Web Appendix for "Constrained randomization and statistical inference for multi-arm parallel cluster randomized controlled trials"

Yunji Zhou, Elizabeth L. Turner, Ryan A. Simmons, Fan Li*

APPENDIX A: COMPARISON OF PREVIOUS STUDIES AND THIS ARTICLE

Web Table 1 shows the comparison between this article and other studies that investigated the extensions of constrained randomization to three-arm parallel cluster randomized controlled trials (cRCTs).[1,2] Specifically, this article provided additional discussion of the implementation of constrained randomization with more design choices, including more extreme cutoff values of acceptable covariate balance and alternative balance metrics (the maximum pairwise $l2$ metric and the maximum Mahalanobis distance metric) for multi-arm cRCTs. For statistical inference under constrained randomization, new randomization tests for the global hypothesis and pairwise hypotheses in the context of multiple treatments were developed and evaluated in addition to linear mixed models. A comparison between the model-based and randomization-based tests was carried out to provide recommendations on the relative performance of each statistical analysis approach under simple and constrained randomization.

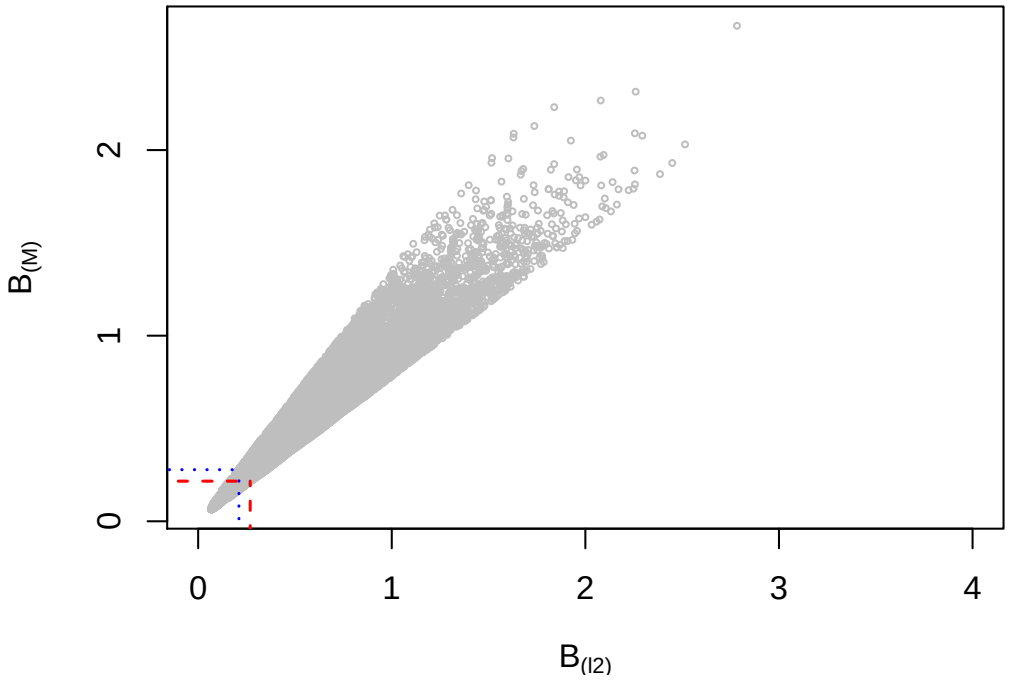
WEB TABLE 1 Comparison of studies examining constrained randomization for multi-arm parallel cluster randomized controlled trials (cRCTs).

	Ciolino et al.[1]	**Watson et al.**[2]	**This article**
Type of study	Parallel cRCT	Parallel cRCT	Parallel cRCT
Number of arms	3	3	3
Study design	Not Applicable	Post only Repeated cross-section Cohort	Post only
# of clusters per arm	10 : 10 : 10 6 : 18 : 18	3 : 3 : 3 ⋮ 11 : 11 : 11	3 : 3 : 3 5 : 5 : 5 10 : 10 : 10
Cluster sizes	Not Applicable	10 to 40	150
Unequal cluster sizes	Not Applicable	TRUE	FALSE
Type of covariates	Continuous	Continuous	Continuous Categorical
Imbalance metrics	min(ANOVA p-values) min(KW p-values) MANOVA p-value min(WRS test p-values) min(t-test p-values)	$l2$ metric	$l2$ metric Mahalanobis distance
Cutoff of acceptable balance	P-value > 0.30	Best 90% to 10%	Best 50%, 10%, 100
Type of outcome	Not Applicable	Continuous	Continuous
Non-normal data	Not Applicable	FALSE	TRUE
Analytical methods	Not Applicable	Linear mixed model	Linear mixed model Randomization test
ICC	Not Applicable	0.001, 0.05	0.01, 0.05, 0.10
Multiplicity adjustment	Not Applicable	FALSE	TRUE

Abbreviations: ICC, intraclass correlation coefficient; ANOVA, analysis of variance; KW, Kruskal–Wallis; MANOVA, multivariate analysis of variance; WRS, Wilcoxon rank-sum.

APPENDIX B: ADDITIONAL ANALYSIS OF THE TESTSMART TRIAL

Web Figure 1 is a scatter plot of the balance scores based on the $l2$ metric ($B_{(l2)}$) against those based on the Mahalanobis distance ($B_{(M)}$) with default weights. The constrained randomization space with $q = 0.1$ under each balance metric is indicated in the plot. Each constrained space reduces the likelihood of outlet/cluster-level covariate imbalance by selecting the most balanced allocation schemes from the complete space.



WEB FIGURE 1 Plot of balance scores from the $l2$ metric ($B_{(l2)}$) against balance scores from the Mahalanobis distance ($B_{(M)}$) in the TESTsmART study. The two constrained randomization spaces ($q = 0.1$) with $B_{(M)}$ and with $B_{(l2)}$ are marked by the long-dashed red lines and the dotted blue lines, respectively.

APPENDIX C: DERIVATION OF TEST STATISTIC (6)

From LMM (3) and the cluster likelihood defined in (4), we have the uniformly most-powerful randomization (UMPR) test statistic for the pairwise hypothesis ($\mathcal{H}_{0,i}$: $\delta_i = 0$) as the joint likelihood $\prod_{j=1}^{G} f(\boldsymbol{Y}_j)$. This test statistic corresponds to the UMPR test because it is independent of the alternative hypothesis $\delta_i = \Delta_i$. The UMPR test statistic can be simplified to test statistic (5) in the special case where $f(Y_{jk}|\gamma_j)$ is $\mathcal{N}(\alpha_{jk}+\gamma_j, \sigma_\epsilon^2)$, and $f(\gamma_j)$ is $\mathcal{N}(0, \sigma_\gamma^2)$.[3] The proof can be found in the Appendix of Braun and Feng[3] and we summarize it here for completeness. In the case where $f(Y_{jk}|\gamma_j)$ is $\mathcal{N}(\alpha_{jk}+\gamma_j, \sigma_\epsilon^2)$, and $f(\gamma_j)$ is $\mathcal{N}(0, \sigma_\gamma^2)$,

the cluster likelihood (4) can be written as

$$
\begin{aligned}
&\int_{-\infty}^{\infty}\prod_{k=1}^{m_j}(2\pi\sigma_\epsilon^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma_\epsilon^2}(y_{jk}-\alpha_{jk}-\gamma)^2\right\}\times(2\pi\sigma_\gamma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma_\gamma^2}(\gamma-0)^2\right\}d\gamma\\
&=\int_{-\infty}^{\infty}(2\pi\sigma_\epsilon^2)^{-m_j/2}\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk}-\gamma)^2\right\}\times(2\pi\sigma_\gamma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma_\gamma^2}(\gamma-0)^2\right\}d\gamma\\
&\propto\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk}-\gamma)^2-\frac{1}{2\sigma_\gamma^2}\gamma^2\right\}d\gamma\\
&\propto\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk})^2+\frac{1}{\sigma_\epsilon^2}\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk})\gamma-\frac{m_j\sigma_\gamma^2+\sigma_\epsilon^2}{2\sigma_\epsilon^2\epsilon_\gamma^2}\gamma^2\right\}d\gamma\\
&\propto\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left[\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk})^2-\frac{\sigma_\gamma^2}{m_j\sigma_\gamma^2+\sigma_\epsilon^2}\left(\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk})\right)^2+\frac{m_j\sigma_\gamma^2+\sigma_\epsilon^2}{\sigma_\gamma^2}\left(\gamma-\frac{\sigma_\gamma^2}{m_j\sigma_\gamma^2+\sigma_\epsilon^2}\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk})\right)^2\right]\right\}d\gamma\\
&\propto\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left[\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk})^2-\frac{\sigma_\gamma^2}{m_j\sigma_\gamma^2+\sigma_\epsilon^2}\left(\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk})\right)^2\right]\right\}\int_{-\infty}^{\infty}\exp\left\{-\frac{\left(\gamma-\frac{\sigma_\gamma^2}{m_j\sigma_\gamma^2+\sigma_\epsilon^2}\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk})\right)^2}{2\frac{\sigma_\epsilon^2\sigma_\gamma^2}{m_j\sigma_\gamma^2+\sigma_\epsilon^2}}\right\}d\gamma\\
&\propto\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left[\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk})^2-\frac{\sigma_\gamma^2}{m_j\sigma_\gamma^2+\sigma_\epsilon^2}\left(\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk})\right)^2\right]\right\}
\end{aligned}
$$

Following Braun and Feng[3], we focus on the exponent due to monotinicity and rewrite the function as

$$
\begin{aligned}
&-\frac{1}{2\sigma_\epsilon^2}\left[\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk})^2-\frac{\sigma_\gamma^2}{m_j\sigma_\gamma^2+\sigma_\epsilon^2}\left(\sum_{k=1}^{m_j}(y_{jk}-\alpha_{jk})\right)^2\right]\\
&=-\frac{1}{2\sigma_\epsilon^2}\left\{\sum_{k=1}^{m_j}\left[y_{jk}-(\lambda+\boldsymbol{x}'_{jk}\boldsymbol{\beta}+\sum_{i'\neq i}\delta_{i'}T_{i'j})-\delta_iT_{ij}\right]^2-\frac{\sigma_\gamma^2}{m_j\sigma_\gamma^2+\sigma_\epsilon^2}\left[\sum_{k=1}^{m_j}y_{jk}-(\lambda+\boldsymbol{x}'_{jk}\boldsymbol{\beta}+\sum_{i'\neq i}\delta_{i'}T_{i'j})-\delta_iT_{ij}\right]^2\right\}\\
&=C_1+C_2+\frac{\delta_iT_{ij}}{\sigma_\epsilon^2}(1-m_jc_j)\sum_{k=1}^{m_j}\left[y_{jk}-\left(\lambda+\boldsymbol{x}'_{jk}\boldsymbol{\beta}+\sum_{i'\neq i}\delta_{i'}T_{i'j}\right)\right],
\end{aligned}
$$

where

$$
\begin{aligned}
c_j&=\sigma_\gamma^2/(\sigma_\epsilon^2+m_j\sigma_\gamma^2),\\
C_1&=-\left\{\sum_{k=1}^{m_j}\left[y_{jk}-\left(\lambda+\boldsymbol{x}'_{jk}\boldsymbol{\beta}+\sum_{i'\neq i}\delta_{i'}T_{i'j}\right)\right]^2+m_j(T_{ij}\delta_i)^2\right\}/2\sigma_\epsilon^2,\\
C_2&=c_j\left\{\left(\sum_{k=1}^{m_j}\left[y_{jk}-\left(\lambda+\boldsymbol{x}'_{jk}\boldsymbol{\beta}+\sum_{i'\neq i}\delta_{i'}T_{i'j}\right)\right]\right)^2+(m_jT_{ij}\delta_i)^2\right\}/2\sigma_\epsilon^2,
\end{aligned}
$$

$T_{i'j}$ is replaced by the *observed* treatment indicator $T^{obs}_{i'j}$ for arm $i' \neq i$, and $T_{ij} \in \{1, -1\}$ is the treatment indicator. By ignoring C_1 and C_2, which are invariant to the treatment assignment because $T^2_{ij} = 1$, we can obtain the cluster specific statistic as

$$\frac{T_{ij}}{\sigma^2_\epsilon + m_j\sigma^2_\gamma} \sum_{k=1}^{m_j} \left(Y_{jk} - \lambda - \boldsymbol{x}'_{jk}\boldsymbol{\beta} - \sum_{i' \neq i} \delta_{i'} T^{obs}_{i'j} \right).$$

By summing over all clusters, we can obtain the overall statistic as

$$\sum_{j=1}^{G} T_{ij} W_j \sum_{k=1}^{m_j} \left(Y_{jk} - \lambda - \boldsymbol{x}'_{jk}\boldsymbol{\beta} - \sum_{i' \neq i} \delta_{i'} T^{obs}_{i'j} \right),$$

where $W_j = (\sigma^2_\epsilon + m_j\sigma^2_\gamma)^{-1}$.

For the global hypothesis ($\mathcal{H}_0$: $\boldsymbol{\delta} = \mathbf{0}$), the joint likelihood $\prod_{j=1}^{G} f(\boldsymbol{Y}_j)$ can still be used, but the resulting statistic does not lead to a UMPR test because it depends on the alternative $\boldsymbol{\delta} = \boldsymbol{\Delta}$. Therefore, we additionally develop a locally most-powerful randomization (LMPR) test based on the marginal likelihood (4), which is rewritten as

$$\boldsymbol{Y}_j \sim \mathcal{N}(\boldsymbol{\alpha}_j = \boldsymbol{T}_j\boldsymbol{\delta} + \boldsymbol{Z}_j\boldsymbol{\eta}, \boldsymbol{\Sigma}_j = \sigma^2_\epsilon \boldsymbol{I} + \sigma^2_\gamma \boldsymbol{J})$$

where $\boldsymbol{T}_j$ is the $m_j \times (C-1)$ matrix of the treatment assignment parameterized as $\{-1, 1\}$, $\boldsymbol{Z}_j$ is the $m_j \times (1+p)$ design matrix including the column vector of ones and the p-dimensional covariates $\boldsymbol{X}_j$, $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ are the corresponding parameter vectors, $\boldsymbol{I}$ is the $m_j \times m_j$ identity matrix, and $\boldsymbol{J}$ is the $m_j \times m_j$ matrix of ones. The full score function is given by

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_\delta \\ \boldsymbol{S}_\eta \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{\delta}} \sum_{j=1}^{G} \log f(\boldsymbol{Y}_j) \\ \frac{\partial}{\partial \boldsymbol{\eta}} \sum_{j=1}^{G} \log f(\boldsymbol{Y}_j) \end{pmatrix},$$

with full information matrix given by

$$\boldsymbol{\mathcal{I}} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{\delta\delta} & \boldsymbol{\mathcal{I}}_{\delta\eta} \\ \boldsymbol{\mathcal{I}}_{\eta\delta} & \boldsymbol{\mathcal{I}}_{\eta\eta} \end{pmatrix} = -E \begin{pmatrix} \frac{\partial^2}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'} \sum_{j=1}^{G} \log f(\boldsymbol{Y}_j) & \frac{\partial^2}{\partial\boldsymbol{\delta}\partial\boldsymbol{\eta}'} \sum_{j=1}^{G} \log f(\boldsymbol{Y}_j) \\ \frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\delta}'} \sum_{j=1}^{G} \log f(\boldsymbol{Y}_j) & \frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'} \sum_{j=1}^{G} \log f(\boldsymbol{Y}_j) \end{pmatrix}.$$

Since we need to estimate the nuisance parameter η, we summarize $\boldsymbol{S}$ using the efficient score statistic and define the locally most-powerful randomization (LMPR) test statistic as

$$Q = \boldsymbol{S}_\delta{}'(\boldsymbol{\mathcal{I}}_{\delta\delta} - \boldsymbol{\mathcal{I}}_{\delta\eta}\boldsymbol{\mathcal{I}}^{-1}_{\eta\eta}\boldsymbol{\mathcal{I}}_{\eta\delta})^{-1}\boldsymbol{S}_\delta$$

To operationalize the LMPR test statistic, we calculate both $\boldsymbol{S}$ and $\boldsymbol{\mathcal{I}}$ evaluated under the global null in the following steps.
(1) We first calculate $\boldsymbol{S}_\delta|_{\delta=\mathbf{0}}$

$$\begin{aligned} \boldsymbol{S}_\delta|_{\delta=\mathbf{0}} &= \frac{\partial}{\partial\boldsymbol{\delta}} \sum_{j=1}^{G} \log f(\boldsymbol{Y}_j)|_{\delta=\mathbf{0}} \\ &= \sum_{j=1}^{G} \boldsymbol{T}'\boldsymbol{\Sigma}^{-1}_j(\boldsymbol{Y}_j - \boldsymbol{T}_j\boldsymbol{\delta} - \boldsymbol{Z}_j\boldsymbol{\eta})|_{\delta=\mathbf{0}} \\ &= \sum_{j=1}^{G} \boldsymbol{T}'_j\boldsymbol{\Sigma}^{-1}_j(\boldsymbol{Y}_j - \boldsymbol{Z}_j\boldsymbol{\eta}), \end{aligned}$$

which can be written in a $(C-1) \times 1$ matrix as

$$\left(\sum_{j=1}^{G} T_{1j}\mathbf{1}'_{m_j}\boldsymbol{\Sigma}^{-1}_j(\boldsymbol{Y}_j - \boldsymbol{Z}_j\boldsymbol{\eta}) \ \cdots \ \sum_{j=1}^{G} T_{(C-1)j}\mathbf{1}'_{m_j}\boldsymbol{\Sigma}^{-1}_j(\boldsymbol{Y}_j - \boldsymbol{Z}_j\boldsymbol{\eta}) \right)'$$

(2) We then derive $\boldsymbol{\mathcal{I}}_{\delta\delta}$

$$\begin{aligned}\boldsymbol{\mathcal{I}}_{\delta\delta} =& - E\left\{\frac{\partial^2}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\sum_{j=1}^{G}\log f(\boldsymbol{Y}_j)\right\}\\ =&E\left(\sum_{j=1}^{G}\boldsymbol{T}_j'\boldsymbol{\Sigma}_j^{-1}\boldsymbol{T}_j\right)\\ =&\sum_{j=1}^{G}E\left(\boldsymbol{T}_j'\boldsymbol{\Sigma}_j^{-1}\boldsymbol{T}_j\right)\end{aligned}$$

The diagonal element of $\boldsymbol{\mathcal{I}}_{\delta\delta}$ is derived as follows

$$\begin{aligned}\sum_{j=1}^{G}E\left(T_{ij}\mathbf{1}'\boldsymbol{\Sigma}_j^{-1}T_{ij}\mathbf{1}\right) =& \sum_{j=1}^{G}E\left(\mathbf{1}'\boldsymbol{\Sigma}_j^{-1}\mathbf{1}\right)\\ =&\sum_{j=1}^{G}\frac{m_j}{\sigma_\epsilon^2}-\frac{m_j^2\sigma_\gamma^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+m_j\sigma_\gamma^2)}\\ =&\sum_{j=1}^{G}\frac{m_j}{\sigma_\epsilon^2+m_j\sigma_\gamma^2}\\ =&\sum_{j=1}^{G}m_jW_j\end{aligned}$$

The off-diagonal element of $\boldsymbol{\mathcal{I}}_{\delta\delta}$ is derived as follows

$$\begin{aligned}\sum_{j=1}^{G}E\left(T_{ij}\mathbf{1}'\boldsymbol{\Sigma}_j^{-1}T_{i'j}\mathbf{1}\right) =& \sum_{j=1}^{G}E\left(T_{ij}T_{i'j}\right)\left(\mathbf{1}'\boldsymbol{\Sigma}_j^{-1}\mathbf{1}\right)\\ =&\sum_{j=1}^{G}\frac{m_jE(T_{ij}T_{i'j})}{\sigma_\epsilon^2+m_j\sigma_\gamma^2}\\ =&\sum_{j=1}^{G}m_jW_jE(T_{ij}T_{i'j})\end{aligned}$$

To calculate $E(T_{ij}T_{i'j})$, we need to obtain the joint probability density function of T_{ij} and $T_{i'j}$ with respect to π_{ij} and $\pi_{i'j}$, which are the known probabilities of a cluster j being assigned to arm i and i', respectively:

$$f(T_{ij},T_{i'j};\pi_{ij},\pi_{i'j}) = \begin{cases}0 & \text{if } T_{ij}=1 \;\&\; T_{i'j}=1\\ \pi_{ij} & \text{if } T_{ij}=1 \;\&\; T_{i'j}=-1\\ 1-\pi_{ij}-\pi_{i'j} & \text{if } T_{ij}=-1 \;\&\; T_{i'j}=-1\\ \pi_{i'j} & \text{if } T_{ij}=-1 \;\&\; T_{i'j}=1\end{cases}$$

Therefore, $E(T_{ij}T_{i'j}) = \sum T_{ij}T_{i'j}f(T_{ij},T_{i'j};\pi_{ij},\pi_{i'j}) = 0-\pi_{ij}+1-\pi_{ij}-\pi_{i'j}-\pi_{i'j} = 1-2\pi_{ij}-2\pi_{i'j}$
(3) Next, we derive $\boldsymbol{\mathcal{I}}_{\eta\eta}$

$$\begin{aligned}\boldsymbol{\mathcal{I}}_{\eta\eta} =& - E\left\{\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'}\sum_{j=1}^{G}\log f(\boldsymbol{Y}_j)\right\}\\ =&\sum_{j=1}^{G}\boldsymbol{Z}_j'\boldsymbol{\Sigma}_j^{-1}\boldsymbol{Z}_j\end{aligned}$$

where $\boldsymbol{\Sigma}_j^{-1}$ can be shown to be equal to $\frac{1}{\sigma_\epsilon^2}\boldsymbol{I}_{m_j} - \frac{m_j\sigma_\gamma^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+m_j\sigma_\gamma^2)}\boldsymbol{J}_{m_j}$ following Appendix A of Li et al.[4]
(4) Next, we derive $\boldsymbol{\mathcal{I}}_{\eta\delta}$

$$\begin{aligned}
\boldsymbol{\mathcal{I}}_{\eta\delta} &= -E\left\{\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\delta}'}\sum_{j=1}^{G}\log f(\boldsymbol{Y}_j)\right\}\\
&= E\left(\sum_{j=1}^{G}\boldsymbol{Z}_j'\boldsymbol{\Sigma}_j^{-1}\boldsymbol{T}_j\right)\\
&= \sum_{j=1}^{G}\boldsymbol{Z}_j'\boldsymbol{\Sigma}_j^{-1}E\left(\boldsymbol{T}_j\right)\\
&= \sum_{j=1}^{G}\boldsymbol{Z}_j'\left(\frac{1}{\sigma_\epsilon^2}E(\boldsymbol{T}_j) - \frac{m_j\sigma_\gamma^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+m_j\sigma_\gamma^2)}\mathbf{1}\left(2\pi_1 - 1 \;\cdots\; 2\pi_{(C-1)} - 1\right)\right)\\
&= \sum_{j=1}^{G}\boldsymbol{Z}_j'\left(\frac{1}{\sigma_\epsilon^2}E(\boldsymbol{T}_j) - \frac{m_j\sigma_\gamma^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+m_j\sigma_\gamma^2)}E(\boldsymbol{T}_j)\right)\\
&= \sum_{j=1}^{G}\boldsymbol{Z}_j'\left(\frac{1}{\sigma_\epsilon^2+m_j\sigma_\gamma^2}E(\boldsymbol{T}_j)\right).
\end{aligned}$$

Since each column of $E(\boldsymbol{T}_j)$ corresponds to a treatment variable T_{ij} and $E(T_{ij}) = \pi_i - (1-\pi_i) = 2\pi_i - 1$, then the ith column of $E(\boldsymbol{T}_j)$ is written by $(2\pi_i - 1)\mathbf{1}_{m_j}$, and we therefore can write

$$\boldsymbol{\mathcal{I}}_{\eta\delta} = \sum_{j=1}^{G}\frac{1}{\sigma_\epsilon^2+m_j\sigma_\gamma^2}\begin{pmatrix}(2\pi_1-1)m_j & (2\pi_2-1)m_j & \cdots & (2\pi_{C-1}-1)m_j\\ (2\pi_1-1)\sum_{k=1}^{m_j}\boldsymbol{x}_{jk} & (2\pi_2-1)\sum_{k=1}^{m_j}\boldsymbol{x}_{jk} & \cdots & (2\pi_{C-1}-1)\sum_{k=1}^{m_j}\boldsymbol{x}_{jk}\end{pmatrix}.$$

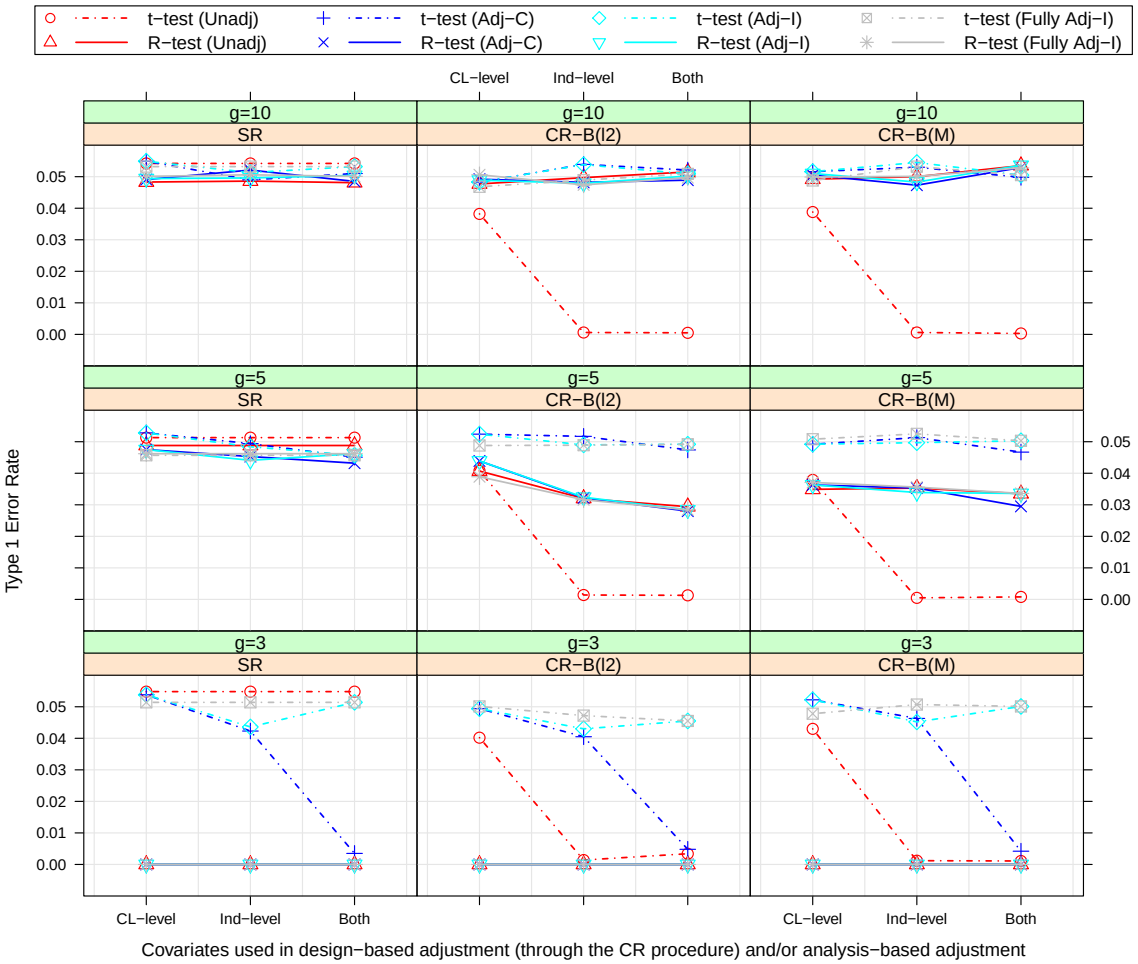
(5) Finaly, we obtain $\boldsymbol{\mathcal{I}}_{\delta\eta} = \boldsymbol{\mathcal{I}}_{\eta\delta}{}'$

APPENDIX D: RESULTS FOR THE PAIRWISE HYPOTHESES

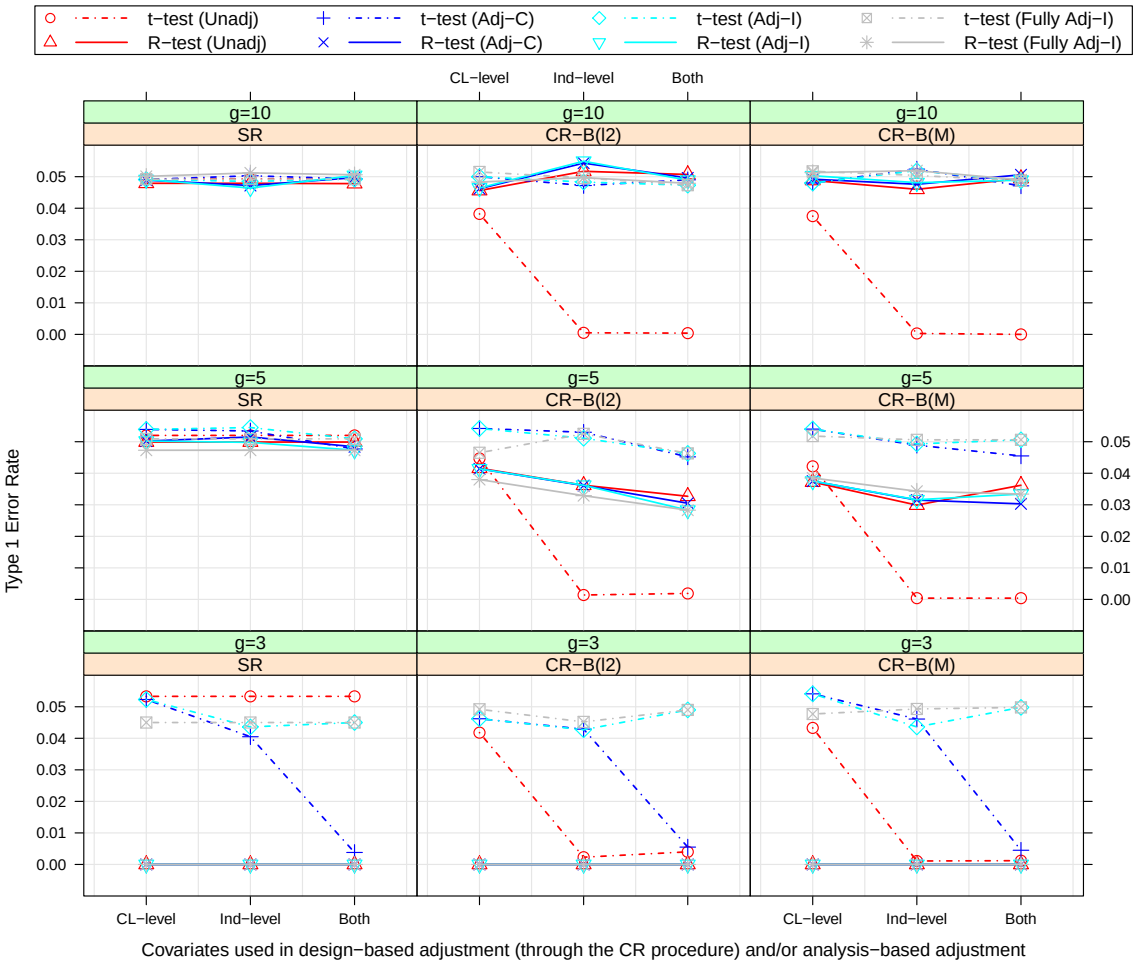
We presented the results for the two pairwise hypotheses ($\mathcal{H}_0$: $\delta_1 = 0$ and $\mathcal{H}_0$: $\delta_2 = 0$) in this section. In Web Table 2 and Web Figures 2-3, we summarized the Monte Carlo type I error rates under simple randomization (SR) and constrained randomization (CR), while in Web Table 3 and Web Figures 4-5, we summarized the corresponding results for power. We held the ICC fixed at 0.05 throughout this section and compared results for $g = 3$, 5, and 10.

WEB TABLE 2 Type I error rates for the pairwise hypotheses ($\mathcal{H}_0$: $\delta_1 = 0$ and $\mathcal{H}_0$: $\delta_2 = 0$) under simple randomization (SR) versus constrained randomization (CR) with candidate set sizes = 50%, 10%, and 100 of the randomization space. All covariates were used in constrained randomization and the adjusted tests; constrained randomization was implemented using the $l2$ metric; ICC = 0.05; alpha level = 5%. The nominal type I error rate is 0.05, and the acceptance range for nominal type I error rate with 10,000 replicates is (0.0457, 0.0543).

			***t*-test**				**Randomization test**			
$\mathcal{H}_0$	# of clusters per arm	Analysis-based adjustment	SR	CR (50%)	CR (10%)	CR (100)	SR	CR (50%)	CR (10%)	CR (100)
$\delta_1 = 0$	$g = 10$	Unadj	0.054	0.009	0.000	0.000	0.048	0.050	0.051	0.050
		Adj-C	0.051	0.052	0.052	0.046	0.048	0.052	0.049	0.048
		Adj-I	0.053	0.052	0.051	0.047	0.050	0.051	0.050	0.049
	$g = 5$	Unadj	0.051	0.012	0.001	0.000	0.049	0.045	0.029	0.000
		Adj-C	0.045	0.043	0.047	0.050	0.043	0.044	0.028	0.001
		Adj-I	0.046	0.050	0.049	0.050	0.046	0.041	0.028	0.000
	$g = 3$	Unadj	0.055	0.018	0.003	0.002	0.000	0.000	0.000	0.000
		Adj-C	0.004	0.005	0.005	0.004	0.000	0.000	0.000	0.000
		Adj-I	0.051	0.050	0.045	0.050	0.000	0.000	0.000	0.000
$\delta_2 = 0$	$g = 10$	Unadj	0.049	0.010	0.000	0.000	0.048	0.052	0.051	0.051
		Adj-C	0.049	0.048	0.049	0.050	0.050	0.049	0.050	0.050
		Adj-I	0.049	0.049	0.047	0.049	0.050	0.052	0.049	0.051
	$g = 5$	Unadj	0.052	0.013	0.002	0.000	0.050	0.043	0.033	0.001
		Adj-C	0.048	0.046	0.045	0.048	0.048	0.042	0.030	0.000
		Adj-I	0.051	0.048	0.046	0.051	0.047	0.043	0.028	0.000
	$g = 3$	Unadj	0.053	0.019	0.004	0.002	0.000	0.000	0.000	0.000
		Adj-C	0.004	0.006	0.005	0.004	0.000	0.000	0.000	0.000
		Adj-I	0.045	0.048	0.049	0.045	0.000	0.000	0.000	0.000



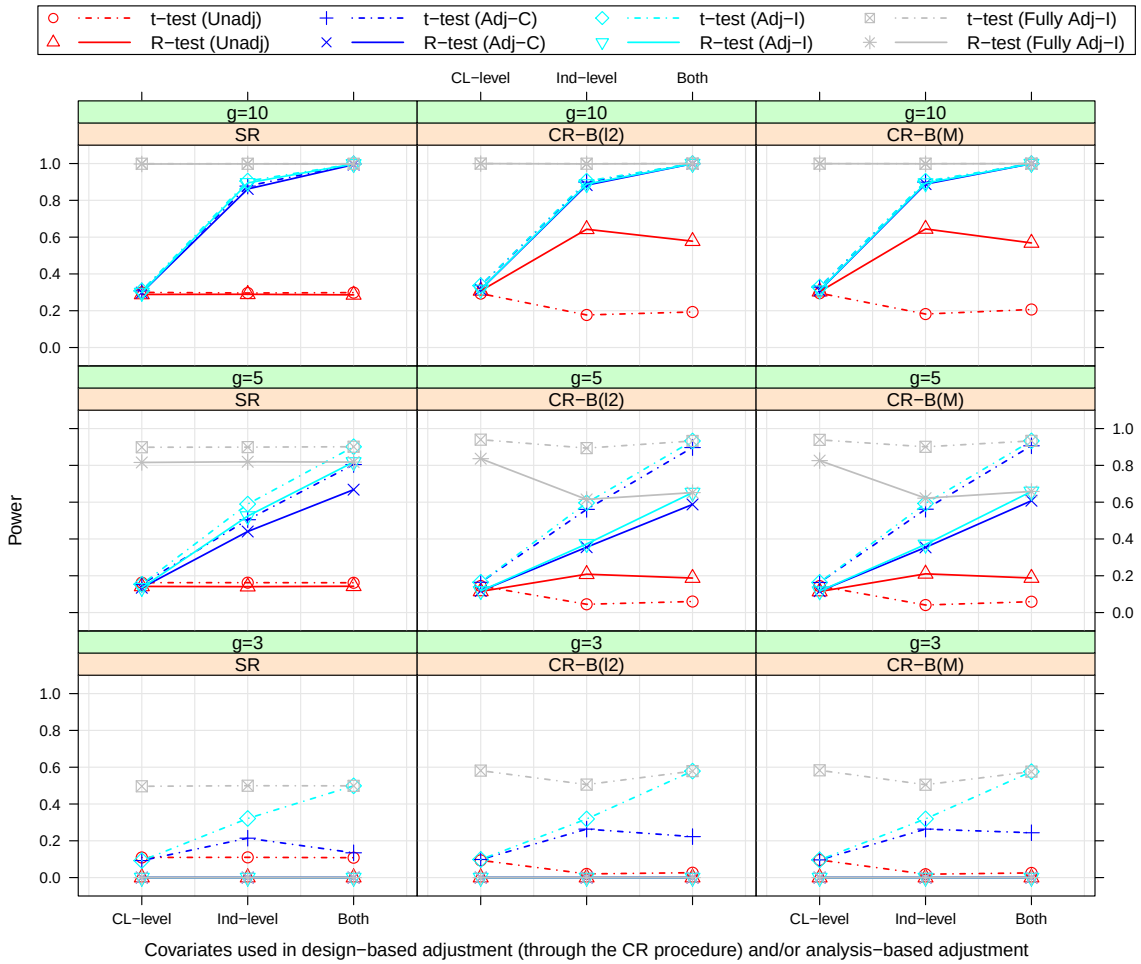
WEB FIGURE 2 Type I error rates for the pairwise hypothesis ($\mathcal{H}_0$: $\delta_1 = 0$) under simple randomization (SR) versus constrained randomization (CR) with 2 balance metrics $B_{(l2)}$ and $B_{(M)}$. CR implemented using covariates indicated on the horizontal axis; candidate set size = 10% under CR; ICC = 0.05; alpha level = 5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).



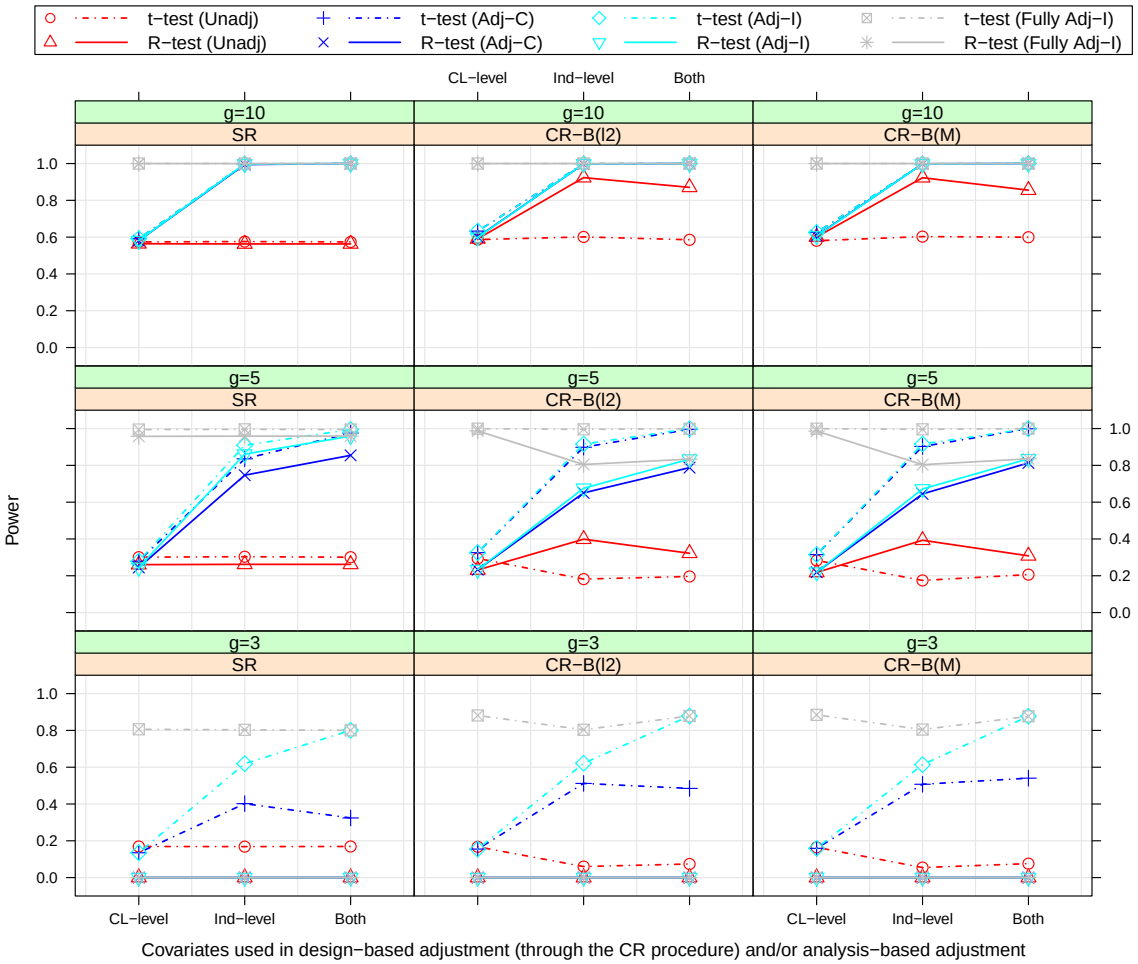
WEB FIGURE 3 Type I error rates for the pairwise hypothesis ($\mathcal{H}_0$: $\delta_2 = 0$) under simple randomization (SR) versus constrained randomization (CR) with 2 balance metrics $B_{(I2)}$ and $B_{(M)}$. CR implemented using covariates indicated on the horizontal axis; candidate set size = 10% under CR; ICC = 0.05; alpha level = 5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).

WEB TABLE 3 Power for the pairwise hypotheses ($\mathcal{H}_0$: $\delta_1 = 0$ and $\mathcal{H}_0$: $\delta_2 = 0$) under simple randomization (SR) versus constrained randomization (CR) with candidate set sizes = 50%, 10%, and 100 of the randomization space. All covariates were used in constrained randomization and the adjusted tests; constrained randomization was implemented using the $l2$ metric; ICC = 0.05; alpha level = 5%; power values corresponding to non-nominal type I errors are shaded out.

			***t*-test**				**Randomization test**			
$\mathcal{H}_0$	# of clusters per arm	Analysis-based adjustment	SR	CR (50%)	CR (10%)	CR (100)	SR	CR (50%)	CR (10%)	CR (100)
$\delta_1 = 0$	$g = 10$	Unadj	0.298	0.258	0.193	0.132	0.286	0.408	0.578	0.800
		Adj-C	0.998	0.999	1.000	1.000	0.994	0.998	0.999	0.999
		Adj-I	0.999	0.999	1.000	1.000	0.998	0.999	0.999	0.999
	$g = 5$	Unadj	0.162	0.108	0.060	0.018	0.143	0.175	0.188	0.004
		Adj-C	0.805	0.853	0.897	0.914	0.668	0.701	0.588	0.007
		Adj-I	0.901	0.923	0.933	0.937	0.818	0.810	0.652	0.007
	$g = 3$	Unadj	0.108	0.064	0.027	0.024	0.000	0.000	0.000	0.000
		Adj-C	0.135	0.176	0.223	0.235	0.000	0.000	0.000	0.000
		Adj-I	0.499	0.546	0.579	0.590	0.000	0.000	0.000	0.000
$\delta_2 = 0$	$g = 10$	Unadj	0.574	0.578	0.586	0.606	0.563	0.712	0.871	0.982
		Adj-C	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Adj-I	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$g = 5$	Unadj	0.301	0.256	0.196	0.120	0.262	0.323	0.322	0.005
		Adj-C	0.976	0.993	0.996	0.998	0.855	0.908	0.788	0.008
		Adj-I	0.995	0.999	0.999	1.000	0.959	0.970	0.834	0.008
	$g = 3$	Unadj	0.169	0.131	0.074	0.064	0.000	0.000	0.000	0.000
		Adj-C	0.324	0.394	0.485	0.500	0.000	0.000	0.000	0.000
		Adj-I	0.800	0.848	0.878	0.882	0.000	0.000	0.000	0.000



WEB FIGURE 4 Power for the pairwise hypothesis ($\mathcal{H}_0$: $\delta_1 = 0$) under simple randomization (SR) versus constrained randomization (CR) with 2 balance metrics $B_{(l2)}$ and $B_{(M)}$. CR implemented using covariates indicated on the horizontal axis; candidate set size = 10% under CR; ICC = 0.05; alpha level = 5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).



WEB FIGURE 5 Power for the pairwise hypothesis ($\mathcal{H}_0$: $\delta_2 = 0$) under simple randomization (SR) versus constrained randomization (CR) with 2 balance metrics $B_{(I2)}$ and $B_{(M)}$. CR implemented using covariates indicated on the horizontal axis; candidate set size = 10% under CR; ICC = 0.05; alpha level = 5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).

APPENDIX E: RESULTS UNDER DIFFERENT INTRACLASS CORRELATION COEFFICIENT

We presented the results under an intraclass correlation coefficient (ICC) of 0.05 in the main text. Two alternative ICCs were considered: 0.10 and 0.01, and presented the results under alternative ICCs in this section. We avoided the comparison of results across different ICCs. Instead, we compared the performance of the design-based and analysis-based adjustment strategies within each level of ICC. In Web Tables 4-5, we summarized the type I error rate and power under ICC = 0.10, while in Web Tables 6-7, we summarized the results under ICC = 0.01. The number of clusters per arm g is held at 5 and the alpha level is held at 5% throughout the comparison in this section. As a counterpart of the χ^2-test, the z-test for the pairwise hypotheses ($\mathcal{H}_0$: $\delta_1 = 0$ and $\mathcal{H}_0$: $\delta_2 = 0$) was not considered further in our simulations because it will carry inflated type I error rate.

WEB TABLE 4 Results under ICC = 0.10: Type I error rates under simple randomization (SR) versus constrained randomization (CR) with candidate set sizes = 50%, 10%, and 100 of the randomization space. All covariates were used in constrained randomization and the adjusted tests; constrained randomization was implemented using the $l2$ metric; $g = 5$. The nominal type I error rate is 0.05, and the acceptable range for nominal type I error rate with 10,000 replicates is (0.0457, 0.0543).

		χ^2-test				F-test/t-test				Randomization test			
$\mathcal{H}_0$	Analysis-based adjustment	SR	CR (50%)	CR (10%)	CR (100)	SR	CR (50%)	CR (10%)	CR (100)	SR	CR (50%)	CR (10%)	CR (100)
δ_1=δ_2=0	Unadj	0.090	0.025	0.004	0.000	0.053	0.010	0.001	0.000	0.052	0.051	0.051	0.043
	Adj-C	0.112	0.108	0.106	0.107	0.053	0.052	0.050	0.050	0.053	0.050	0.050	0.041
	Adj-I	0.100	0.099	0.095	0.095	0.054	0.052	0.049	0.050	0.054	0.050	0.048	0.040
δ_1=0	Unadj	—	—	—	—	0.052	0.014	0.002	0.000	0.046	0.039	0.035	0.000
	Adj-C	—	—	—	—	0.053	0.052	0.051	0.050	0.047	0.042	0.030	0.000
	Adj-I	—	—	—	—	0.053	0.053	0.049	0.050	0.050	0.041	0.031	0.000

WEB TABLE 5 Results under ICC = 0.10: Power under simple randomization (SR) versus constrained randomization (CR) with candidate set sizes = 50%, 10%, and 100 of the randomization space. All covariates were used in constrained randomization and the adjusted tests; constrained randomization was implemented using the $l2$ metric; $g = 5$; power values corresponding to non-nominal type I errors are shaded out.

		χ^2-test				F-test/t-test				Randomization test			
$\mathcal{H}_0$	Analysis-based adjustment	SR	CR (50%)	CR (10%)	CR (100)	SR	CR (50%)	CR (10%)	CR (100)	SR	CR (50%)	CR (10%)	CR (100)
δ_1=δ_2=0	Unadj	0.341	0.273	0.201	0.123	0.240	0.175	0.111	0.056	0.239	0.327	0.428	0.529
	Adj-C	0.914	0.945	0.964	0.978	0.814	0.858	0.899	0.923	0.719	0.804	0.870	0.835
	Adj-I	0.961	0.972	0.978	0.981	0.914	0.928	0.941	0.951	0.869	0.903	0.927	0.863
δ_1=0	Unadj	—	—	—	—	0.171	0.125	0.077	0.036	0.152	0.184	0.184	0.003
	Adj-C	—	—	—	—	0.604	0.650	0.699	0.737	0.500	0.508	0.423	0.005
	Adj-I	—	—	—	—	0.718	0.742	0.758	0.776	0.631	0.615	0.483	0.006

WEB TABLE 6 Results under ICC = 0.01: Type I error rates under simple randomization (SR) versus constrained randomization (CR) with candidate set sizes = 50%, 10%, and 100 of the randomization space. All covariates were used in constrained randomization and the adjusted tests; constrained randomization was implemented using the $l2$ metric; $g = 5$. The nominal type I error rate is 0.05, and the acceptable range for nominal type I error rate with 10,000 replicates is (0.0457, 0.0543).

		χ^2-test				F-test/t-test				Randomization test			
$\mathcal{H}_0$	Analysis-based adjustment	SR	CR (50%)	CR (10%)	CR (100)	SR	CR (50%)	CR (10%)	CR (100)	SR	CR (50%)	CR (10%)	CR (100)
$\delta_1=\delta_2=0$	Unadj	0.093	0.019	0.001	0.000	0.056	0.007	0.000	0.000	0.055	0.053	0.049	0.038
	Adj-C	0.024	0.023	0.024	0.022	0.004	0.003	0.004	0.003	0.048	0.046	0.051	0.040
	Adj-I	0.086	0.087	0.090	0.096	0.041	0.044	0.043	0.050	0.045	0.050	0.048	0.042
$\delta_1=0$	Unadj	—	—	—	—	0.054	0.012	0.001	0.000	0.050	0.046	0.030	0.000
	Adj-C	—	—	—	—	0.010	0.010	0.011	0.007	0.046	0.047	0.032	0.001
	Adj-I	—	—	—	—	0.042	0.049	0.049	0.050	0.042	0.043	0.032	0.000

WEB TABLE 7 Results under ICC = 0.01: Power under simple randomization (SR) versus constrained randomization (CR) with candidate set sizes = 50%, 10%, and 100 of the randomization space. All covariates were used in constrained randomization and the adjusted tests; constrained randomization was implemented using the $l2$ metric; $g = 5$; power values corresponding to non-nominal type I errors are shaded out.

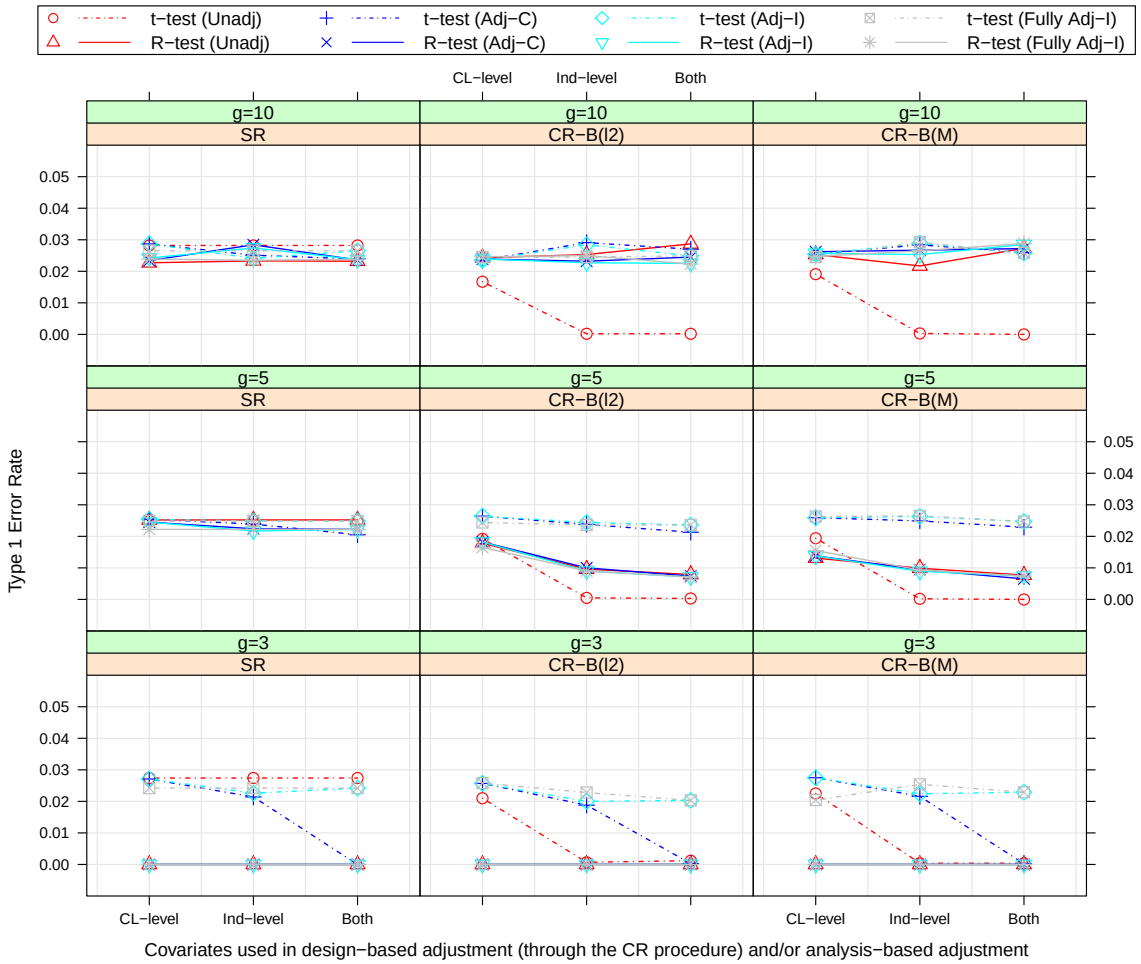
		χ^2-test				F-test/t-test				Randomization test			
$\mathcal{H}_0$	Analysis-based adjustment	SR	CR (50%)	CR (10%)	CR (100)	SR	CR (50%)	CR (10%)	CR (100)	SR	CR (50%)	CR (10%)	CR (100)
$\delta_1=\delta_2=0$	Unadj	0.283	0.209	0.111	0.035	0.192	0.119	0.052	0.011	0.188	0.293	0.409	0.571
	Adj-C	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.941	0.989	0.999	0.999
	Adj-I	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.992	1.000	1.000	1.000
$\delta_1=0$	Unadj	—	—	—	—	0.143	0.100	0.041	0.007	0.125	0.171	0.182	0.003
	Adj-C	—	—	—	—	0.985	0.997	0.999	1.000	0.887	0.936	0.823	0.008
	Adj-I	—	—	—	—	0.999	1.000	1.000	1.000	0.974	0.976	0.853	0.008

APPENDIX F: RESULTS UNDER MULTIPLICITY ADJUSTMENT

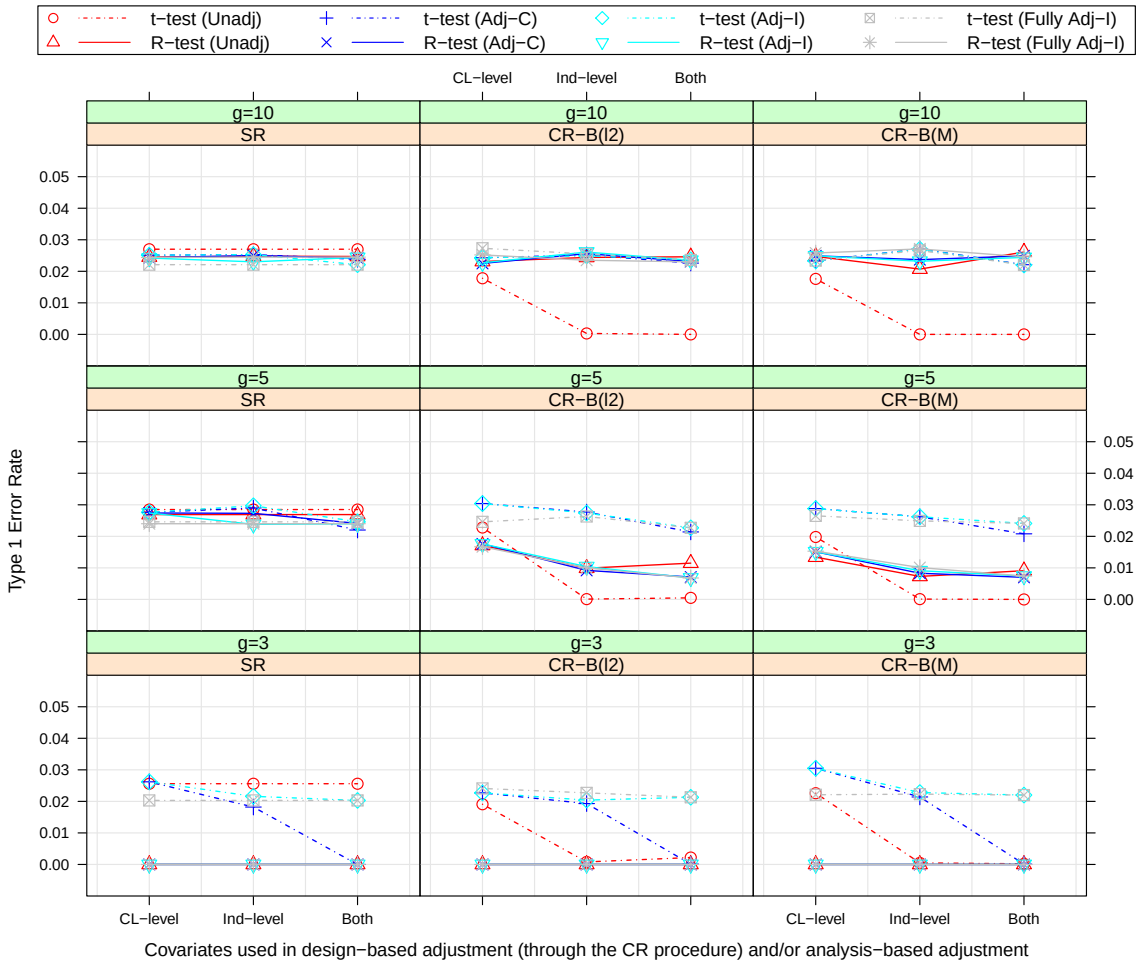
We performed a conservative Bonferroni adjustment[5] for the tests of the two pairwise hypotheses, which is suggested for the scenario where the two active treatment arms consist of the same treatment given at different doses and the overall effectiveness will be concluded if any one of the treatment doses shows a significant effect relative to 'standard of care'. In Web Table 8 and Web Figures 6-7, we summarized the Monte Carlo type I error rates for the pairwise hypothesis ($\mathcal{H}_0$: $\delta_1 = 0$ and $\mathcal{H}_0$: $\delta_2 = 0$) under simple randomization (SR) and constrained randomization (CR). In Web Table 9 and Web Figures 8-9, we summarized the results for power. Alpha level is held at 2.5% throughout the comparison in this section.

WEB TABLE 8 Results under multiplicity adjustment: Type I error rates for the pairwise hypotheses ($\mathcal{H}_0$: $\delta_1 = 0$ and $\mathcal{H}_0$: $\delta_2 = 0$) under simple randomization (SR) versus constrained randomization (CR) with candidate set sizes = 50%, 10%, and 100 of the randomization space. All covariates were used in constrained randomization and the adjusted tests; constrained randomization was implemented using the $l2$ metric; ICC = 0.05; alpha level = 2.5%. The nominal type I error rate is 0.025, and the acceptance range for nominal type I error rate with 10,000 replicates is (0.0215, 0.0285).

			***t*-test**				**Randomization test**			
$\mathcal{H}_0$	# of clusters per arm	Analysis-based adjustment	SR	CR (50%)	CR (10%)	CR (100)	SR	CR (50%)	CR (10%)	CR (100)
$\delta_1 = 0$	$g = 10$	Unadj	0.028	0.003	0.000	0.000	0.023	0.026	0.029	0.024
		Adj-C	0.024	0.027	0.027	0.024	0.024	0.024	0.024	0.024
		Adj-I	0.026	0.027	0.025	0.024	0.024	0.026	0.022	0.025
	$g = 5$	Unadj	0.025	0.004	0.000	0.000	0.025	0.021	0.008	0.000
		Adj-C	0.020	0.021	0.021	0.022	0.022	0.022	0.007	0.000
		Adj-I	0.025	0.024	0.024	0.026	0.022	0.018	0.007	0.000
	$g = 3$	Unadj	0.027	0.006	0.001	0.001	0.000	0.000	0.000	0.000
		Adj-C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Adj-I	0.024	0.024	0.020	0.026	0.000	0.000	0.000	0.000
$\delta_2 = 0$	$g = 10$	Unadj	0.027	0.004	0.000	0.000	0.025	0.025	0.025	0.027
		Adj-C	0.024	0.023	0.023	0.025	0.024	0.023	0.023	0.024
		Adj-I	0.022	0.024	0.024	0.025	0.024	0.024	0.024	0.024
	$g = 5$	Unadj	0.028	0.005	0.000	0.000	0.027	0.020	0.011	0.000
		Adj-C	0.022	0.022	0.021	0.025	0.024	0.019	0.007	0.000
		Adj-I	0.025	0.026	0.023	0.027	0.024	0.018	0.007	0.000
	$g = 3$	Unadj	0.026	0.008	0.002	0.001	0.000	0.000	0.000	0.000
		Adj-C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Adj-I	0.020	0.024	0.021	0.022	0.000	0.000	0.000	0.000



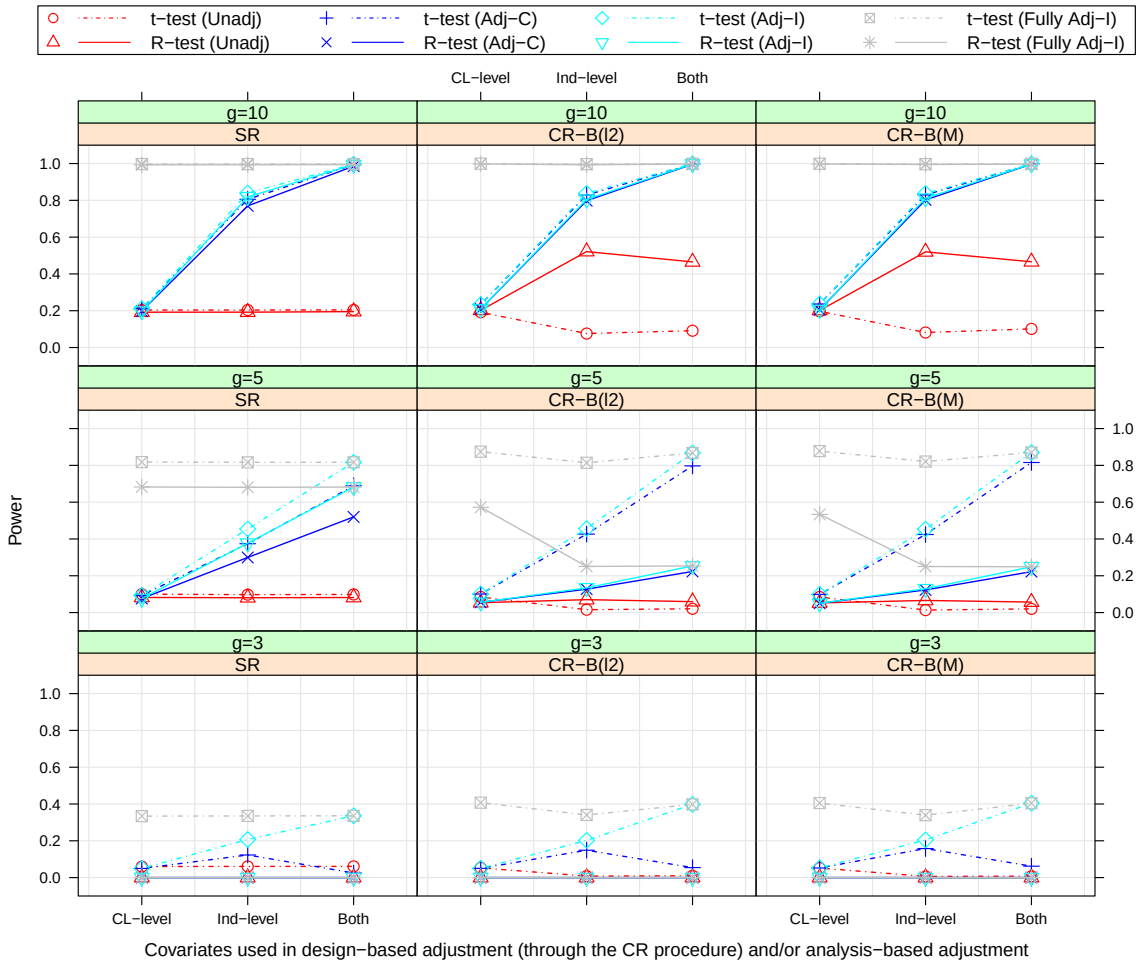
WEB FIGURE 6 Results under multiplicity adjustment: type I error rates for the pairwise hypothesis ($\mathcal{H}_0$: $\delta_1 = 0$) under simple randomization (SR) versus constrained randomization (CR) with 2 balance metrics $B_{(I2)}$ and $B_{(M)}$. CR implemented using covariates indicated on the horizontal axis; candidate set size = 10% under CR; ICC = 0.05; alpha level = 2.5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).



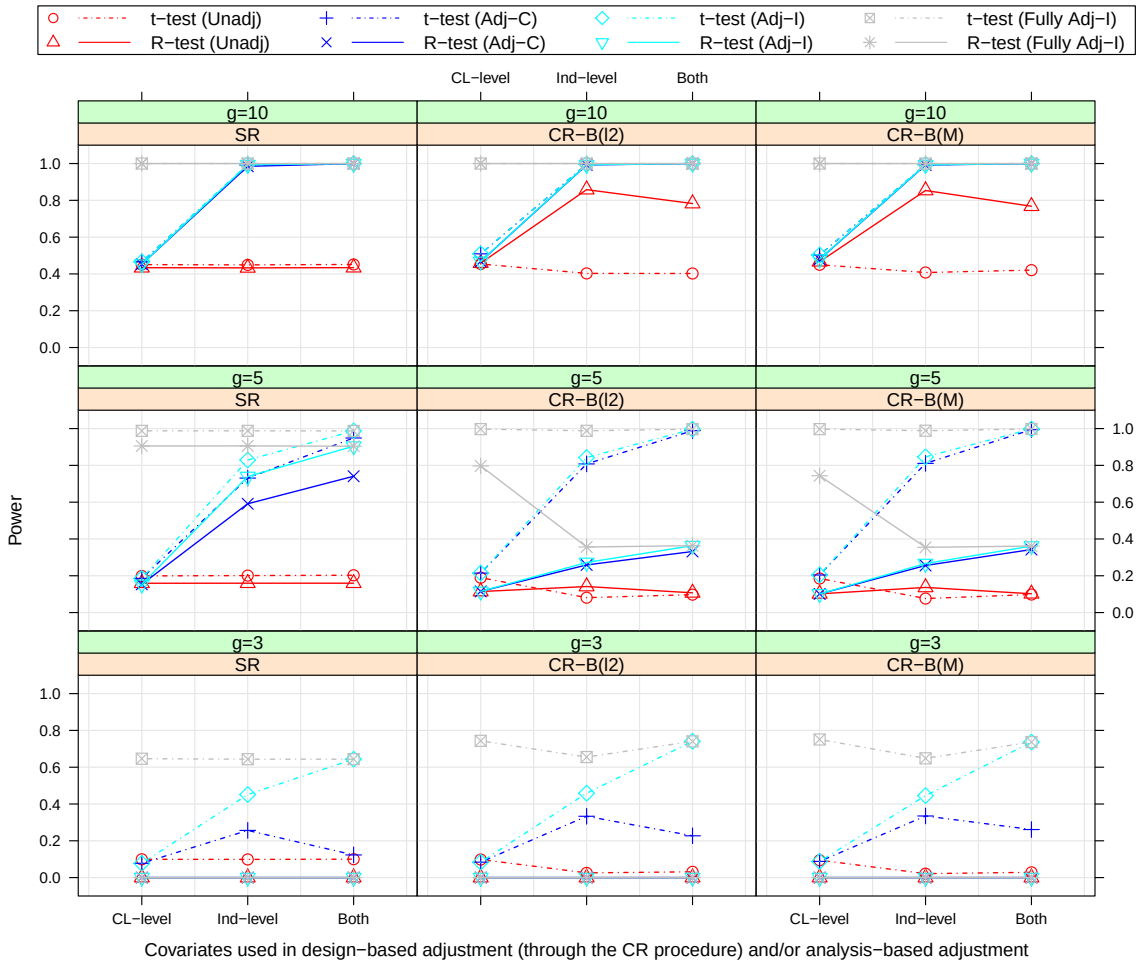
WEB FIGURE 7 Results under multiplicity adjustment: type I error rates for the pairwise hypothesis ($\mathcal{H}_0$: $\delta_2 = 0$) under simple randomization (SR) versus constrained randomization (CR) with 2 balance metrics $B_{(I2)}$ and $B_{(M)}$. CR implemented using covariates indicated on the horizontal axis; candidate set size = 10% under CR; ICC = 0.05; alpha level = 2.5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).

WEB TABLE 9 Results under multiplicity adjustment: Power for the pairwise hypotheses ($\mathcal{H}_0$: $\delta_1 = 0$ and $\mathcal{H}_0$: $\delta_2 = 0$) under simple randomization (SR) versus constrained randomization (CR) with candidate set sizes = 50%, 10%, and 100 of the randomization space. All covariates were used in constrained randomization and the adjusted tests; constrained randomization was implemented using the $l2$ metric; ICC = 0.05; alpha level = 2.5%; power values corresponding to non-nominal type I errors are shaded out.

			***t*-test**				**Randomization test**			
$\mathcal{H}_0$	# of clusters per arm	Analysis-based adjustment	SR	CR (50%)	CR (10%)	CR (100)	SR	CR (50%)	CR (10%)	CR (100)
$\delta_1 = 0$	$g = 10$	Unadj	0.204	0.155	0.091	0.042	0.195	0.313	0.466	0.699
		Adj-C	0.994	0.998	0.998	0.999	0.986	0.994	0.996	0.997
		Adj-I	0.997	0.998	0.999	0.999	0.994	0.997	0.998	0.998
	$g = 5$	Unadj	0.098	0.055	0.021	0.005	0.082	0.097	0.059	0.000
		Adj-C	0.689	0.748	0.797	0.828	0.520	0.497	0.223	0.000
		Adj-I	0.817	0.845	0.867	0.876	0.682	0.621	0.253	0.000
	$g = 3$	Unadj	0.061	0.033	0.010	0.010	0.000	0.000	0.000	0.000
		Adj-C	0.025	0.040	0.054	0.059	0.000	0.000	0.000	0.000
		Adj-I	0.336	0.378	0.398	0.409	0.000	0.000	0.000	0.000
$\delta_2 = 0$	$g = 10$	Unadj	0.452	0.436	0.403	0.378	0.434	0.605	0.782	0.949
		Adj-C	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Adj-I	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$g = 5$	Unadj	0.203	0.151	0.098	0.044	0.160	0.195	0.107	0.000
		Adj-C	0.949	0.978	0.989	0.995	0.741	0.747	0.332	0.000
		Adj-I	0.987	0.995	0.997	0.998	0.904	0.873	0.364	0.000
	$g = 3$	Unadj	0.100	0.066	0.032	0.028	0.000	0.000	0.000	0.000
		Adj-C	0.123	0.164	0.227	0.239	0.000	0.000	0.000	0.000
		Adj-I	0.644	0.701	0.741	0.744	0.000	0.000	0.000	0.000



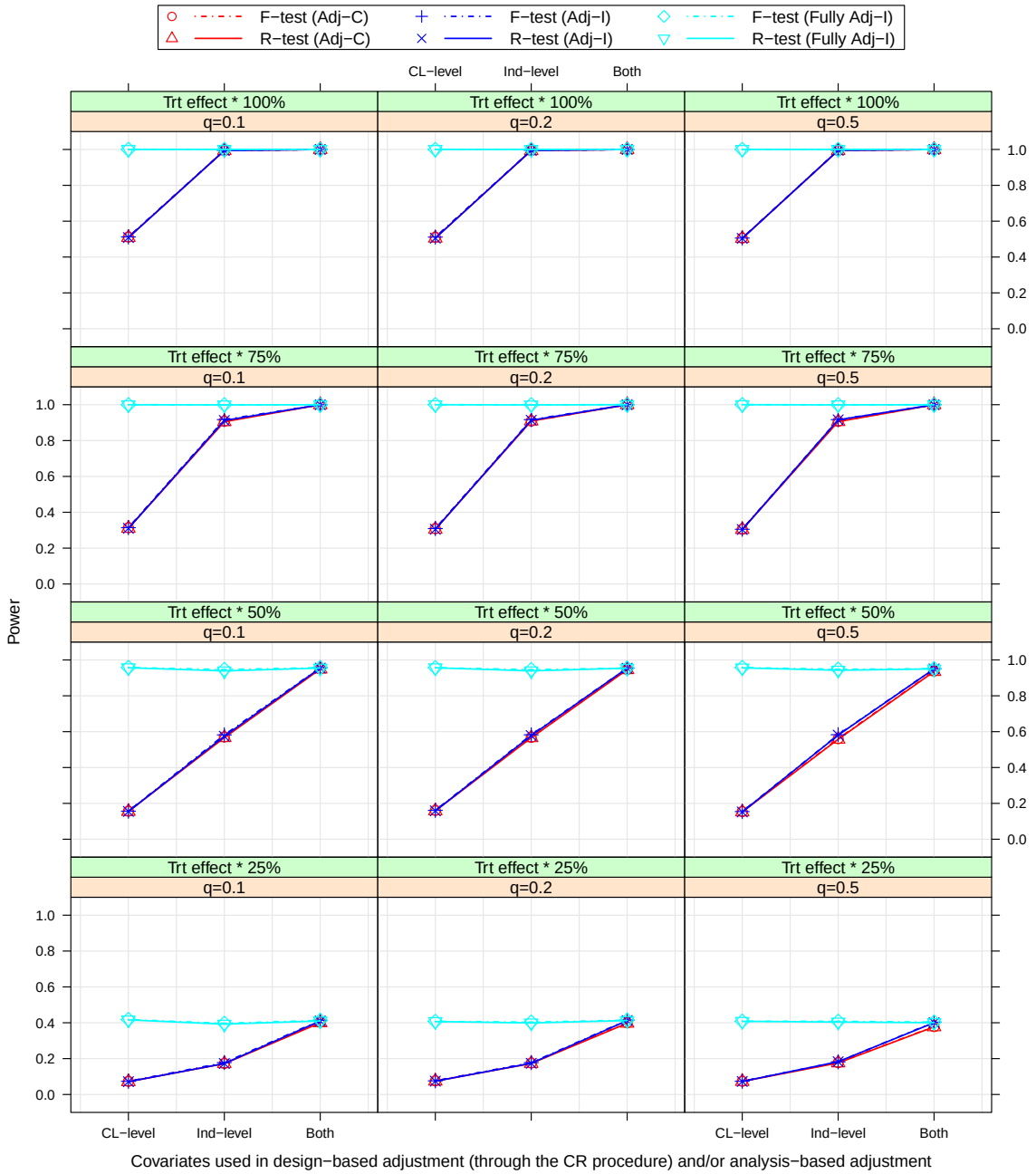
WEB FIGURE 8 Results under multiplicity adjustment: power for the pairwise hypothesis ($\mathcal{H}_0$: $\delta_1 = 0$) under simple randomization (SR) versus constrained randomization (CR) with 2 balance metrics $B_{(l2)}$ and $B_{(M)}$. CR implemented using covariates indicated on the horizontal axis; candidate set size = 10% under CR; ICC = 0.05; alpha level = 2.5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).



WEB FIGURE 9 Results under multiplicity adjustment: power for the pairwise hypothesis ($\mathcal{H}_0$: $\delta_2 = 0$) under simple randomization (SR) versus constrained randomization (CR) with 2 balance metrics $B_{(I2)}$ and $B_{(M)}$. CR implemented using covariates indicated on the horizontal axis; candidate set size = 10% under CR; ICC = 0.05; alpha level = 2.5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).

APPENDIX G: ADDITIONAL SIMULATION RESULTS UNDER REDUCED EFFECT SIZES

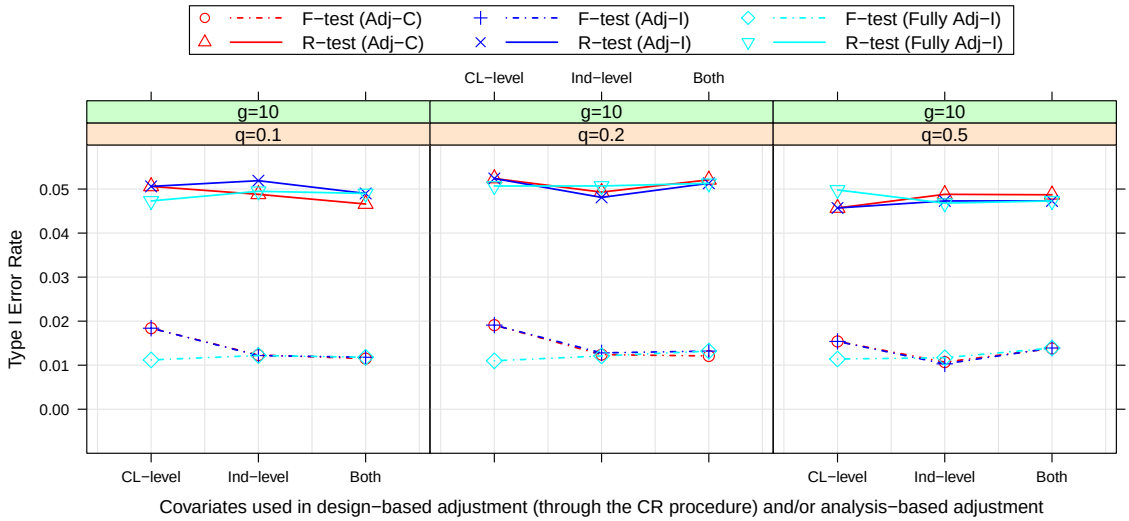
We presented the simulation results with $g = 3, 5, 10$ in the main text. It was shown that the adjusted model-based tests and the randomization-based tests (UMPR/LMPR tests) provided a similar level of power when the number of clusters per arm is large. However, with $g = 10$, the power for both types of tests reaches the maximum (100% power), making the comparison uninformative. Therefore, we presented additional simulation results for power with effect sizes reduced to 75%, 50%, and 25% of the original magnitude to better compare the power of the UMPR/LMPR tests with that of the model-based tests. We showed the power for the global hypothesis in Web Figure 10. Results for the pairwise hypotheses are qualitatively similar, thus omitted for presentation. These results demonstrate that the randomization test is asymptotically equivalent to the model-based test in terms of statistical power.



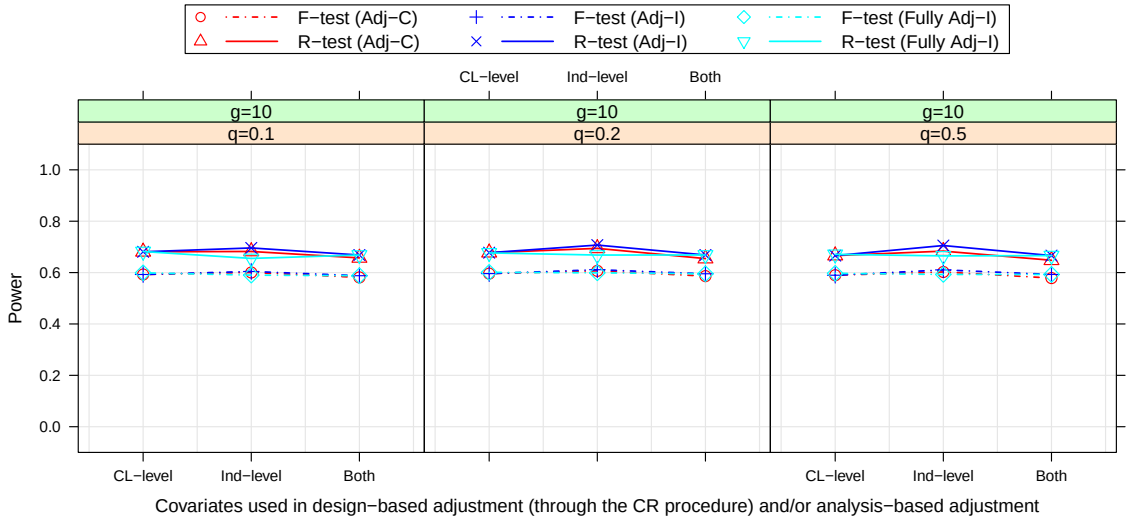
WEB FIGURE 10 Results under reduced effect sizes (75%, 50%, and 25% of the original magnitude): power for the global hypothesis ($\mathcal{H}_0$: $\delta_1 = \delta_2 = 0$) under constrained randomization (CR) with $B_{(M)}$ balance metric and $q = 0.1$, 0.2, and 0.5. CR implemented using covariates indicated on the horizontal axis; ICC = 0.05; alpha level = 5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).

APPENDIX H: ADDITIONAL SIMULATION RESULTS UNDER NON-NORMAL DATA GENERATION PROCESSES

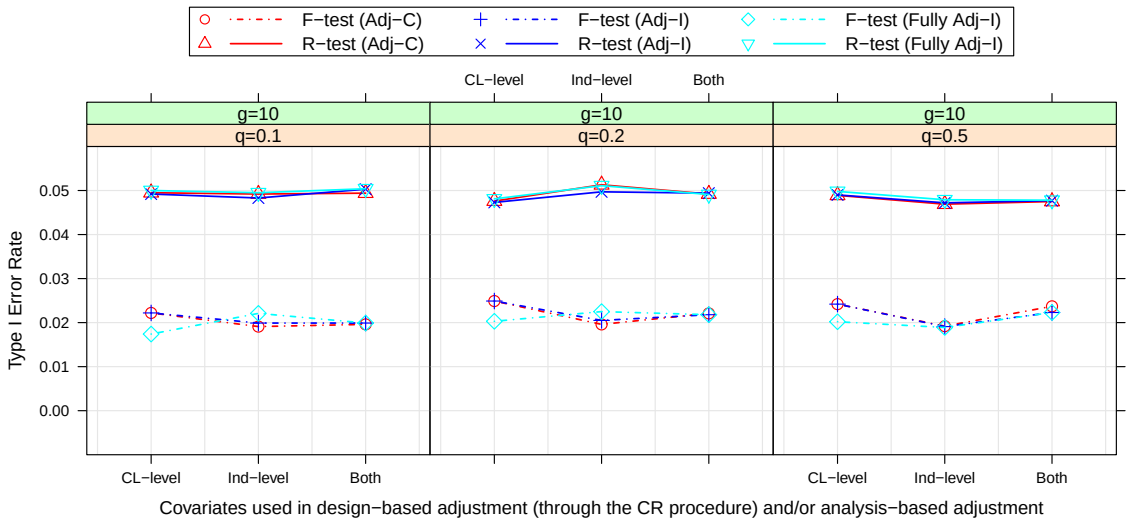
To evaluate the robustness of the analytical methods under comparison against violations of distributional assumptions, we generated non-normal data in a similar way to Small et al.[6], specifying the random cluster effect γ_j and error term ϵ_{jk} to follow the standard Cauchy distribution, respectively. Note that the standard Cauchy distribution has a thicker tail than the normal distributions. Treatment effects were set to 8 for both treatment arms to evaluate power. Number of clusters per arm g was held at 10. Results of type I error rates and power under Cauchy residual and random cluster effect are presented in Web Figures 11-18. In terms of type I error rates, the model-based tests are overly conservative in either scenario, while the randomization test (both UMPR and LMPR, depending on the hypothesis of interest) maintains the nominal test size. For power, both tests are negatively affected by the violations of normality assumptions. However, the randomization test can be more powerful than the model-based test. In summary, the randomization test is a more flexible and robust alternative to the model-based test when distributional assumptions are not guaranteed to hold.



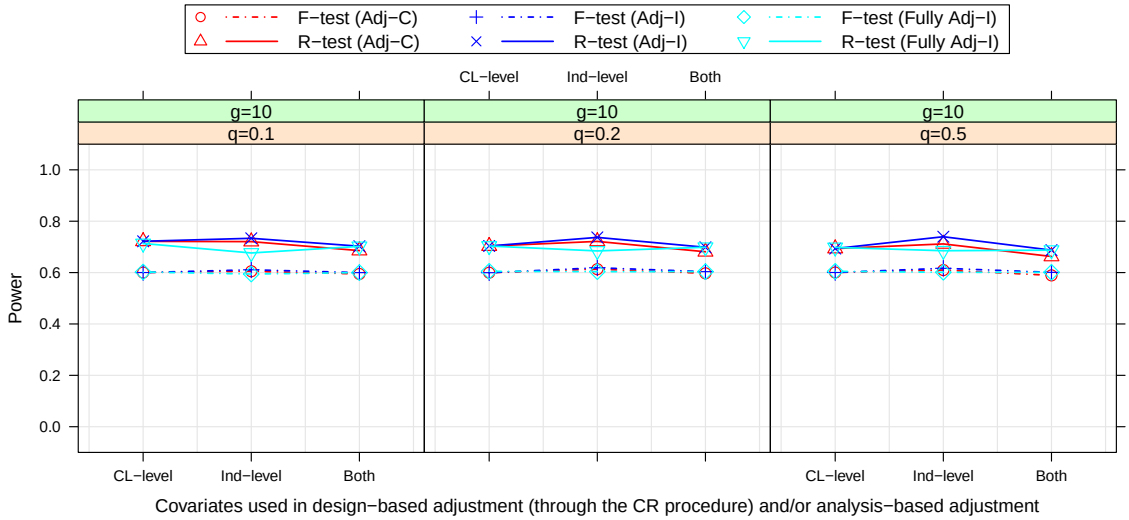
WEB FIGURE 11 Results under Cauchy residual: type I error rate for the global hypothesis ($\mathcal{H}_0$: $\delta_1 = \delta_2 = 0$) under constrained randomization (CR) with $B_{(M)}$ balance metric and $q = 0.1$, 0.2, and 0.5. CR implemented using covariates indicated on the horizontal axis; alpha level = 5%; R-test: randomization test; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).



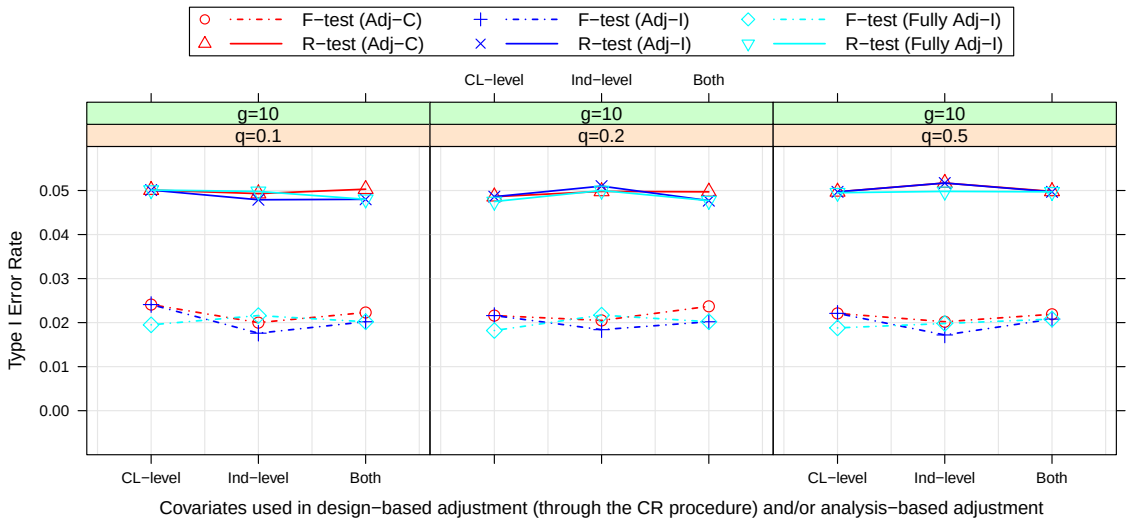
WEB FIGURE 12 Results under Cauchy residual: power for the global hypothesis ($\mathcal{H}_0$: $\delta_1 = \delta_2 = 0$) under constrained randomization (CR) with $B_{(M)}$ balance metric and $q = 0.1$, 0.2, and 0.5. CR implemented using covariates indicated on the horizontal axis; alpha level = 5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).



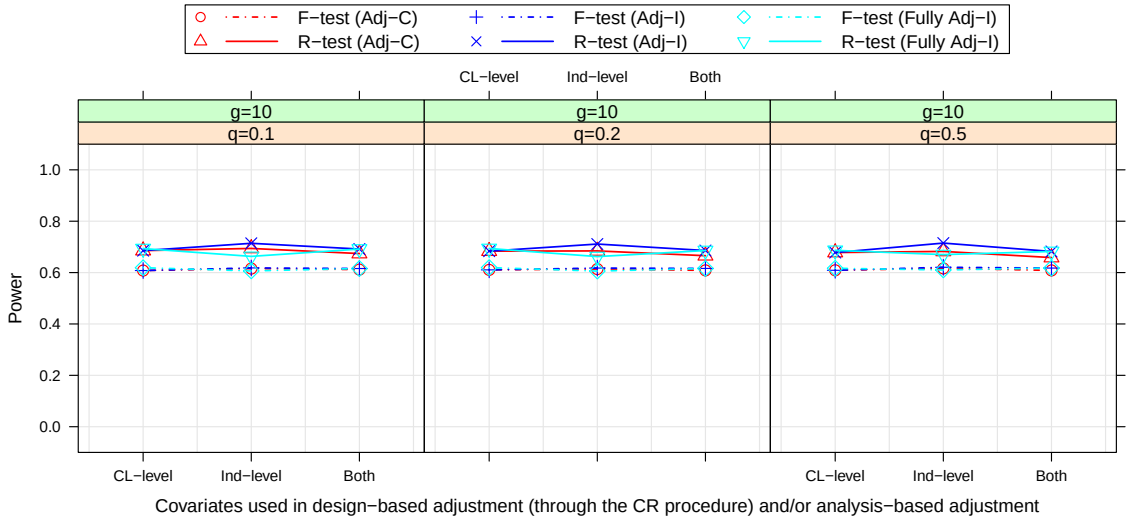
WEB FIGURE 13 Results under Cauchy residual: type I error rate for the pairwise hypothesis ($\mathcal{H}_0$: $\delta_1 = 0$) under constrained randomization (CR) with $B_{(M)}$ balance metric and $q = 0.1$, 0.2, and 0.5. CR implemented using covariates indicated on the horizontal axis; alpha level = 5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).



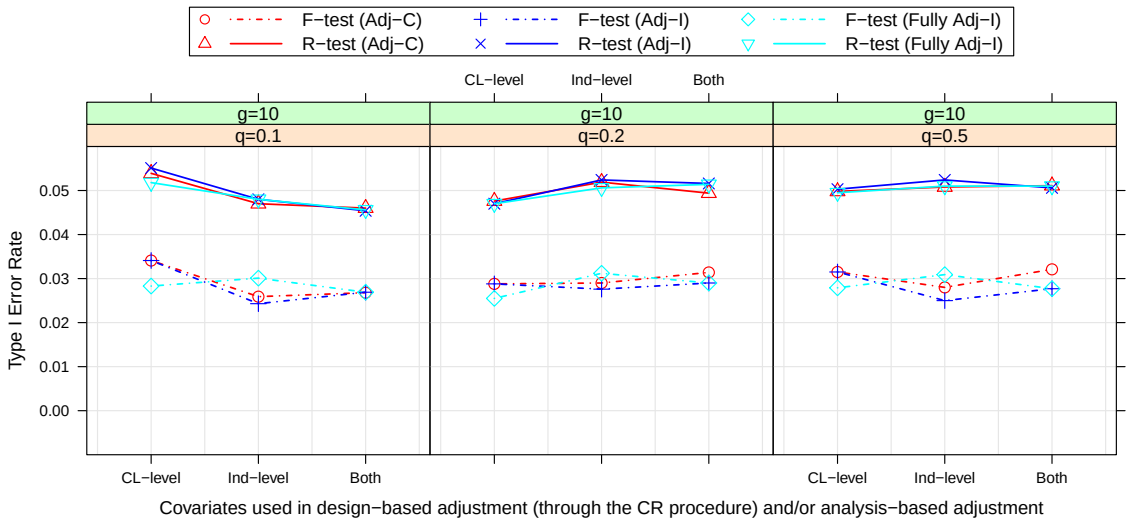
WEB FIGURE 14 Results under Cauchy residual: power for the pairwise hypothesis ($\mathcal{H}_0$: $\delta_1 = 0$) under constrained randomization (CR) with $B_{(M)}$ balance metric and $q = 0.1$, 0.2, and 0.5. CR implemented using covariates indicated on the horizontal axis; alpha level = 5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).



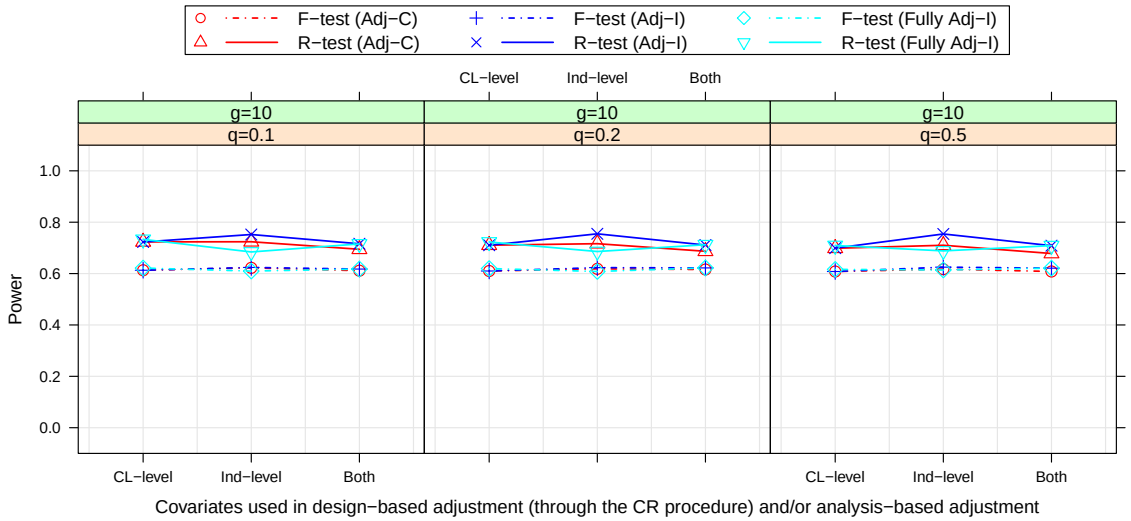
WEB FIGURE 15 Results under Cauchy random cluster effect: type I error rate for the global hypothesis ($\mathcal{H}_0$: $\delta_1 = \delta_2 = 0$) under constrained randomization (CR) with $B_{(M)}$ balance metric and $q = 0.1$, 0.2, and 0.5. CR implemented using covariates indicated on the horizontal axis; alpha level = 5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).



WEB FIGURE 16 Results under Cauchy random cluster effect: power for the global hypothesis ($\mathcal{H}_0$: $\delta_1 = \delta_2 = 0$) under constrained randomization (CR) with $B_{(M)}$ balance metric and $q = 0.1$, 0.2, and 0.5. CR implemented using covariates indicated on the horizontal axis; alpha level = 5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).



WEB FIGURE 17 Results under Cauchy random cluster effect: type I error rate for the pairwise hypothesis ($\mathcal{H}_0$: $\delta_1 = 0$) under constrained randomization (CR) with $B_{(M)}$ balance metric and $q = 0.1$, 0.2, and 0.5. CR implemented using covariates indicated on the horizontal axis; alpha level = 5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).



WEB FIGURE 18 Results under Cauchy random cluster effect: power for the pairwise hypothesis (H_0: $\delta_1 = 0$) under constrained randomization (CR) with $B_{(M)}$ balance metric and $q = 0.1$, 0.2, and 0.5. CR implemented using covariates indicated on the horizontal axis; alpha level = 5%; R-test: randomization test; CL-level: cluster-level covariates, $\boldsymbol{x}_j$; Ind-level: individual-level covariates, $\boldsymbol{z}_{jk}$; Unadj: unadjusted test; Adj-C: test adjusted for the covariates on the horizontal axis (with individual-level covariates aggregated at the cluster level); Adj-I: test adjusted for the covariates on the horizontal axis (with actual individual-level covariates); Fully Adj-I: test adjusted for all four covariates (with actual individual-level covariates).

References

1. Ciolino JD, Diebold A, Jensen JK, Rouleau GW, Koloms KK, Tandon D. Choosing an imbalance metric for covariate-constrained randomization in multiple-arm cluster-randomized trials. *Trials* 2019; 20(1): 293.

2. Watson SI, Girling A, Hemming K. Design and analysis of three-arm parallel cluster randomized trials with small numbers of clusters. *Statistics in Medicine* 2020. doi: https://doi.org/10.1002/sim.8828

3. Braun TM, Feng Z. Optimal permutation tests for the analysis of group randomized trials. *Journal of the American Statistical Association* 2001; 96(456): 1424–1432.

4. Li F, Turner EL, Preisser JS. Sample size determination for GEE analyses of stepped wedge cluster randomized trials. *Biometrics* 2018; 74(4): 1450–1458.

5. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods.. *American journal of public health* 1996; 86(5): 726–728.

6. Small DS, Ten Have TR, Rosenbaum PR. Randomization inference in a group–randomized trial of treatments for depression: covariate adjustment, noncompliance, and quantile effects. *Journal of the American Statistical Association* 2008; 103(481): 271–279.