h^{(8)}_{10}(n) - h^{(8)}_{01}(n),
\]

where \( \text{pr}(T^{(AT)}_{ij1} + T^{(AT)}_{ij2} \geq 1) \) is derived in Appendix B.1, and \( \beta_k(\cdot) \) and \( \gamma_k(\cdot) \) are defined in Function 7 and 8 above. Thus, we can obtain the efficiency of ATM.

**Pooling Sensitivity \( \text{PS}_{e,k} \)**

Similar as \( \text{PS}^{(AT)}_{e,k} \), the pooling sensitivity of ATM for \( k \)th infection, \( \text{PS}^{(ATM)}_{e,k} \), is

\[
\text{PS}^{(ATM)}_{e,k} = \text{pr} \left( Y_{ijk} = 1, T^{(ATM)}_{ij1} + T^{(ATM)}_{ij2} \geq 1 \middle| \tilde{Y}_{ijk} = 1 \right)
= \sum_{m_{k'} = 0}^{1} \text{pr} \left( M_1 + M_2 \geq 1 \middle| \tilde{M}_{k'} = m_{k'}, \tilde{Y}_{ijk} = 1 \right)
\times \text{pr} \left( Y_{ijk} = 1, T^{(AT)}_{ij1} + T^{(AT)}_{ij2} \geq 1, \tilde{Y}_{ijk} = 1 \right). \quad (B.9)
\]

We now derive Equation (B.9). Firstly, it is straightforward to see that the term on the RHS of Equation (B.9), \( \text{pr}(M_1 + M_2 \geq 1 | \tilde{M}_{k'} = m_{k'}, \tilde{Y}_{ijk} = 1) \), equals 1 –
\( \overline{S}_{c,k}S_{e,k'}^m \overline{S}_{p,k'}^{1-m} \) for \( m_{k'} = 0, 1 \). Denote the other term in Equation (B.9), \( \text{pr}(Y_{ijk} = 1, T_{ijk}^{(AT)} + T_{ijk}^{(AT)} \geq 1, \overline{M}_{k'} = m_{k'} | Y_{ijk} = 1) \) as \( h^{(9)}_{m_{k'}}(n; k) \). Then when \( m_{k'} = 0 \),

\[
h^{(9)}_0(n; k) = \text{pr} \left( Y_{ijk} = 1, T_{ijk}^{(AT)} = 1, \overline{M}_{k'} = 0 \left| \bar{Y}_{ijk} = 1 \right. \right)
+ \text{pr} \left( Y_{ijk} = 1, T_{ijk}^{(AT)} = 1, \overline{M}_{k'} = 0 \left| \bar{Y}_{ijk} = 1 \right. \right)
- \text{pr} \left( Y_{ijk} = 1, T_{ijk}^{(AT)} = 1, T_{ijk}^{(AT)} = 1, \overline{M}_{k'} = 0 \left| \bar{Y}_{ijk} = 1 \right. \right),
\]

(B.10)

The first term on the RHS of Equation (B.10) is,

\[
\text{pr} \left( Y_{ijk} = 1, T_{ijk}^{(AT)} = 1, \overline{M}_{k'} = 0 \left| \bar{Y}_{ijk} = 1 \right. \right) = \pi_k^{-1} \eta_k (1) \pi_k^{n-1} \overline{S}_{c,k}^3
+ 2\pi_k^{-1} \eta_k (1) \pi_k^{n-1} \left[ \eta_k (n) \overline{S}_{e,k} + \eta_0 (n) S_{p,k} \right]^{n-1} S_{e,k}^2 \overline{S}_{e,k},
\]

and the second term on the RHS of Equation (B.10) is,

\[
\text{pr} \left( Y_{ijk} = 1, T_{ijk}^{(AT)} = 1, \overline{M}_{k'} = 0 \left| \bar{Y}_{ijk} = 1 \right. \right)
= \pi_k^{-1} \eta_k (1) \pi_k^{n-1} \left( \overline{S}_{c,k} S_{e,k}^2 + 2S_{e,k} \overline{S}_{p,k'} S_{p,k'}^n \right).
\]

Lastly, the third term on the RHS of Equation (B.10) is

\[
\text{pr} \left( Y_{ijk} = 1, T_{ijk}^{(AT)} = 1, T_{ijk}^{(AT)} = 1, \overline{M}_{k'} = 0 \left| \bar{Y}_{ijk} = 1 \right. \right)
= \pi_k^{-1} \eta_k (1) \pi_k^{n-1} S_{c,k}^3 \overline{S}_{p,k'} \left( \overline{S}_{p,k'} + 2S_{p,k'}^n \right)
+ 2\pi_k^{-1} \eta_k (1) \pi_k^{n-1} \left[ \eta_k (n) \overline{S}_{e,k} + \eta_0 (n) S_{p,k} \right]^{n-1} S_{e,k}^2 \overline{S}_{e,k} \overline{S}_{p,k'} \left[ \overline{S}_{p,k'} + 2S_{p,k'}^n \right].
\]

When \( m_{k'} = 1 \), \( h^{(9)}_1(n; k) = PS_{c,k}^{(AT)} - h^{(9)}_0(n; k) \), where \( PS_{c,k}^{(AT)} \) is derived in Appendix B.1 and \( h^{(9)}_0(n; k) \) is defined above, for \( k = 1, 2 \). This completes the derivation of \( PS_{c,k}^{(AT)} \).

**Pooling Specificity \( PS_{p,k} \)**

Similar as in Appendix B.1, the pooling specificity of ATM for the \( k \)th infection, denoted by \( PS_{p,k}^{(AT)} \), for \( k = 1, 2 \), is found by solving

\[
1 - PS_{p,k}^{(AT)} = \overline{S}_{p,k} \pi_k^{-1} \left[ \text{EFF}(ATM) - \frac{1}{n^2} - \frac{2}{n} f_b^- (n) - \pi_k S_{e,k}^{-1} PS_{e,k}^{(AT)} \right].
\]
where $f_k^-(n)$ is defined in Function 6, $\text{EFF(AM)}$ and $\text{PS}_{e,k}^{(ATM)}$ are derived in the first two sections of Appendix B.2.

### B.3 Additional tables from Section 3.4 and Section 3.5

**Table B.1:** Values of $\text{PS}_{p.k}$ and $\text{NPV}_k$ for the cases identified in Table 3.1 when $S_{e,k} = 0.95$ and $S_{p,j} = 0.99$. These are calculated from expressions in Tebbs et al. (2013) and Section B.1.

<table>
<thead>
<tr>
<th>Case</th>
<th>Cell probabilities</th>
<th>$\mathcal{A}$</th>
<th>$n^*$</th>
<th>$\text{PS}_{p.1}$</th>
<th>$\text{PS}_{p.2}$</th>
<th>$\text{NPV}_1$</th>
<th>$\text{NPV}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>$p_{00} = 0.870$, $p_{10} = 0.123$ $p_{01} = 0.004$, $p_{11} = 0.003$</td>
<td>IPP 4</td>
<td>0.997</td>
<td>0.996</td>
<td>0.986</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT 6</td>
<td>0.998</td>
<td>0.997</td>
<td>0.981</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>(ii)</td>
<td>$p_{00} = 0.890$, $p_{10} = 0.104$ $p_{01} = 0.003$, $p_{11} = 0.003$</td>
<td>IPP 4</td>
<td>0.997</td>
<td>0.996</td>
<td>0.989</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT 7</td>
<td>0.998</td>
<td>0.997</td>
<td>0.984</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>(iii)</td>
<td>$p_{00} = 0.918$, $p_{10} = 0.078$ $p_{01} = 0.002$, $p_{11} = 0.002$</td>
<td>IPP 4</td>
<td>0.998</td>
<td>0.997</td>
<td>0.992</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT 8</td>
<td>0.998</td>
<td>0.997</td>
<td>0.988</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>(iv)</td>
<td>$p_{00} = 0.930$, $p_{10} = 0.066$ $p_{01} = 0.003$, $p_{11} = 0.001$</td>
<td>IPP 5</td>
<td>0.997</td>
<td>0.997</td>
<td>0.993</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT 9</td>
<td>0.998</td>
<td>0.998</td>
<td>0.990</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>(v)</td>
<td>$p_{00} = 0.950$, $p_{10} = 0.047$ $p_{01} = 0.002$, $p_{11} = 0.001$</td>
<td>IPP 5</td>
<td>0.998</td>
<td>0.998</td>
<td>0.995</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT 10</td>
<td>0.999</td>
<td>0.998</td>
<td>0.993</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>(vi)</td>
<td>$p_{00} = 0.970$, $p_{10} = 0.028$ $p_{01} = 0.001$, $p_{11} = 0.001$</td>
<td>IPP 7</td>
<td>0.998</td>
<td>0.998</td>
<td>0.997</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT 14</td>
<td>0.999</td>
<td>0.999</td>
<td>0.996</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>
Table B.2: Values of PS$_{p,k}$ and NPV$_k$ for the cases identified in Table 3.2 when $S_{c,k} = 0.95$ and $S_{p,j} = 0.99$. These are calculated from expressions in Tebbs et al. (2013) and Section B.1.

<table>
<thead>
<tr>
<th>Case</th>
<th>Cell probabilities</th>
<th>A</th>
<th>n*</th>
<th>PS$_{p,1}$</th>
<th>PS$_{p,2}$</th>
<th>NPV$_1$</th>
<th>NPV$_2$</th>
</tr>
</thead>
</table>
| (i)  | $p_{00} = 0.870$, $p_{10} = 0.064$  
      |       | IPP 4 | 0.996 | 0.996 | 0.994 | 0.994 |
|      | $p_{01} = 0.063$, $p_{11} = 0.003$  
      |       | AT 7 | 0.997 | 0.997 | 0.991 | 0.991 |
| (ii) | $p_{00} = 0.890$, $p_{10} = 0.054$  
      |       | IPP 4 | 0.997 | 0.997 | 0.995 | 0.995 |
|      | $p_{01} = 0.053$, $p_{11} = 0.003$  
      |       | AT 8 | 0.998 | 0.998 | 0.992 | 0.993 |
| (iii)| $p_{00} = 0.918$, $p_{10} = 0.040$  
      |       | IPP 4 | 0.997 | 0.997 | 0.996 | 0.996 |
|      | $p_{01} = 0.040$, $p_{11} = 0.002$  
      |       | AT 9 | 0.998 | 0.998 | 0.994 | 0.994 |
| (iv) | $p_{00} = 0.930$, $p_{10} = 0.034$  
      |       | IPP 5 | 0.997 | 0.997 | 0.997 | 0.997 |
|      | $p_{01} = 0.034$, $p_{11} = 0.002$  
      |       | AT 10 | 0.998 | 0.998 | 0.995 | 0.995 |
| (v)  | $p_{00} = 0.950$, $p_{10} = 0.025$  
      |       | IPP 5 | 0.998 | 0.998 | 0.998 | 0.998 |
|      | $p_{01} = 0.024$, $p_{11} = 0.001$  
      |       | AT 12 | 0.999 | 0.999 | 0.996 | 0.997 |
| (vi) | $p_{00} = 0.970$, $p_{10} = 0.015$  
      |       | IPP 7 | 0.998 | 0.998 | 0.999 | 0.999 |
|      | $p_{01} = 0.014$, $p_{11} = 0.001$  
      |       | AT 16 | 0.999 | 0.999 | 0.998 | 0.998 |

Table B.3: Values of PS$_{p,k}$ and NPV$_k$ for the cases identified in Table 3.3 when $S_{c,k} = 0.95$ and $S_{p,j} = 0.99$. These are calculated from expressions in Tebbs et al. (2013), Section B.1 and Section B.2.

<table>
<thead>
<tr>
<th>Case</th>
<th>Cell probabilities</th>
<th>A</th>
<th>n*</th>
<th>PS$_{p,1}$</th>
<th>PS$_{p,2}$</th>
<th>NPV$_1$</th>
<th>NPV$_2$</th>
</tr>
</thead>
</table>
| (i)  | $p_{00} = 0.9900$, $p_{10} = 0.0048$  
      |       | IPP 11 | 1.000 | 1.000 | 0.999 | 0.999 |
|      | $p_{01} = 0.0048$, $p_{11} = 0.0004$  
      |       | AT 20 | 1.000 | 1.000 | 0.999 | 0.999 |
|      |       | ATM 5 | 1.000 | 1.000 | 0.999 | 0.999 |
| (ii) | $p_{00} = 0.9930$, $p_{10} = 0.0033$  
      |       | IPP 13 | 0.999 | 0.999 | 1.000 | 1.000 |
|      | $p_{01} = 0.0034$, $p_{11} = 0.0003$  
      |       | AT 20 | 1.000 | 1.000 | 1.000 | 1.000 |
|      |       | ATM 6 | 1.000 | 1.000 | 1.000 | 1.000 |
| (iii)| $p_{00} = 0.9960$, $p_{10} = 0.0019$  
      |       | IPP 17 | 0.999 | 0.999 | 1.000 | 1.000 |
|      | $p_{01} = 0.0019$, $p_{11} = 0.0002$  
      |       | AT 20 | 1.000 | 1.000 | 1.000 | 1.000 |
|      |       | ATM 7 | 1.000 | 1.000 | 1.000 | 1.000 |
| (iv) | $p_{00} = 0.9990$, $p_{10} = 0.0004$  
      |       | IPP 33 | 1.000 | 1.000 | 1.000 | 1.000 |
|      | $p_{01} = 0.0005$, $p_{11} = 0.0001$  
      |       | AT 20 | 1.000 | 1.000 | 1.000 | 1.000 |
|      |       | ATM 10 | 1.000 | 1.000 | 1.000 | 1.000 |
Table B.4: Classification accuracy measures for 2011 chlamydia and gonorrhea data collected in Region X IPP. For each state/specimen stratum, the sample size $N$, and the classification accuracy measures, which are averaged over $B = 1000$ replications, are included. The optimal configuration $n^*$, the averaged number of tests, and the savings compared to the optimal IPP algorithm is given in Table 3.5.

<table>
<thead>
<tr>
<th>State</th>
<th>Specimen</th>
<th>Algorithm</th>
<th>PS_{e1}</th>
<th>PS_{e2}</th>
<th>PS_{p1}</th>
<th>PS_{p2}</th>
<th>PPV_1</th>
<th>PPV_2</th>
<th>NPV_1</th>
<th>NPV_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>Swab</td>
<td>IPP</td>
<td>0.890</td>
<td>0.988</td>
<td>0.993</td>
<td>0.996</td>
<td>0.909</td>
<td>0.565</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>0.842</td>
<td>0.988</td>
<td>0.994</td>
<td>0.997</td>
<td>0.918</td>
<td>0.606</td>
<td>0.989</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>0.811</td>
<td>0.987</td>
<td>0.995</td>
<td>0.997</td>
<td>0.919</td>
<td>0.611</td>
<td>0.986</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Urine</td>
<td>IPP</td>
<td>0.898</td>
<td>0.869</td>
<td>0.997</td>
<td>0.998</td>
<td>0.969</td>
<td>0.688</td>
<td>0.990</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>0.854</td>
<td>0.885</td>
<td>0.997</td>
<td>0.998</td>
<td>0.970</td>
<td>0.711</td>
<td>0.986</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>0.821</td>
<td>0.882</td>
<td>0.997</td>
<td>0.998</td>
<td>0.970</td>
<td>0.717</td>
<td>0.982</td>
<td>0.999</td>
</tr>
<tr>
<td>Idaho</td>
<td>Swab</td>
<td>IPP</td>
<td>0.889</td>
<td>0.900</td>
<td>0.993</td>
<td>0.996</td>
<td>0.913</td>
<td>0.378</td>
<td>0.991</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>0.841</td>
<td>0.991</td>
<td>0.995</td>
<td>0.997</td>
<td>0.928</td>
<td>0.432</td>
<td>0.988</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>0.802</td>
<td>0.991</td>
<td>0.995</td>
<td>0.997</td>
<td>0.928</td>
<td>0.441</td>
<td>0.985</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Urine</td>
<td>IPP</td>
<td>0.899</td>
<td>0.879</td>
<td>0.997</td>
<td>0.998</td>
<td>0.961</td>
<td>0.531</td>
<td>0.991</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>0.853</td>
<td>0.896</td>
<td>0.997</td>
<td>0.998</td>
<td>0.966</td>
<td>0.572</td>
<td>0.987</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>0.818</td>
<td>0.889</td>
<td>0.997</td>
<td>0.998</td>
<td>0.966</td>
<td>0.581</td>
<td>0.984</td>
<td>1.000</td>
</tr>
<tr>
<td>Oregon</td>
<td>Swab</td>
<td>IPP</td>
<td>0.890</td>
<td>0.988</td>
<td>0.994</td>
<td>0.996</td>
<td>0.910</td>
<td>0.420</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>0.841</td>
<td>0.989</td>
<td>0.995</td>
<td>0.997</td>
<td>0.929</td>
<td>0.488</td>
<td>0.989</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>0.803</td>
<td>0.989</td>
<td>0.996</td>
<td>0.997</td>
<td>0.930</td>
<td>0.498</td>
<td>0.986</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Urine</td>
<td>IPP</td>
<td>0.898</td>
<td>0.878</td>
<td>0.998</td>
<td>0.998</td>
<td>0.968</td>
<td>0.641</td>
<td>0.991</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>0.854</td>
<td>0.900</td>
<td>0.998</td>
<td>0.998</td>
<td>0.971</td>
<td>0.678</td>
<td>0.988</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>0.820</td>
<td>0.892</td>
<td>0.998</td>
<td>0.998</td>
<td>0.971</td>
<td>0.686</td>
<td>0.985</td>
<td>1.000</td>
</tr>
<tr>
<td>Washington</td>
<td>Swab</td>
<td>IPP</td>
<td>0.890</td>
<td>0.987</td>
<td>0.994</td>
<td>0.996</td>
<td>0.906</td>
<td>0.523</td>
<td>0.993</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>0.840</td>
<td>0.988</td>
<td>0.995</td>
<td>0.997</td>
<td>0.920</td>
<td>0.576</td>
<td>0.989</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>0.809</td>
<td>0.988</td>
<td>0.995</td>
<td>0.997</td>
<td>0.921</td>
<td>0.585</td>
<td>0.987</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Urine</td>
<td>IPP</td>
<td>0.899</td>
<td>0.882</td>
<td>0.997</td>
<td>0.998</td>
<td>0.960</td>
<td>0.546</td>
<td>0.992</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT</td>
<td>0.853</td>
<td>0.895</td>
<td>0.998</td>
<td>0.998</td>
<td>0.968</td>
<td>0.606</td>
<td>0.989</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATM</td>
<td>0.817</td>
<td>0.890</td>
<td>0.998</td>
<td>0.998</td>
<td>0.968</td>
<td>0.614</td>
<td>0.986</td>
<td>1.000</td>
</tr>
</tbody>
</table>