Supplementary Material for "Group testing case identification with biomarker information"

Dewei Wang[1], Christopher S. McMahan[2], Joshua M. Tebbs[1,*], and Christopher R. Bilder[3]
[1]Department of Statistics, University of South Carolina, Columbia, SC 29208, U.S.A.
[2]Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, U.S.A.
[3]Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, U.S.A.
[*]*email*: tebbs@stat.sc.edu

Web Appendix A. *Efficiency derivation from Section 3.1.* The efficiency of an S-stage hierarchical algorithm $H(n_1 : n_2 : \cdots : n_S)$ is given by

$$\text{EFF} = \frac{1}{n_1} + \sum_{s=1}^{S-1} \frac{1}{n_{s+1}} \text{pr}\left(\prod_{s'=1}^{s} Z_{\mathcal{P}_{s'1}} = 1\right), \tag{A.1}$$

an expression known in the case identification literature (see, Kim et al., 2007 and the references therein). Under classical assumptions; i.e., (a) the sensitivity and specificity are unaffected by pool size and (b) testing outcomes on pools containing common individuals are independent conditional on the true pool statuses, the probability in Equation (A.1) is

$$\text{pr}\left(\prod_{s'=1}^{s} Z_{\mathcal{P}_{s'1}} = 1\right) = q^{n_1}(1-S_p)^s + \sum_{s'=1}^{s-1}(q^{n_{s'}+1} - q^{n_{s'}})S_e^{s'}(1-S_p)^{s-s'} + (1-q^{n_s})S_e^s;$$

see Kim et al. (2007). Within our biomarker framework, calculating this probability is much more difficult. For any pool $\mathcal{P}_{s'1}$, recall that the corresponding testing response $Z_{\mathcal{P}_{s'1}}$ is related to the measured biomarker level $\mathcal{C}_{\mathcal{P}_{s'1}}$ through $Z_{\mathcal{P}_{s'1}} = I(\mathcal{C}_{\mathcal{P}_{s'1}} > \tau_{\mathcal{P}_{s'1}})$. Therefore,

$$\text{pr}\left(\prod_{s'=1}^{s} Z_{\mathcal{P}_{s'1}} = 1\right) = \text{pr}(\mathcal{C}_{\mathcal{P}_{s1}} > \tau_{\mathcal{P}_{s1}}, \mathcal{C}_{\mathcal{P}_{s-1,1}} > \tau_{\mathcal{P}_{s-1,1}}, ..., \mathcal{C}_{\mathcal{P}_{11}} > \tau_{\mathcal{P}_{11}}). \tag{A.2}$$

The density function of $\mathcal{C}_{\mathcal{P}_{s'1}}$ depends on the number of positive individuals in $\mathcal{P}_{s'1}$. Because of the nested structure $\mathcal{P}_{s1} \subset \mathcal{P}_{s-1,1} \subset \cdots \subset \mathcal{P}_{11}$ in a hierarchical algorithm, we can calculate the probability in Equation (A.2) by going through all possible numbers of positive individuals that could be contained in $\mathcal{P}_{s1}$, $\mathcal{P}_{s-1,1} \setminus \mathcal{P}_{s1}$, ..., $\mathcal{P}_{11} \setminus \mathcal{P}_{21}$. This probability equals

$$\sum_{m_s=0}^{n_s} \sum_{m_{s-1}=0}^{n_{s-1}-n_s} \cdots \sum_{m_1=0}^{n_1-n_2} \text{pr}\left(\mathcal{C}_{\mathcal{P}_{s1}} > \tau_{\mathcal{P}_{s1}}, \mathcal{C}_{\mathcal{P}_{s-1,1}} > \tau_{\mathcal{P}_{s-1,1}}, ..., \mathcal{C}_{\mathcal{P}_{11}} > \tau_{\mathcal{P}_{11}} \middle| \sum_{i\in\mathcal{P}_{s1}} T_i = m_s, \right.$$

$$\left. \sum_{i\in\mathcal{P}_{s-1,1}\setminus\mathcal{P}_{s1}} T_i = m_{s-1}, ..., \sum_{i\in\mathcal{P}_{11}\setminus\mathcal{P}_{21}} T_i = m_1 \right)$$

$$\times \text{pr}\left(\sum_{i\in\mathcal{P}_{s1}} T_i = m_s\right) \text{pr}\left(\sum_{i\in\mathcal{P}_{s-1,1}\setminus\mathcal{P}_{s1}} T_i = m_{s-1}\right) \cdots \text{pr}\left(\sum_{i\in\mathcal{P}_{11}\setminus\mathcal{P}_{21}} T_i = m_1\right). \tag{A.3}$$

Conditioning on the event $\{\sum_{i\in\mathcal{P}_{s1}} T_i = m_s, \sum_{i\in\mathcal{P}_{s-1,1}\setminus\mathcal{P}_{s1}} T_i = m_{s-1}, ..., \sum_{i\in\mathcal{P}_{11}\setminus\mathcal{P}_{21}} T_i = m_1\}$; i.e., there are m_s positive individuals in $\mathcal{P}_{s1}$ and $\sum_{s''=s'}^{s} m_{s''}$ positive individuals in $\mathcal{P}_{s'1}$ for $s' = s-1, s-2, ..., 1$, we denote the joint probability density function of $(\mathcal{C}_{\mathcal{P}_{s1}}, \mathcal{C}_{\mathcal{P}_{s-1,1}}, ..., \mathcal{C}_{\mathcal{P}_{11}})'$ by $f_{\mathcal{C}_{\mathcal{P}_{s1}},\mathcal{C}_{\mathcal{P}_{s-1,1}},...,\mathcal{C}_{\mathcal{P}_{11}}}(u_s, u_{s-1}, ..., u_1|m_s, m_{s-1}, ..., m_1)$, which can be written as

$$\int_{\mathbb{R}^{n_1}} \prod_{s'=1}^{s} f_\epsilon\left(u_{s'}\middle|\frac{1}{n_{s'}}\sum_{i=1}^{n_{s'}} v_i\right) \prod_{i=1}^{m_s} f_{\widetilde{\mathcal{C}}^+}(v_i) \prod_{i=m_s+1}^{n_s} f_{\widetilde{\mathcal{C}}^-}(v_i) \prod_{i=n_s+1}^{n_s+m_{s-1}} f_{\widetilde{\mathcal{C}}^+}(v_i) \prod_{i=n_s+m_{s-1}+1}^{n_{s-1}} f_{\widetilde{\mathcal{C}}^-}(v_i)$$
$$\cdots \prod_{i=n_2+1}^{n_2+m_1} f_{\widetilde{\mathcal{C}}^+}(v_i) \prod_{i=n_2+m_1+1}^{n_1} f_{\widetilde{\mathcal{C}}^-}(v_i) dv_1 dv_2 \cdots dv_{n_1},$$

where, as in the manuscript, it is understood that products like $\prod_{i=a}^{b} f_{\widetilde{\mathcal{C}}^+}(v_i)$ and $\prod_{i=a}^{b} f_{\widetilde{\mathcal{C}}^-}(v_i)$, $a > b$, are vacuous. Consequently, the conditional probability in (A.3) is equal to

$$\int_{\tau_{\mathcal{P}_{11}}}^{\infty}\int_{\tau_{\mathcal{P}_{21}}}^{\infty}\cdots\int_{\tau_{\mathcal{P}_{s1}}}^{\infty} f_{\mathcal{C}_{\mathcal{P}_{s1}},\mathcal{C}_{\mathcal{P}_{s-1,1}},...,\mathcal{C}_{\mathcal{P}_{11}}}(u_s, u_{s-1}, ..., u_1|m_s, m_{s-1}, ..., m_1) du_s du_{s-1} \cdots du_1.$$

The s unconditional probabilities in (A.3) are $\text{pr}(\sum_{i\in\mathcal{P}_{s1}} T_i = m_s) = \binom{n_s}{m_s} p^{m_s} q^{n_s - m_s}$ and

$$\text{pr}\left(\sum_{i\in\mathcal{P}_{s'1}\setminus\mathcal{P}_{s'+1,1}} T_i = m_{s'}\right) = \binom{n_{s'} - n_{s'+1}}{m_{s'}} p^{m_{s'}} q^{n_{s'} - n_{s'+1} - m_{s'}}$$

for $s' = 1, 2, ..., s-1$. This completes the derivation of $\text{EFF}\{H(n_1 : n_2 : \cdots : n_S)\}$.

Web Appendix B. *Derivations of PSE and PSP from Section 3.2; closed-form calculations under normality.* In an S-stage hierarchical algorithm $H(n_1 : n_2 : \cdots : n_S)$, the pooling sensitivity and pooling specificity are given by $\text{PSE} = \text{pr}(\prod_{s=1}^{S} Z_{\mathcal{P}_{s1}} = 1|T_1 = 1)$ and $\text{PSP} = 1 - \text{pr}(\prod_{s=1}^{S} Z_{\mathcal{P}_{s1}} = 1|T_1 = 0)$, respectively. For $m_S \in \{0, 1\}$, note that we can write $\text{pr}(\prod_{s=1}^{S} Z_{\mathcal{P}_{s1}} = 1|T_1 = m_S)$ as

$$\sum_{m_{S-1}=0}^{n_{S-1}-n_S} \cdots \sum_{m_1=0}^{n_1-n_2} \text{pr}\left(\mathcal{C}_{\mathcal{P}_{S1}} > \tau_{\mathcal{P}_{S1}}, \mathcal{C}_{\mathcal{P}_{S-1,1}} > \tau_{\mathcal{P}_{S-1,1}}, ..., \mathcal{C}_{\mathcal{P}_{11}} > \tau_{\mathcal{P}_{11}} \middle| \sum_{i\in\mathcal{P}_{S1}} T_i = m_S, \right.$$
$$\left. \sum_{i\in\mathcal{P}_{S-1,1}\setminus\mathcal{P}_{S1}} T_i = m_{S-1}, ..., \sum_{i\in\mathcal{P}_{11}\setminus\mathcal{P}_{21}} T_i = m_1\right)$$
$$\times \text{pr}\left(\sum_{i\in\mathcal{P}_{S-1,1}\setminus\mathcal{P}_{S1}} T_i = m_{S-1}\right) \cdots \text{pr}\left(\sum_{i\in\mathcal{P}_{11}\setminus\mathcal{P}_{21}} T_i = m_1\right).$$

The conditional probability above is a special case of the conditional probability in (A.3). The unconditional probabilities are the same binomial calculations as in Web Appendix A.

Simulation details: In an S-stage hierarchical algorithm $H(n_1 : n_2 : \cdots : n_S)$, the efficiency (EFF), the pooling sensitivity (PSE), and the pooling specificity (PSP) can be calculated

exactly when $f_{\widetilde{\mathcal{C}}^+}$, $f_{\widetilde{\mathcal{C}}^-}$, and f_ϵ are all normal densities. Otherwise, the integral expressions in Web Appendix A are very likely to be intractable. Instead of implementing high-dimensional numerical integration, we use Monte Carlo simulation to estimate EFF, PSE, and PSP. Estimates of PPV and NPV in Section 3.2 can then be constructed easily.

Estimating PSE and PSP in an S-stage hierarchical algorithm $H(n_1 : n_2 : \cdots : n_S)$ involves a simple modification to our simulation algorithm presented in Section 3.1 in the manuscript:

SIMULATION PROCEDURE TO ESTIMATE PSE AND PSP

1. Set $T_1 = 1(0)$ if estimating PSE (PSP). Generate $T_2, T_3, ..., T_{n_1} \sim$ iid Bernoulli(p). Generate $\widetilde{\mathcal{C}}_i \sim f_{\widetilde{\mathcal{C}}_i|T_i=t}(u) = tf_{\widetilde{\mathcal{C}}^+}(u) + (1-t)f_{\widetilde{\mathcal{C}}^-}(u)$, $i = 1, 2, ..., n_1$.

2. (Stage 1). Calculate $\widetilde{\mathcal{C}}_{\mathcal{P}_{11}} = n_1^{-1} \sum_{i \in \mathcal{P}_{11}} \widetilde{\mathcal{C}}_i$ and generate $\mathcal{C}_{\mathcal{P}_{11}}$ from $f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{11}})$.

   (a) If $Z_{\mathcal{P}_{11}} = I(\mathcal{C}_{\mathcal{P}_{11}} > \tau_{\mathcal{P}_{11}}) = 0$, stop and classify the first individual in $\mathcal{P}_{11}$ as negative.

   (b) If $Z_{\mathcal{P}_{11}} = I(\mathcal{C}_{\mathcal{P}_{11}} > \tau_{\mathcal{P}_{11}}) = 1$, continue to the next stage.

3. (Stage 2). Calculate $\widetilde{\mathcal{C}}_{\mathcal{P}_{21}} = n_2^{-1} \sum_{i \in \mathcal{P}_{21}} \widetilde{\mathcal{C}}_i$ and generate $\mathcal{C}_{\mathcal{P}_{21}}$ from $f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{21}})$.

   (a) If $Z_{\mathcal{P}_{21}} = I(\mathcal{C}_{\mathcal{P}_{21}} > \tau_{\mathcal{P}_{21}}) = 0$, stop and classify the first individual in $\mathcal{P}_{11}$ as negative.

   (b) If $Z_{\mathcal{P}_{21}} = I(\mathcal{C}_{\mathcal{P}_{11}} > \tau_{\mathcal{P}_{21}}) = 1$, continue to the next stage.

4. (Stage 3). Calculate $\widetilde{\mathcal{C}}_{\mathcal{P}_{31}} = n_3^{-1} \sum_{i \in \mathcal{P}_{31}} \widetilde{\mathcal{C}}_i$, generate $\mathcal{C}_{\mathcal{P}_{31}}$ from $f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{31}})$, and calculate $Z_{\mathcal{P}_{31}} = I(\mathcal{C}_{\mathcal{P}_{31}} > \tau_{\mathcal{P}_{31}})$. Continue this overall process until individual testing is performed in stage S.

5. (Stage S). Set $\widetilde{\mathcal{C}}_{\mathcal{P}_{S1}} = \widetilde{\mathcal{C}}_1$, generate $\mathcal{C}_{\mathcal{P}_{S1}}$ from $f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{S1}})$. If $Z_{\mathcal{P}_{S1}} = I(\mathcal{C}_{\mathcal{P}_{S1}} > \tau_{\mathcal{P}_{S1}}) = 0$, stop and classify the first individual in $\mathcal{P}_{11}$ as negative; otherwise, stop and classify the first individual in $\mathcal{P}_{11}$ as positive.

We implement this procedure B times and estimate PSE and PSP using

$$\widehat{\text{PSE}} = \frac{1}{B} \sum_{b=1}^{B} D_b,$$

when $T_1 \overset{\text{set}}{=} 1$ and

$$\widehat{\text{PSP}} = \frac{1}{B} \sum_{b=1}^{B} (1 - D_b),$$

when $T_1 \overset{\text{set}}{=} 0$, respectively. In these expressions, $D_b = 1(0)$ if the first individual in $\mathcal{P}_{11}$ is classified as positive (negative) in the bth implementation, $b = 1, 2, ..., B$.

Exact calculations under normality: We show how to calculate EFF, PSE, and PSP exactly for an S-stage hierarchical algorithm $H(n_1 : n_2 : \cdots : n_S)$ under normal biomarker

and measurement error assumptions. Suppose $\widetilde{\mathcal{C}}^+ \sim \mathcal{N}(\mu_+, \sigma_+^2)$, $\widetilde{\mathcal{C}}^- \sim \mathcal{N}(\mu_-, \sigma_-^2)$, and $\mathcal{C}|\widetilde{\mathcal{C}} \sim \mathcal{N}(\widetilde{\mathcal{C}}, \sigma_\epsilon^2)$. For individuals in the master pool $\mathcal{P}_{11}$, collect the true biomarker levels and the true disease statuses into $\widetilde{\boldsymbol{\mathcal{C}}} = (\widetilde{\mathcal{C}}_1, \widetilde{\mathcal{C}}_2, ..., \widetilde{\mathcal{C}}_{n_1})'$ and $\mathbf{T} = (T_1, T_2, ..., T_{n_1})'$, respectively, so that $\widetilde{\boldsymbol{\mathcal{C}}}|\mathbf{T} = \mathbf{t} \sim \mathcal{N}_{n_1}(\boldsymbol{\mu}(\mathbf{t}), \boldsymbol{\Sigma}(\mathbf{t}))$, where $\mathbf{t}$ is a binary vector, $\boldsymbol{\mu}(\mathbf{t})$ is an $n_1 \times 1$ vector whose ith component is $t_i\mu_+ + (1-t_i)\mu_-$, and $\boldsymbol{\Sigma}(\mathbf{t})$ is an $n_1 \times n_1$ diagonal matrix with ith diagonal element equal to $t_i\sigma_+^2 + (1-t_i)\sigma_-^2$. For $s = 1, 2, ..., S$, let $\widetilde{\boldsymbol{\mathcal{C}}}_s = (\widetilde{\mathcal{C}}_{\mathcal{P}_{11}}, \widetilde{\mathcal{C}}_{\mathcal{P}_{21}}, ..., \widetilde{\mathcal{C}}_{\mathcal{P}_{s1}})'$ and $\boldsymbol{\mathcal{C}}_s = (\mathcal{C}_{\mathcal{P}_{11}}, \mathcal{C}_{\mathcal{P}_{21}}, ..., \mathcal{C}_{\mathcal{P}_{s1}})'$ denote the true and measured biomarker levels of $\mathcal{P}_{11}, \mathcal{P}_{21}, ..., \mathcal{P}_{s1}$, respectively. We now derive the distribution of $\boldsymbol{\mathcal{C}}_s|\mathbf{T} = \mathbf{t}$. Let $\mathbf{d}_s$ be a vector whose first n_s elements are $1/n_s$ and whose remaining elements are 0. Set $\mathbf{D}_s = (\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_s)$. Then we can write $\widetilde{\mathcal{C}}_{\mathcal{P}_{sl}} = \mathbf{d}_s'\widetilde{\boldsymbol{\mathcal{C}}}$ and $\widetilde{\boldsymbol{\mathcal{C}}}_s = \mathbf{D}_s'\widetilde{\boldsymbol{\mathcal{C}}}$. Thus, $\widetilde{\boldsymbol{\mathcal{C}}}_s|\mathbf{T} = \mathbf{t} \sim \mathcal{N}_s(\mathbf{D}_s'\boldsymbol{\mu}(\mathbf{t}), \mathbf{D}_s'\boldsymbol{\Sigma}(\mathbf{t})\mathbf{D}_s)$ and

$$\boldsymbol{\mathcal{C}}_s|\mathbf{T} = \mathbf{t} \sim \mathcal{N}_s(\mathbf{D}_s'\boldsymbol{\mu}(\mathbf{t}), \mathbf{D}_s'\boldsymbol{\Sigma}(\mathbf{t})\mathbf{D}_s + \sigma_\epsilon^2\mathbf{I}_s), \tag{B.1}$$

where $\mathbf{I}_s$ is the s-dimensional identity matrix.

Calculating the efficiency in an S-stage hierarchical algorithm $H(n_1 : n_2 : \cdots : n_S)$, in general, is not possible because of the difficult conditional probability in Equation (A.3); i.e.,

$$\text{pr}\left(\mathcal{C}_{\mathcal{P}_{s1}} > \tau_{\mathcal{P}_{s1}}, \mathcal{C}_{\mathcal{P}_{s-1,1}} > \tau_{\mathcal{P}_{s-1,1}}, ..., \mathcal{C}_{\mathcal{P}_{11}} > \tau_{\mathcal{P}_{11}} \middle| \sum_{i\in\mathcal{P}_{s1}} T_i = m_s, \sum_{i\in\mathcal{P}_{s-1,1}\setminus\mathcal{P}_{s1}} T_i = m_{s-1}, ..., \sum_{i\in\mathcal{P}_{11}\setminus\mathcal{P}_{21}} T_i = m_1\right). \tag{B.2}$$

However, under normal biomarker and measurement error assumptions, the probability in (B.2) can be calculated using the multivariate normal distribution in (B.1). After setting

$$\mathbf{t}_s = (\mathbf{1}_{m_s}', \mathbf{0}_{n_s-m_s}', \mathbf{1}_{m_{s-1}}', \mathbf{0}_{n_{s-1}-n_s-m_{s-1}}', ..., \mathbf{1}_{m_1}', \mathbf{0}_{n_1-n_2-m_1}')',$$

where $\mathbf{1}_a$ ($\mathbf{0}_a$) is an a-dimensional vector of ones (zeros), one can integrate the conditional density in (B.1) over the set $(\tau_{\mathcal{P}_{s1}}, \infty) \times (\tau_{\mathcal{P}_{s-1,1}}, \infty) \times \cdots \times (\tau_{\mathcal{P}_{11}}, \infty)$. This can be done easily using the R package `mvtnorm`. Because PSE and PSP involve the same type of conditional probability as in (B.2), these can be calculated exactly as well. Our R programs available at `www.chrisbilder.com/grouptesting` will calculate EFF, PSE, and PSP exactly under normal biomarker and measurement error assumptions. In non-normal biomarker and/or measurement error applications, our programs use simulation to estimate these quantities.

As a small example, we have calculated EFF, PSE, and PSP exactly as described above and also using the simulation approach described in Section 3.1 in the manuscript when $p = 0.10$ for $H(4:1)$ and $H(6:3:1)$, the most efficient two- and three-stage hierarchial algorithms under classical assumptions when $p = 0.10$. Our biomarker and measurement error assumptions are the same as in Figure 2 in the manuscript; i.e., $\widetilde{\mathcal{C}}^- \sim \mathcal{N}(3, 0.25)$, $\widetilde{\mathcal{C}}^+ \sim \mathcal{N}(6, 1)$, and $\mathcal{C}|\widetilde{\mathcal{C}} \sim \mathcal{N}(\widetilde{\mathcal{C}}, 0.0025)$. Table B.1 (next page) contains the exact calculations and the estimated calculations using our simulation algorithms with $B = 1{,}000{,}000$ Monte Carlo data sets. Exact and estimated values are close as one might expect.

Table B.1: Exact and simulated operating characteristics when $p = 0.10$ for two- and three-stage hierarchical algorithms. A description of these calculations is given on the previous page.

	$H(4:1)$		$H(6:3:1)$	
	Exact	Simulated	Exact	Simulated
EFF	0.613	0.613	0.587	0.589
PSE	0.900	0.900	0.860	0.862
PSP	0.992	0.992	0.994	0.994
PPV	0.928	0.927	0.940	0.937
NPV	0.989	0.989	0.985	0.985

Web Appendix C. *Two-dimensional array testing.* We extend our simulation methodology to estimate the operating characteristics of two-dimensional array testing as described in Section 3.3 in the manuscript. A two-dimensional array testing algorithm with R rows and C columns is denoted by $A(R \times C)$.

SIMULATION PROCEDURE TO ESTIMATE EFF$\{A(R \times C)\}$

1. Generate $\{T_{r,c} : r = 1, 2, ..., R,\ c = 1, 2, ..., C\} \sim$ iid Bernoulli(p). Generate $\widetilde{\mathcal{C}}_{r,c} \sim f_{\widetilde{\mathcal{C}}_{r,c}|T_{r,c}=t}(u) = tf_{\widetilde{\mathcal{C}}^+}(u) + (1-t)f_{\widetilde{\mathcal{C}}^-}(u)$, $r = 1, 2, ..., R,\ c = 1, 2, ..., C$.

2. (Stage 1). Calculate $\widetilde{\mathcal{C}}_{\mathcal{P}_{r+}} = C^{-1}\sum_{c=1}^{C}\widetilde{\mathcal{C}}_{r,c}$, generate $\mathcal{C}_{\mathcal{P}_{r+}}$ from $f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{r+}})$, and compute $Z_{\mathcal{P}_{r+}} = I(\mathcal{C}_{\mathcal{P}_{r+}} > \tau_{\mathcal{P}_{r+}})$, $r = 1, 2, ..., R$. Calculate $\widetilde{\mathcal{C}}_{\mathcal{P}_{+c}} = R^{-1}\sum_{r=1}^{R}\widetilde{\mathcal{C}}_{r,c}$, generate $\mathcal{C}_{\mathcal{P}_{+c}}$ from $f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{+c}})$, and compute $Z_{\mathcal{P}_{+c}} = I(\mathcal{C}_{\mathcal{P}_{+c}} > \tau_{\mathcal{P}_{+c}})$, $c = 1, 2, ..., C$.

   (a) If the set

$$\mathcal{M} = \left\{(r,c) : Z_{\mathcal{P}_{r+}} = Z_{\mathcal{P}_{+c}} = 1 \text{ or } Z_{\mathcal{P}_{r+}} = 1,\ \sum_{c'=1}^{C} Z_{\mathcal{P}_{+c'}} = 0 \right.$$
$$\left. \text{or } \sum_{r'=1}^{R} Z_{\mathcal{P}_{r'+}} = 0,\ Z_{\mathcal{P}_{+c}} = 1\right\}$$

   is empty, stop and classify all individuals in the array as negative.

   (b) If $\mathcal{M}$ is not empty, classify all individuals not in $\mathcal{M}$ as negative and continue to the next stage.

3. (Stage 2). For each $(r,c) \in \mathcal{M}$, generate $\mathcal{C}_{r,c}$ from $f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{r,c})$.

   (a) If $Z_{r,c} = I(\mathcal{C}_{r,c} > \tau) = 0$, classify the (r,c)th individual in the array as negative.

   (b) If $Z_{r,c} = I(\mathcal{C}_{r,c} > \tau) = 1$, classify the (r,c)th individual in the array as positive.

We implement this procedure B times and estimate the efficiency of $A(R \times C)$ using

$$\widehat{\text{EFF}} = \frac{1}{RCB} \sum_{b=1}^{B} M_b$$

where M_b is the number of tests observed in the bth replication. The variance of the number of tests per individual, which we denote by $\text{var}\{A(R \times C)\}$, can be estimated using the sample variance of $M_1/RC, M_2/RC, ..., M_B/RC$. Unlike hierarchical algorithms, $\text{EFF}\{A(R \times C)\}$ cannot be calculated exactly even under normality. We estimate the pooling sensitivity (PSE) and pooling specificity (PSP) of $A(R \times C)$ using the following simulation procedure:

SIMULATION PROCEDURE TO ESTIMATE PSE AND PSP

1. Set $T_{1,1} = 1(0)$ if estimating PSE (PSP). Generate

   $$\{T_{r,c} : r = 1, 2, ..., R,\ c = 1, 2, ..., C,\ (r, c) \neq (1, 1)\} \sim \text{iid Bernoulli}(p).$$

   Generate $\widetilde{\mathcal{C}}_{r,c} \sim f_{\widetilde{\mathcal{C}}_{r,c}|T_{r,c}=t}(u) = tf_{\widetilde{\mathcal{C}}^+}(u) + (1-t)f_{\widetilde{\mathcal{C}}^-}(u)$, $r = 1, 2, ..., R,\ c = 1, 2, ..., C$.

2. (Stage 1). Calculate $\widetilde{\mathcal{C}}_{\mathcal{P}_{r+}} = C^{-1} \sum_{c=1}^{C} \widetilde{\mathcal{C}}_{r,c}$, generate $\mathcal{C}_{\mathcal{P}_{r+}}$ from $f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{r+}})$, and compute $Z_{\mathcal{P}_{r+}} = I(\mathcal{C}_{\mathcal{P}_{r+}} > \tau_{\mathcal{P}_{r+}})$, $r = 1, 2, ..., R$. Calculate $\widetilde{\mathcal{C}}_{\mathcal{P}_{+c}} = R^{-1} \sum_{r=1}^{R} \widetilde{\mathcal{C}}_{r,c}$, generate $\mathcal{C}_{\mathcal{P}_{+c}}$ from $f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{\mathcal{P}_{+c}})$, and compute $Z_{\mathcal{P}_{+c}} = I(\mathcal{C}_{\mathcal{P}_{+c}} > \tau_{\mathcal{P}_{+c}})$, $c = 1, 2, ..., C$.

   (a) If $Z_{\mathcal{P}_{1+}} = Z_{\mathcal{P}_{+1}} = 1$, or $Z_{\mathcal{P}_{1+}} = 1, \sum_{c'=1}^{C} Z_{\mathcal{P}_{+c'}} = 0$, or $\sum_{r'=1}^{R} Z_{\mathcal{P}_{r'+}} = 0, Z_{\mathcal{P}_{+1}} = 1$, continue to the next stage.

   (b) Otherwise, stop and classify the $(1, 1)$th individual in the array as negative.

3. (Stage 2). Generate $\mathcal{C}_{1,1}$ from $f_\epsilon(\cdot|\widetilde{\mathcal{C}}_{1,1})$. Classify the $(1, 1)$th individual in the array as positive if $Z_{1,1} = I(\mathcal{C}_{1,1} > \tau) = 1$; otherwise, classify the $(1, 1)$th individual as negative.

We implement this procedure B times and estimate PSE and PSP using

$$\widehat{\text{PSE}} = \frac{1}{B} \sum_{b=1}^{B} D_b,$$

when $T_{1,1} \overset{\text{set}}{=} 1$ and

$$\widehat{\text{PSP}} = \frac{1}{B} \sum_{b=1}^{B} (1 - D_b),$$

when $T_{1,1} \overset{\text{set}}{=} 0$, respectively. In these expressions, $D_b = 1(0)$ if the $(1, 1)$th individual is classified as positive (negative) in the bth implementation, $b = 1, 2, ..., B$.

Web Appendix D. *Additional tables in Section 5.* The following tables accompany Table 1 in the manuscript for the Wein and Zenios (1996) and Zenios and Wein (1998) application and the May et al. (2010) application in Section 5. Table D.1 is constructed for $p = 0.01$; Table D.2 is constructed for $p = 0.10$.

Table D.1: Operating characteristics for Application 1 (Zenios and Wein, 1998) and Application 2 (May et al., 2010) when $p = 0.01$. Efficiency (EFF), standard deviation of the number of tests per individual (SD), and accuracy probabilities (PSE, PSP, PPV, and NPV) are provided. The threshold τ^* maximizes Youden's index for individual testing. The threshold $\tau^*_{\mathcal{P}}$ is calculated as in Section 4. Classical operating characteristics are calculated exactly from Kim et al. (2007). Biomarker-based characteristics are estimated using $B = 1{,}000{,}000$ Monte Carlo data sets. For each application, individual testing values of S_e and S_p are provided.

			Biomarker-based evaluations			
			τ^*	τ^*/pool size	$\tau^*_{\mathcal{P}}$	Classical
Application 1 $S_e > 0.999; S_p > 0.999$	$H(11:1)$	EFF (SD)	0.191 (0.301)	0.808 (0.450)	0.196 (0.307)	0.196 (0.306)
		PSE	0.962	0.999	0.984	>0.999
		PSP	>0.999	>0.999	>0.999	>0.999
		PPV	>0.999	>0.999	>0.999	>0.999
		NPV	>0.999	>0.999	>0.999	>0.999
	$H(14:7:1)$	EFF (SD)	0.152 (0.219)	0.646 (0.441)	0.156 (0.223)	0.158 (0.226)
		PSE	0.936	0.998	0.973	>0.999
		PSP	>0.999	>0.999	>0.999	>0.999
		PPV	>0.999	>0.999	>0.999	>0.999
		NPV	>0.999	>0.999	>0.999	>0.999
	$A(25 \times 25)$	EFF (SD)	0.120 (0.030)	0.763 (0.091)	0.135 (0.036)	0.135 (0.038)
		PSE	0.779	0.991	0.910	>0.999
		PSP	>0.999	>0.999	>0.999	>0.999
		PPV	>0.999	>0.999	>0.999	>0.999
		NPV	0.998	>0.999	>0.999	>0.999
Application 2 $S_e = 0.989; S_p = 0.980$	$H(11:1)$	EFF (SD)	0.131 (0.196)	0.756 (0.472)	0.261 (0.376)	0.212 (0.326)
		PSE	0.393	0.988	0.901	0.978
		PSP	>0.999	0.981	0.993	0.98
		PPV	0.837	0.343	0.574	0.820
		NPV	0.994	>0.999	>0.999	>0.999
	$H(16:8:1)$	EFF (SD)	0.091 (0.134)	0.629 (0.374)	0.181 (0.257)	0.160 (0.232)
		PSE	0.310	0.988	0.843	0.966
		PSP	>0.999	0.982	0.996	0.999
		PPV	0.874	0.353	0.705	0.881
		NPV	0.993	>0.999	0.998	>0.999
	$A(24 \times 24)$	EFF (SD)	0.092 (0.013)	1.068 (0.025)	0.165 (0.046)	0.141 (0.039)
		PSE	0.225	0.989	0.767	0.967
		PSP	>0.999	0.981	0.997	>0.999
		PPV	0.940	0.339	0.720	0.913
		NPV	0.992	>0.999	0.998	>0.999

Table D.2: Operating characteristics for Application 1 (Zenios and Wein, 1998) and Application 2 (May et al., 2010) when $p = 0.10$. Efficiency (EFF), standard deviation of the number of tests per individual (SD), and accuracy probabilities (PSE, PSP, PPV, and NPV) are provided. The threshold τ^* maximizes Youden's index for individual testing. The threshold $\tau^*_{\mathcal{P}}$ is calculated as in Section 4. Classical operating characteristics are calculated exactly from Kim et al. (2007). Biomarker-based characteristics are estimated using $B = 1{,}000{,}000$ Monte Carlo data sets. For each application, individual testing values of S_e and S_p are provided.

			Biomarker-based evaluations			
			τ^*	τ^*/pool size	$\tau^*_{\mathcal{P}}$	Classical
Application 1 $S_e > 0.999; S_p > 0.999$	$H(4:1)$	EFF (SD)	0.593 (0.475)	0.821 (0.495)	0.594 (0.475)	0.594 (0.475)
		PSE	0.998	>0.999	0.999	>0.999
		PSP	>0.999	>0.999	>0.999	>0.999
		PPV	>0.999	>0.999	>0.999	>0.999
		NPV	>0.999	>0.999	>0.999	>0.999
	$H(6:3:1)$	EFF (SD)	0.591 (0.472)	0.780 (0.460)	0.592 (0.472)	0.594 (0.472)
		PSE	0.994	>0.999	0.997	>0.999
		PSP	>0.999	>0.999	>0.999	>0.999
		PPV	>0.999	>0.999	>0.999	>0.999
		NPV	>0.999	>0.999	>0.999	>0.999
	$A(7 \times 7)$	EFF (SD)	0.578 (0.178)	0.928 (0.185)	0.581 (0.179)	0.583 (0.180)
		PSE	0.986	>0.999	0.994	>0.999
		PSP	>0.999	>0.999	>0.999	>0.999
		PPV	>0.999	>0.999	>0.999	>0.999
		NPV	0.999	>0.999	>0.999	>0.999
Application 2 $S_e = 0.989; S_p = 0.980$	$H(4:1)$	EFF (SD)	0.487 (0.425)	0.692 (0.497)	0.614 (0.481)	0.603 (0.478)
		PSE	0.725	0.987	0.963	0.978
		PSP	0.996	0.994	0.990	0.994
		PPV	0.952	0.950	0.913	0.950
		NPV	0.970	0.998	0.996	0.998
	$H(6:3:1)$	EFF (SD)	0.424 (0.419)	0.703 (0.474)	0.600 (0.476)	0.594 (0.470)
		PSE	0.626	0.985	0.944	0.968
		PSP	0.998	0.986	0.993	0.996
		PPV	0.966	0.887	0.935	0.966
		NPV	0.960	0.998	0.994	0.996
	$A(7 \times 7)$	EFF (SD)	0.408 (0.107)	0.774 (0.203)	0.587 (0.177)	0.586 (0.179)
		PSE	0.533	0.988	0.926	0.968
		PSP	0.998	0.982	0.992	0.995
		PPV	0.970	0.859	0.930	0.960
		NPV	0.951	0.999	0.992	0.996

Web Appendix E. *Sensitivity analysis described in Section 6.* We illustrate how to perform a sensitivity analysis to investigate the impact of model misspecification. Recall that our simulation-based approach to calculate the operating characteristics of case identification algorithms requires one to specify $f_{\widetilde{\mathcal{C}}^+}$, $f_{\widetilde{\mathcal{C}}^-}$, and f_ϵ. It is therefore natural to wonder what happens when one or more of these distributions is misspecified.

Performing a sensitivity analysis within our framework is easy, because Monte Carlo simulation is used to incorporate biomarker and measurement error information. To illustrate, suppose

$$\begin{aligned} \widetilde{\mathcal{C}}^+ &\sim \mathcal{G}(10, 20) \\ \widetilde{\mathcal{C}}^- &\sim \mathcal{G}(1, 0.25) \\ \mathcal{C}_\mathcal{P}|\widetilde{\mathcal{C}}_\mathcal{P} &\sim \mathcal{N}(\widetilde{\mathcal{C}}_\mathcal{P}, 0.01), \end{aligned}$$

where $\mathcal{G}(\mu, \sigma^2)$ denotes a two-parameter gamma distribution with mean μ and variance σ^2. Note that the model for $\mathcal{C}_\mathcal{P}|\widetilde{\mathcal{C}}_\mathcal{P}$ assumes additive mean-zero error. For this collection of distributions, the threshold that maximizes Youden's index for individual testing is $\tau^* = 2.671$, which provides values of $S_e = 0.988$ and $S_p = 0.994$.

To examine the impact of model misspecification, suppose $\widetilde{\mathcal{C}}^+$ is wrongly specified to have a $\mathcal{G}(10 - \delta, 20)$ distribution where $\delta \geq 0$. We consider values $\delta \in \{0, 0.1, 0.2, ..., 1.9, 2\}$ and assume that the models for $\widetilde{\mathcal{C}}^-$ and $\mathcal{C}_\mathcal{P}|\widetilde{\mathcal{C}}_\mathcal{P}$ are correct (obviously, these could be misspecified too). Note that when $\delta = 0$, there is no model misspecification, and, as δ increases, the violation becomes more severe. For example, when $\delta = 2$, $S_e = 0.948$ and $S_p = 0.982$.

In this illustration, we consider Dorfman's two-stage hierarchical algorithm $H(5:1)$ when the population prevalence is $p = 0.05$. Similar results would be expected for other algorithms and for other values of p. For each value of δ, we first calculated our Youden-index type threshold for pools $\tau^*_\mathcal{P}$, defined in Section 4. We then estimated the expected number of tests per individual (EFF) and the standard deviation of the number of tests per individual (SD) for $H(5:1)$ using the simulation procedure described in Section 3.1 with $B = 1{,}000{,}000$ Monte Carlo data sets. The pooling sensitivity (PSE) and pooling specificity (PSP) were estimated using the simulation procedure in Web Appendix B, and predictive values PPV and NPV were estimated using the formulas in Section 3.2.

Figure E.1 (next page) shows the results of our sensitivity analysis. Recall that our goal is to assess the impact of model misspecification (i.e., when $\delta > 0$) on the operating characteristics EFF, SD, and the four classification accuracy probabilities. The horizontal lines in each subfigure denote the values under no misspecification; i.e., when $\delta = 0$. One notes that the values of EFF and SD are largely unaffected by model misspecification in this example. In addition, PSP and NPV are largely unaffected, likely because the distribution for $\widetilde{\mathcal{C}}^-$ is not misspecified. The biggest impact is seen in the values of PSE and PPV, which decrease as model misspecification becomes more severe.

Our research web site `www.chrisbilder.com/grouptesting` contains the R programs we used to perform this sensitivity study. Our programs can be easily altered to consider other case identification algorithms, other values of p, and other choices for the biomarker and measurement error distributions.

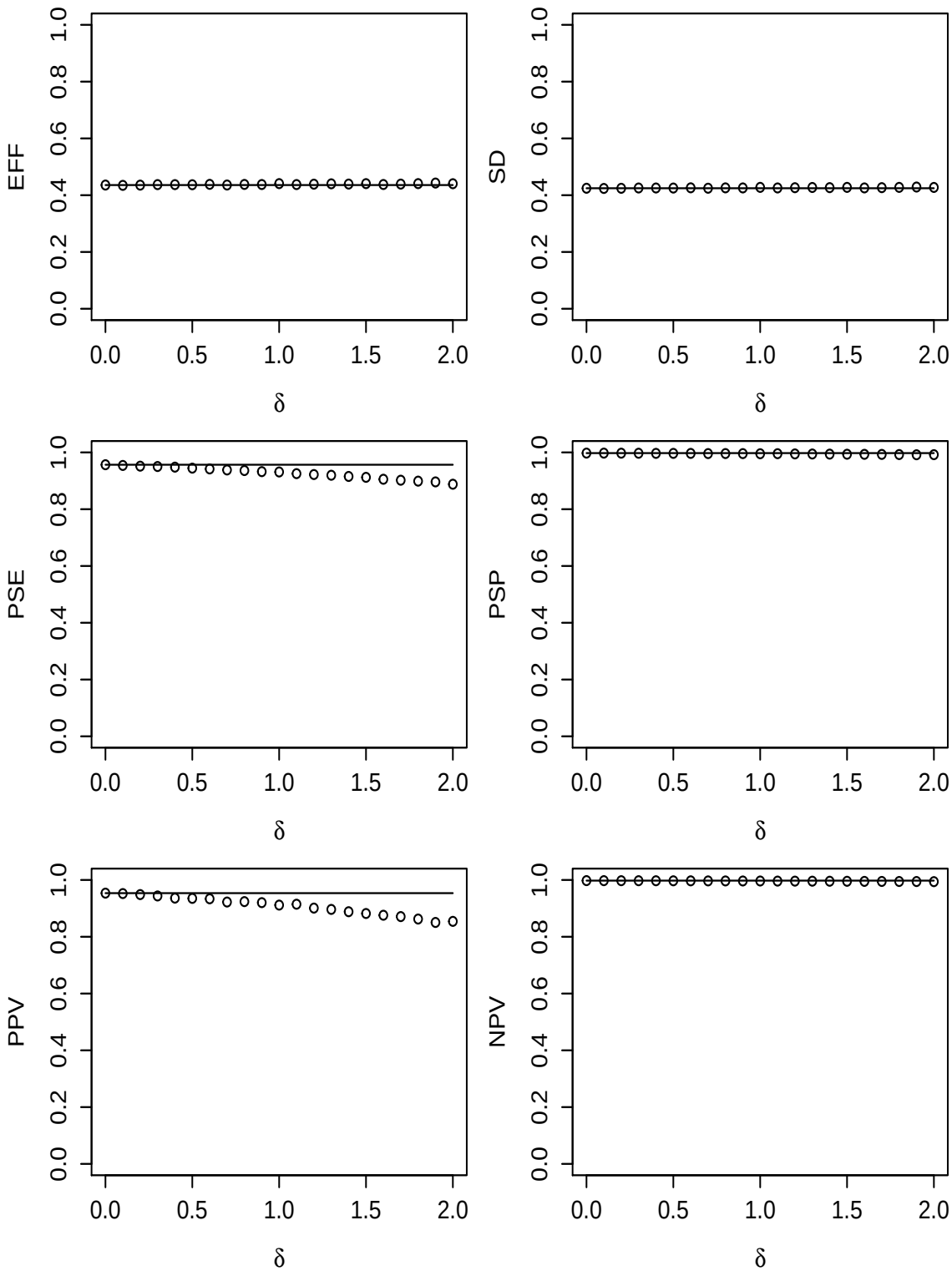

Figure E.1: Sensitivity analysis results (see description on the last page). Estimated operating characteristics of $H(5:1)$ when $p = 0.05$. Values of $\delta > 0$ correspond to model misspecification for $\widetilde{\mathcal{C}}^{+}$.