Web-based Supplementary Materials for

"Group Testing for Case Identification with Correlated Responses"

Samuel D. Lendle, Michael G. Hudgens*, and Bahjat F. Qaqish

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599, U.S.A.

* email: mhudgens@bios.unc.edu

Web Appendix A

Proof of Lemma 1.

For $n = 1, 2, \ldots$, let $\{\sigma_n\}$ be a sequence of real numbers with $0 \leqslant \sigma_n < 1$ for all n such that σ_n converges to 1, i.e. $\lim_{n\to\infty} \sigma_n = 1$. For $n = 1, 2, \ldots$, let $\dot{X}_n$ be the sum of m exchangeable binary random variables, each with mean p and with pairwise correlation σ_n. Let $Z_n = \dot{X}_n/m$ so $E(Z_n) = p$, $\text{var}(Z_n) = m^{-2}\text{var}(\dot{X}_n) = m^{-1}p(1-p)\{1 + (m-1)\sigma_n\}$, and $\lim_{n\to\infty} \text{var}(Z_n) = p - p^2$, implying $\lim_{n\to\infty} E(Z_n^2) = p$.

Let $A_n = \{0 < Z_n < 1\}$, $\text{pr}(A_n) = \alpha_n$, $\text{pr}(Z_n = 1) = \beta_n$, $E(Z_n \mid A_n) = \mu_n$ and $E(Z_n^2 \mid A_n) = \nu_n$. For all n, $E(Z_n) = E(Z_n \mid A_n)\text{pr}(A_n) + \text{pr}(Z_n = 1) = \mu_n\alpha_n + \beta_n = p$. For all n,

$$\mu_n - \nu_n = \sum_{i=1}^{m-1} \left[\left\{ \frac{i(m-i)}{m^2} \right\} \text{pr}(Z_n = i/m \mid A_n) \right] \geqslant \frac{m-1}{m^2},$$

so $\nu_n \leqslant \mu_n - (m-1)/m^2$. Let $c = (m-1)/m^2$. This implies $E(Z_n^2) = E(Z_n^2 \mid A_n)\text{pr}(A_n) + \text{pr}(Z_n = 1) = \nu_n\alpha_n + \beta_n \leqslant (\mu_n - c)\alpha_n + \beta_n = \mu_n\alpha_n + \beta_n - c\alpha_n = p - c\alpha_n$. Because c is a positive constant and $\lim_{n\to\infty} E(Z_n^2) = p$, $\lim_{n\to\infty} \alpha_n = 0$. Therefore, $\lim_{n\to\infty} \text{pr}(0 < \dot{X}_n < m) = 0$.

For all n, $\mu_n \leqslant (m-1)/m < 1$ so $\lim_{n\to\infty} \mu_n\alpha_n = 0$. Since $\mu_n\alpha_n + \beta_n = p$, it follows $\lim_{n\to\infty} \beta_n = p$. This implies $\lim_{n\to\infty} \text{pr}(Z_n = 0) = 1 - p$, so $\lim_{n\to\infty} \text{pr}(\dot{X}_n = 0) = 1 - p$ and $\lim_{n\to\infty} \text{pr}(\dot{X}_n = m) = p$.

Proof of Lemma 2.

$$
\begin{aligned}
\mathrm{pr}(\dot{X}' = \dot{x}') &= \sum_{\dot{x}=\dot{x}'}^{m-(m'-\dot{x}')} \mathrm{pr}(\dot{X}' = \dot{x}', \dot{X} = \dot{x}) \\
&= \sum_{\dot{x}=\dot{x}'}^{m-(m'-\dot{x}')} \mathrm{pr}(\dot{X} = \dot{x})\mathrm{pr}(\dot{X}' = \dot{x}' \mid \dot{X} = \dot{x}) \\
&= \sum_{\dot{x}=\dot{x}'}^{m-(m'-\dot{x}')} E_\pi \left\{ \binom{m}{\dot{x}} \pi^{\dot{x}}(1-\pi)^{m-\dot{x}} \right\} \frac{\binom{m'}{\dot{x}'}\binom{m-m'}{\dot{x}-\dot{x}'}}{\binom{m}{\dot{x}}} \\
&= E_\pi \left\{ \binom{m'}{\dot{x}'} \pi^{\dot{x}'}(1-\pi)^{m'-\dot{x}'} \times \sum_{\dot{x}=\dot{x}'}^{m-(m'-\dot{x}')} \left(\binom{m-m'}{\dot{x}-\dot{x}'} \pi^{\dot{x}-\dot{x}'}(1-\pi)^{m-m'-(\dot{x}-\dot{x}')} \right) \right\} \\
&= E_\pi \left\{ \binom{m'}{\dot{x}'} \pi^{\dot{x}'}(1-\pi)^{m'-\dot{x}'} \times \sum_{\dot{x}=0}^{m-m'} \left(\binom{m-m'}{\dot{x}} \pi^{\dot{x}}(1-\pi)^{m-m'-\dot{x}} \right) \right\} \\
&= E_\pi \left\{ \binom{m'}{\dot{x}'} \pi^{\dot{x}'}(1-\pi)^{m'-\dot{x}'} \right\}
\end{aligned}
$$

Proof of Lemma 3. Suppose $m' = 1$, so $E(\pi) = \mathrm{pr}(\dot{X}' = 1) = p$. When $m' = 1$, $\mathrm{pr}(\dot{X}' = 1) = \mathrm{pr}(X_i = 1) = E(X_i)$ for all i, so $E(X_i) = p$ by Lemma 2. Suppose $m' = 2$, so $E(\pi^2) = \mathrm{pr}(\dot{X}' = 2) = \sigma p(1-p) + p^2$. For all $i \neq j$, $E(X_i X_j) = \mathrm{pr}(X_i = 1, X_j = 1) = \mathrm{pr}(\dot{X}' = 2)$ and

$$
\begin{aligned}
\mathrm{cor}(X_i, X_j) &= \frac{E(X_i X_j) - E(X_i)E(X_j)}{\sqrt{\mathrm{var}(X_i)\mathrm{var}(X_j)}} \\
&= \frac{\sigma p(1-p) + p^2 - p^2}{p(1-p)} \\
&= \sigma
\end{aligned}
$$

Web Appendix B

In this section we elaborate on three key assumptions made in the main paper regarding test error. For further discussion of these assumptions see Kim et al. (2007) and Kim and Hudgens (2009).

I. Given a pool contains at least one positive unit, the probability the pool tests positive equals S_e, where S_e is the test sensitivity.

II. Given a pool contains no positive units, the probability the pool tests negative equals S_p, where S_p is the test specificity.

III. Conditional on the true status of a pool, the test result for that pool is independent of the true status and test result of any other pool.

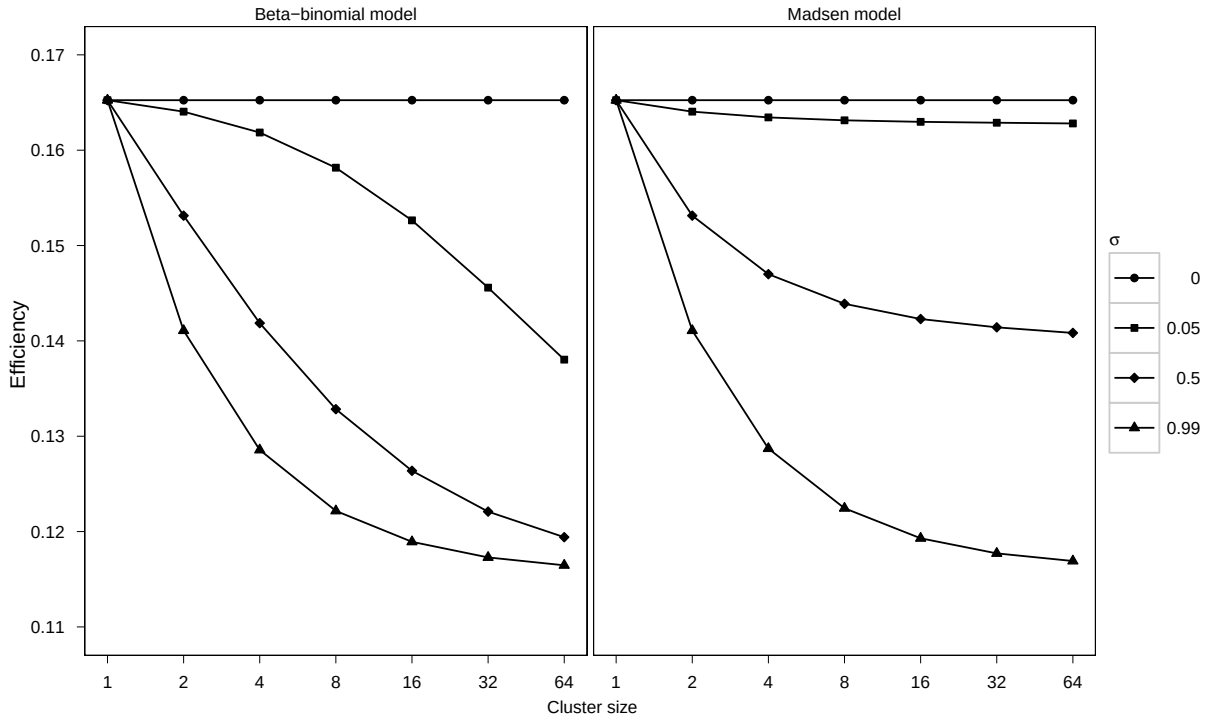
Assumption I implies that the test sensitivity is independent of the number of specimens composing a pool and of the number of positive specimens within a pool. Models that allow for sensitivity to depend on pool size (e.g., to account for possible dilution effects) are not considered. In light of assumption I, the results in the main paper can be viewed as appropriate in settings where the largest pool sizes are small enough that appreciable dilution effects are unlikely. Similarly, assumption II implies test specificity is independent of pool size. This assumption would be dubious in settings where there is synergism or additive effects (Xie et al. 2001), i.e., where two negative units may produce a false positive result when placed in the same pool. Extensions of the results in the main paper relaxing assumptions I and II that allow for sensitivity and specificity to depend on pool size (e.g., as in Johnson et al. 1991) should be straightforward and could be considered for future research.

Web Appendix C

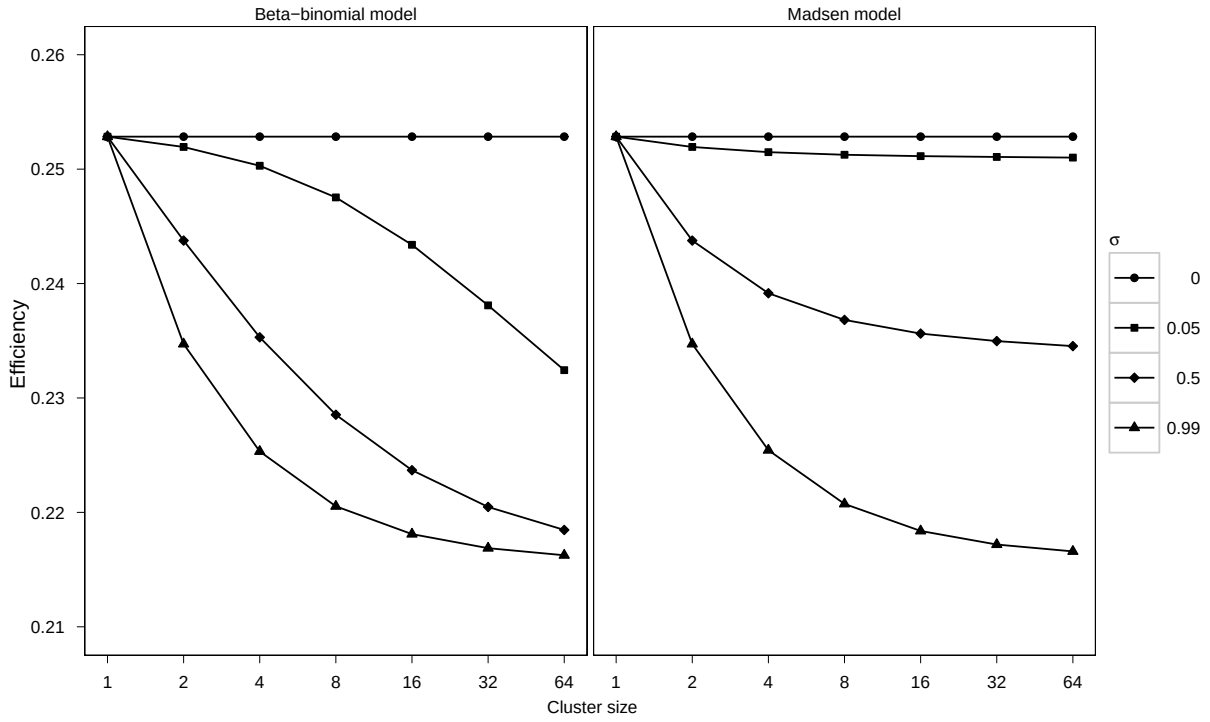
In this section we describe an investigation regarding the robustness of the efficiencies derived in the main paper when Assumption 2 (i.e., units within a cluster are exchangeable) does not hold. A simulation study was conducted where clusters of units had an auto-regressive (AR) correlation structure. In particular, using the method described in Section 2.2 of Lunn and Davies (1998), clusters were simulated such that for a cluster of m ordered units $X_1, \ldots, X_m$ the correlation between any two units was $\text{cor}(X_i, X_j) = \theta^{|i-j|}$ for $i, j \in \{1, \ldots, m\}$. Efficiencies of the three stage nested hierarchical procedure (as in Figure 3 of the main paper) and the rectangular matrix algorithm (as in Figure 4 of the main paper) when clusters have an AR correlation structure were calculated empirically via simulation. The empirical AR efficiencies were compared with the efficiency expected if it was assumed (incorrectly) that units within a cluster had an exchangeable correlation structure with correlation σ equal to the average true correlation between units, i.e., $\sigma = \sum_{i=1}^{m-1}(m-i)\theta^i / \binom{m}{2}$.

The simulation results are given in Web Figures 9 and 10. These results show that assuming an exchangeable correlation structure can yield efficiency estimates that are relatively close to the true AR efficiencies. For example, looking at Web Figure 10, for a 1×16 rectangular matrix procedure and average correlation between units of 0.2, the AR efficiency equals 0.29 whereas the efficiency assuming an exchangeable correlation structure equals 0.30. In contrast, a naive approach that assumes no correlation between units would estimate the efficiency to be 0.45. For the 4×4 rectangular matrix procedure the true AR efficiency and the efficiency assuming an exchangeable correlation structure are nearly identical (black and gray dashed lines in Web Figure 10). For the three stage hierarchical procedure efficiencies presented in Web Figure 9 the approximation assuming exchangeable correlation is less accurate although the bias is modest.

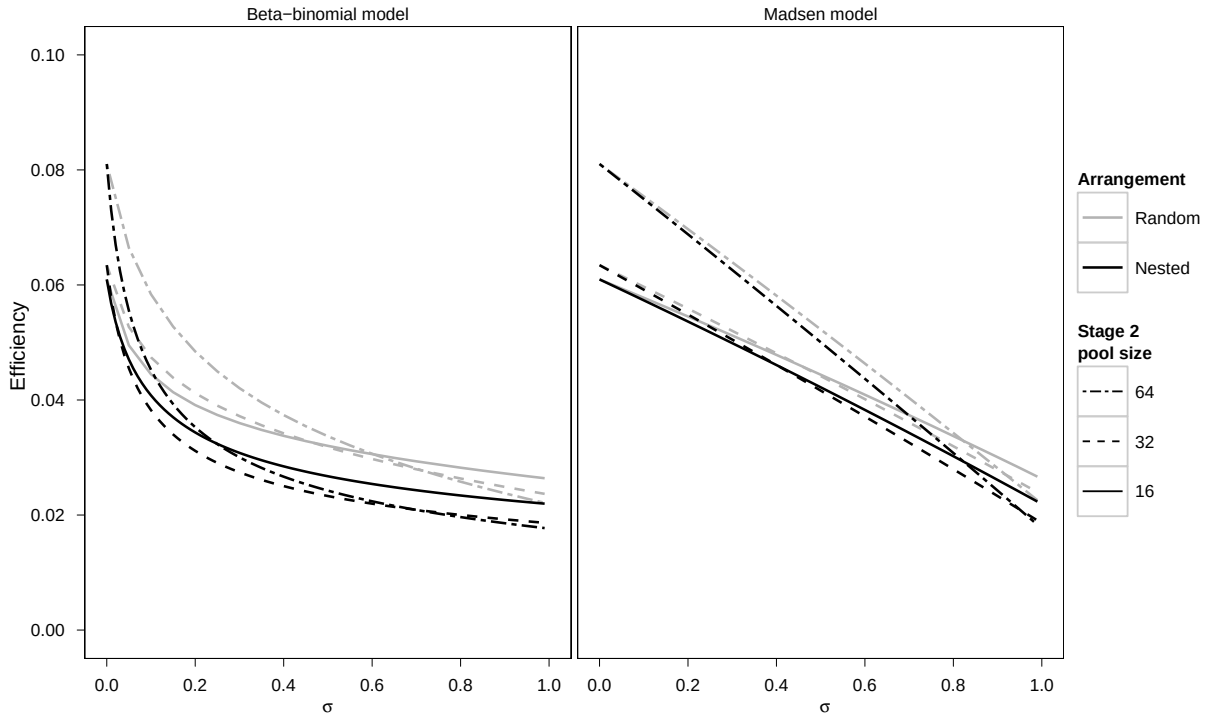
Web Figures



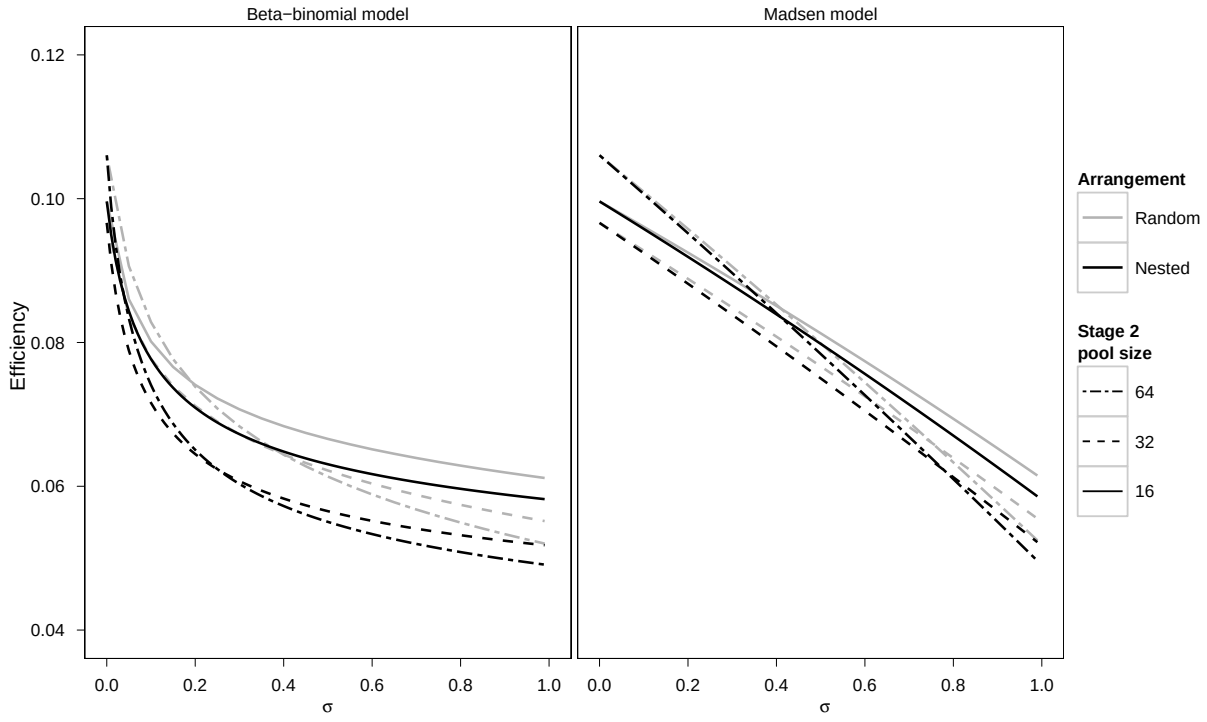
Web Figure 1: Efficiencies for a two stage hierarchical procedure where $S_e = S_p = 0.9$, $n_1 = 64$ and $p = 0.001$ by cluster size m, pairwise correlation σ, and model



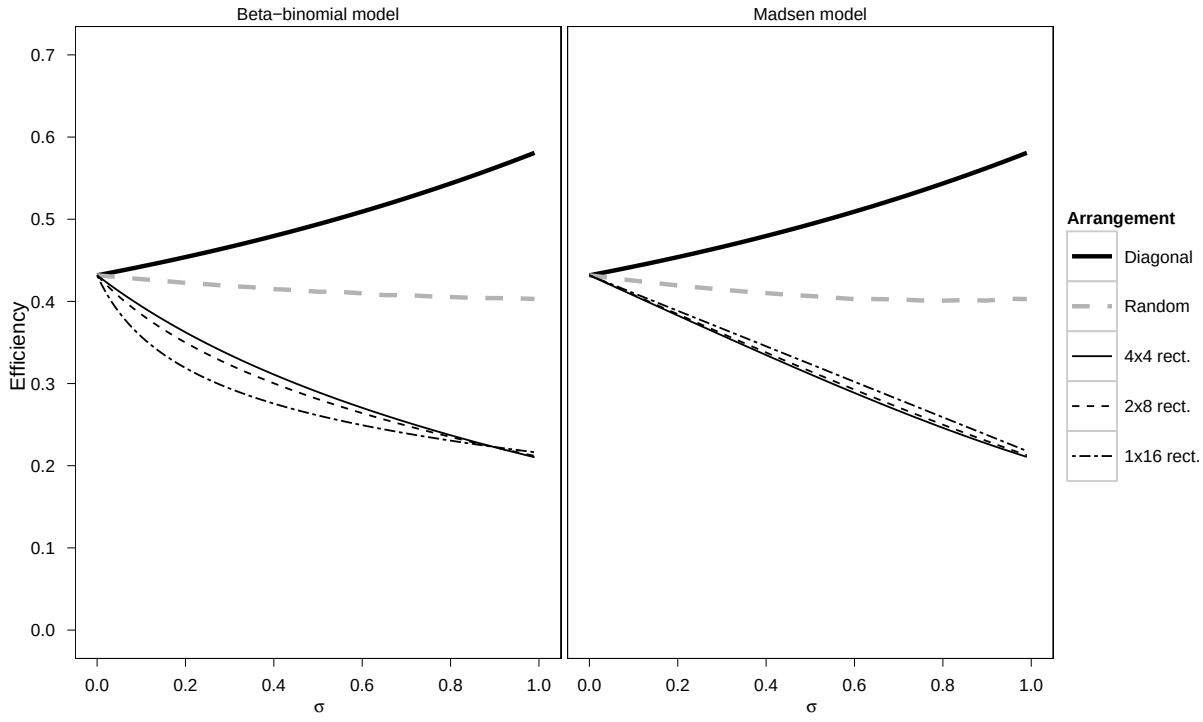
Web Figure 2: Efficiencies for a two stage hierarchical procedure where $S_e = S_p = 0.8$, $n_1 = 64$ and $p = 0.001$ by cluster size m, pairwise correlation σ, and model



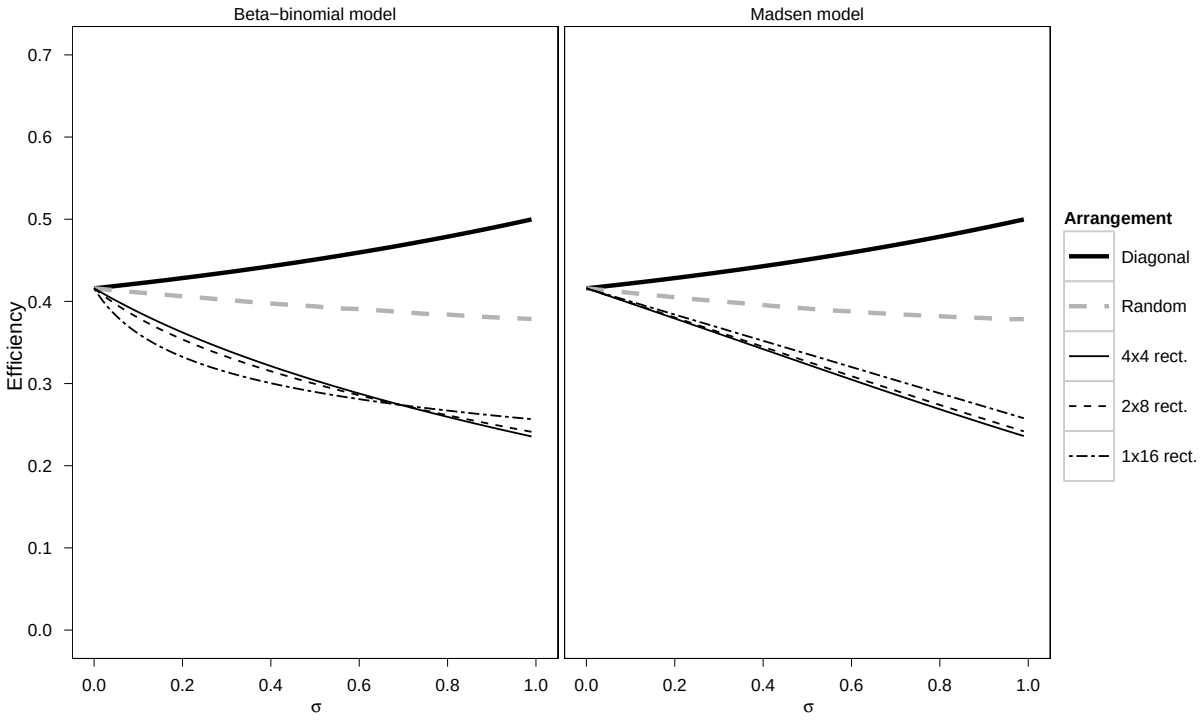
Web Figure 3: Efficiencies for three stage hierarchical procedures where $S_e = S_p = 0.9$, $n_1 = 256$, $p = 0.001$, and $m = 32$ by pairwise correlation σ, stage two pool size n_2, arrangement, and model



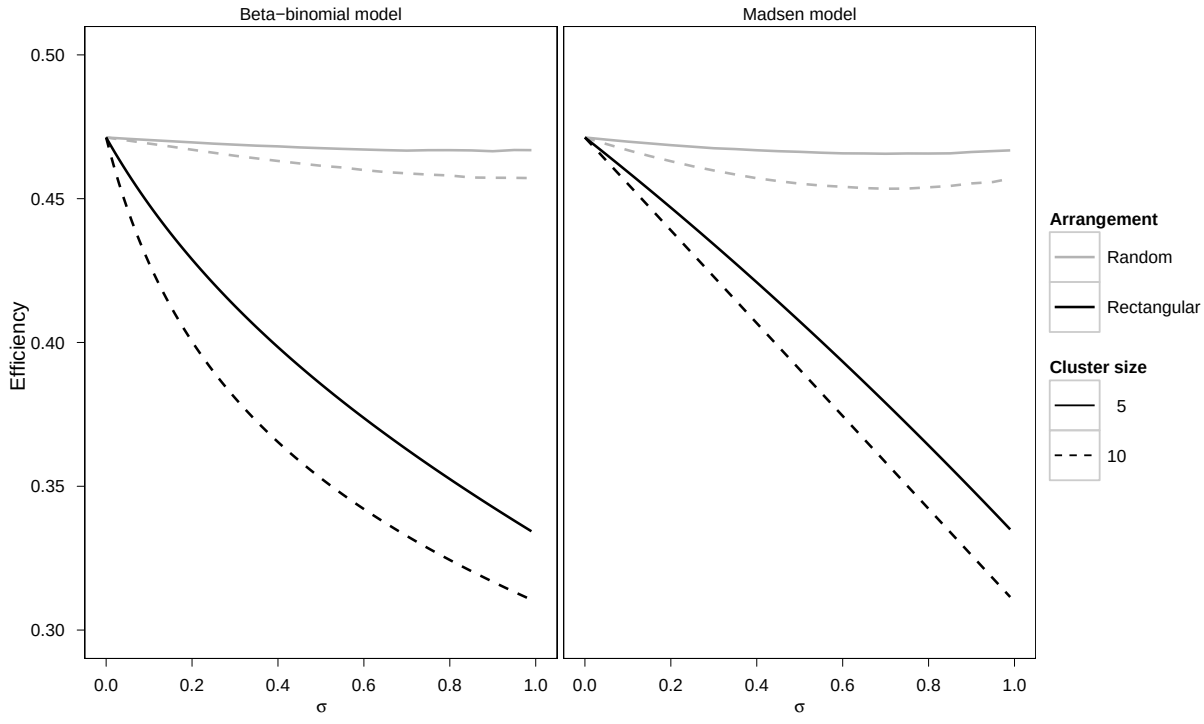
Web Figure 4: Efficiencies for three stage hierarchical procedures where $S_e = S_p = 0.8$, $n_1 = 256$, $p = 0.001$, and $m = 32$ by pairwise correlation σ, stage two pool size n_2, arrangement, and model



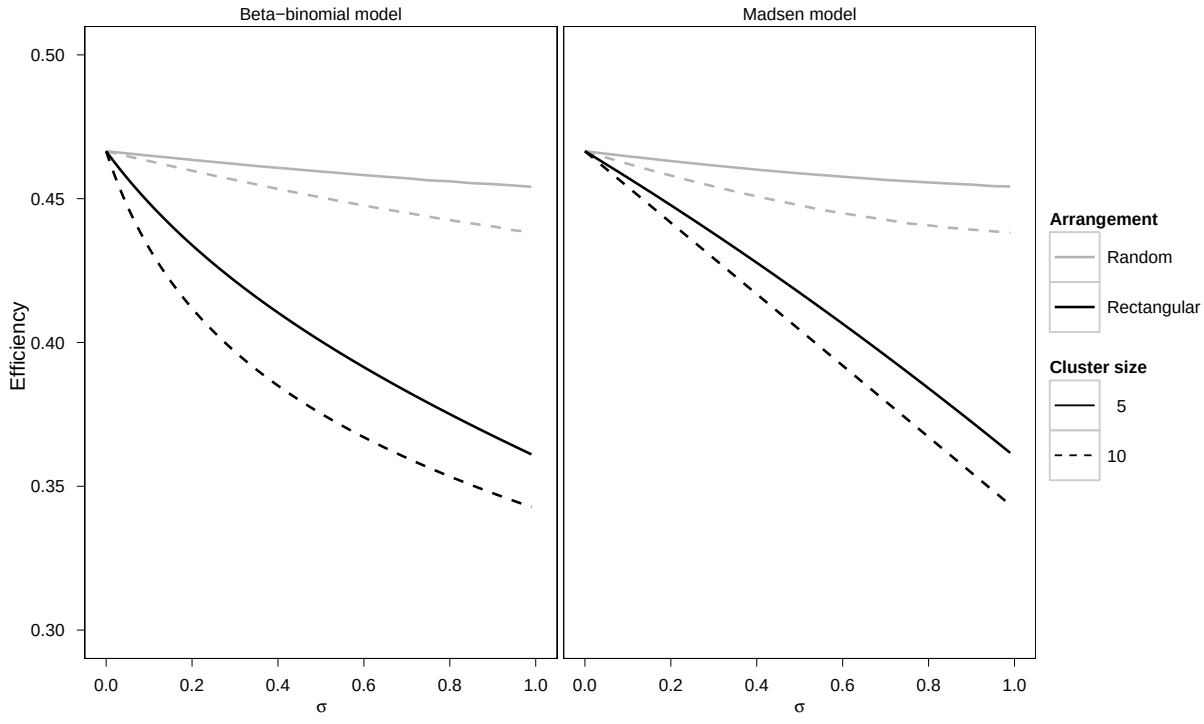
Web Figure 5: Efficiencies for a 16×16 matrix procedure where $p = 0.05$, $S_e = S_p = 0.9$ and clusters are of size $m = 16$ by arrangement, pairwise correlation σ, and model



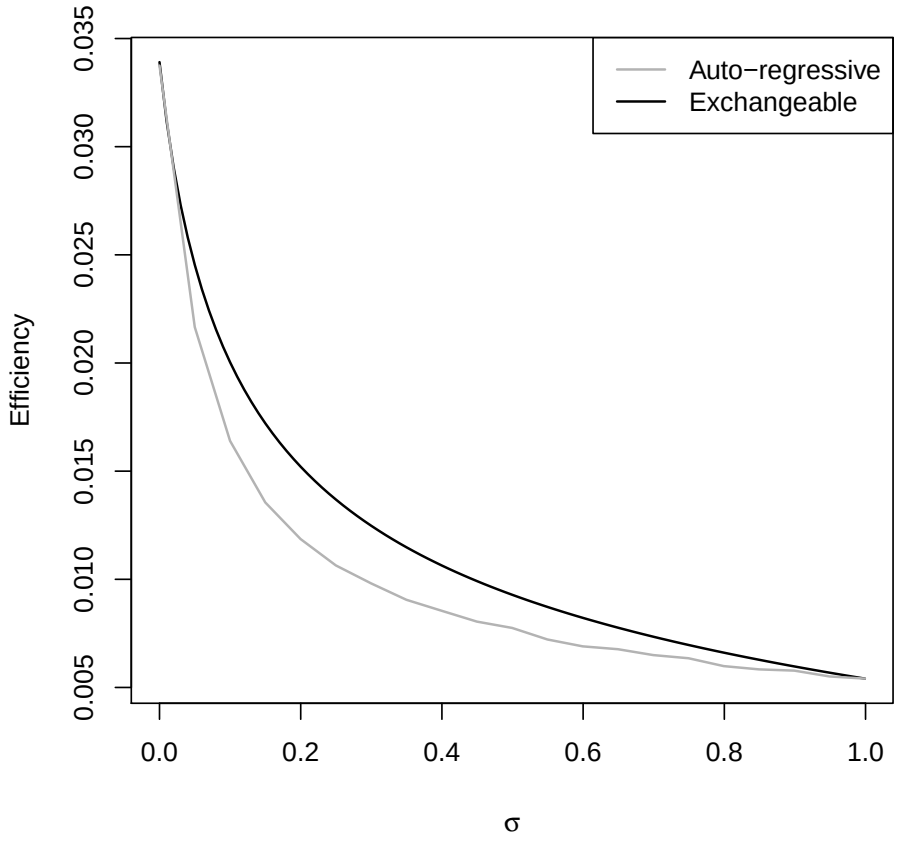
Web Figure 6: Efficiencies for a 16×16 matrix procedure where $p = 0.05$, $S_e = S_p = 0.8$ and clusters are of size $m = 16$ by arrangement, pairwise correlation σ, and model



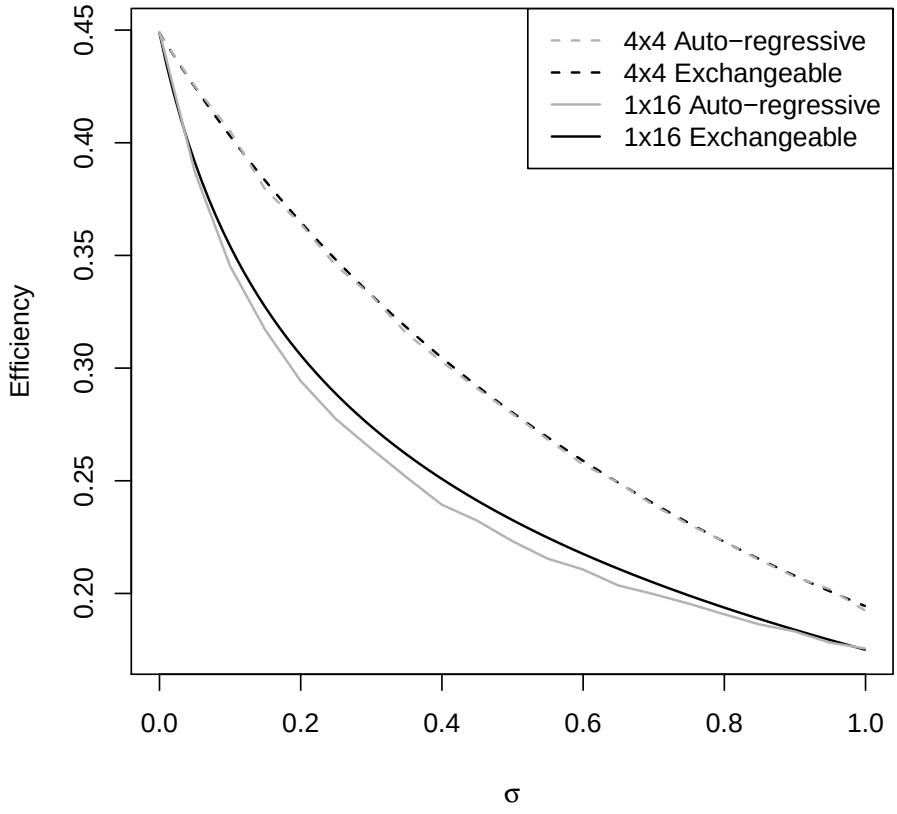
Web Figure 7: Efficiencies for a 9×10 matrix procedure where $p = 0.07$, $S_e = S_p = 0.9$ by pairwise correlation σ, cluster size m, arrangement, and model



Web Figure 8: Efficiencies for a 9×10 matrix procedure where $p = 0.07$, $S_e = S_p = 0.8$ by pairwise correlation σ, cluster size m, arrangement, and model



Web Figure 9: Efficiencies for a three stage nested hierarchical procedure where $S_e = S_p = 1$, $n_1 = 256$, $n_2 = 16$, $p = 0.001$, and $m = 32$ as described in Web Appendix C.



Web Figure 10: Efficiencies for a 16×16 matrix procedure where $S_e = S_p = 1$, $p = 0.05$, and $m = 16$ as described in Web Appendix C. The 4×4 efficiencies (gray and black dashed lines) are nearly identical.

References

Lunn, A. D., and Davies, S. J. (1998). A note on generating correlated binary variables. *Biometrika* **85,** 487–490.

Xie, M., K. Tatsuoka, J. Sacks, and S. S. Young (2001). Group testing with blockers and synergism. *Journal of the American Statistical Association 96*(453), 92–102.