INFERENCe FOR A TWO-STAGE ENRICHMENT DESIGN

BY ZHANTAO LIN¹, NANCY FLOURNOY²,† AND WILLIAM F. ROSENBERGER¹,*

¹Department of Statistics, George Mason University, zLin4@masonlive.gmu.edu; *wrosenbe@gmu.edu
²Department of Statistics, University of Missouri, Columbia, †flournoyn@missouri.edu

Two-stage enrichment designs can be used to target the benefiting population in clinical trials based on patients’ biomarkers. In the case of continuous biomarkers, we show that using a bivariate model that treats biomarkers as random variables more accurately identifies a treatment-benefiting enriched population than assuming biomarkers are fixed. Additionally, we show that under the bivariate model, the maximum likelihood estimators (MLEs) follow a randomly scaled mixture of normal distributions. Using random normings, we obtain asymptotically standard normal MLEs and construct hypothesis tests. Finally, in a simulation study, we demonstrate that our proposed design is more powerful than a single stage design when outcomes and biomarkers are correlated; the model-based estimators have smaller bias and mean square error (MSE) than weighted average estimators.

1. Introduction.

1.1. Enrichment strategies and biomarkers. Precision medicine can be thought of as the tailoring of treatments based on individual characteristics, or biomarkers. Enrichment is a key strategy in the development of clinical trials for precision medicine. Enrichment designs rely on the selection of patients for clinical trials using one or more biomarkers, and aim to demonstrate the safety and/or effectiveness of the drug in selected populations. In some situations, an appropriate binary biomarker can be easily evaluated and used to restrict the entry of patients into phase III trials, together with the pathophysiology of the disease and the mechanism of drug action (e.g. Fong et al., 2009; Audeh et al., 2010; Chapman et al., 2011). Simon and Maitournam (2004), Maitournam and Simon (2005), and Mandrekar and Sargent (2009) studied such targeted designs. However, in other cases, we do not know their predictive strength in humans, although we may have some candidate biomarkers. In this case, the population may be enriched inappropriately, which leads to a too-narrow benefiting population or exposure to an ineffective treatment. Therefore, adaptive enrichment designs, which use the accrued information on previous subjects’ response to treatments to find the benefiting population, have been considered.

Adaptive enrichment designs have been proposed when subgroups of patients are prespecified (e.g. Russek-Cohen and Simon, 1997; Jennison and Turnbull, 2007; Wang, O’Neill and Hung, 2007; Wang, James Hung and O’Neill, 2009; Rosenblum and van der Laan, 2011; Friede, Parsons and Stallard, 2012; Stallard et al., 2014; Rosenblum, Fang and Liu, 2020). Multivariate normal models are applied to account for outcomes of different subgroups of patients and to conduct inference. This type of design is useful only in the case of limited subgroups (Simon, 2015).

When a candidate biomarker is measured on a continuous scale without a known dichotomy, a threshold that identifies the enrichment group must be defined. Simon and Simon (2013, 2017) and Spencer et al. (2016) provided several enrichment designs that include

MSC 2010 subject classifications: Primary 00X00, 00X00; secondary 00X00

Keywords and phrases: Adaptive designs, design of experiments, inference for stochastic processes, precision medicine, random biomarker, threshold determination.
procedures for identifying the threshold and for testing the existence of a benefiting subpopulation; they showed their designs to be more powerful than non-adaptive procedures. In the designs of Simon and Simon (2013, 2017), the threshold for future use is selected at an interim analysis, based on first stage data using prespecified criteria. Spencer et al. (2016) chose the threshold for the study’s enriched second stage from several candidate cutpoints; then they chose a threshold for future use from a continuous range of possibilities based on data from both stages. At the end of their designs, rejecting the hypothesis that treatment effects are the same for all candidate biomarker values indicates that there exists a subpopulation for which the treatment effect is significantly different from that of the full population. But this test does not establish that the treatment effect and biomarkers are correlated. This means that the threshold may identify a subpopulation with a distinctive biomarker profile; but if the biomarker is, at most, weakly related to the outcome of primary interest, the treatment may be efficacious to a larger population than is prescribed by the threshold. Following Simon and Simon (2013), Simon and Simon (2017) and Zhang et al. (2018) estimate and test the treatment effect for the target population identified with data from the first stage. For inference at the end of the study, the randomness of biomarkers is not considered. To take the randomness of biomarkers into account, we address this problem differently.

1.2. Modelling and inference. We derive closed-form test statistics and the asymptotic distributions of estimators for our proposed design. A procedure is presented in Section 2.2 for selecting a threshold that more precisely identifies the benefiting subpopulation. With the bivariate model, the treatment effect for the full population and the correlation between the outcome and a single continuous biomarker may be estimated simultaneously. The correlation provides an explicit measure of the value of the biomarker as a surrogate for predicting the treatment effect in the subpopulation and it can be used to predict what the treatment effect would be in the full population. In addition, the threshold is estimated directly on a continuous scale at the interim analysis and this estimate can be updated at the end of the study to provide better subpopulation selection than can be obtained from the interim analysis.

The treatment’s target population will be defined by the threshold estimated at the study end rather than at the interim analysis (Simon and Simon, 2013, 2017; Zhang et al., 2018). Finally, the treatment effect for the target population is estimated based on data from both stages, and a confidence interval for the treatment effect is available for the target population.

Even though sample sizes $n_1$ and $n_2$ of the two stages are fixed at the beginning of the study, the distributional properties of statistics are complicated because using the random first stage threshold to screen second stage patients induces dependence. This important distributional feature disappears when approximations are made by letting both sample sizes tend to infinity; while letting $n_1$ alone tend to infinity makes stage 2 data irrelevant. Lane, Yao and Flournoy (2014) show that only allowing the second stage sample size to tend to infinity provides better approximations to finite sample distributions than letting both sample sizes tend to infinity, and this is the approach taken in this paper. Consequently, the joint asymptotic distribution of MLEs from such enrichment studies is a randomly scaled mixture of normal distributions, which is shown in Lemma 3.1 in Section 3.2. We derive the asymptotic distributions of test statistics using random information measures as described by Barndorff-Nielsen and Sørensen (1994). The net result is a complete package for design and analysis in closed-form, which should facilitate its use in practice.

1.3. Organization. The remainder of this article is organized as follows. In Section 2, the design procedure based on the bivariate model is described. In Section 3, we describe a set of hypothesis tests that identify the subpopulation for stage 2, if it exists, and give the
closed-form test statistics and treatment effect estimators that provide a complete template for analyzing such a trial. The proposed design is compared with a single stage design in the simulation study under different scenarios in Section 4. The extension of the proposed design to a parallel group design considering prognostic and predictive biomarkers is discussed in Section 5. We conclude this article with a discussion in Section 6. Proofs of all results are given in the Appendix.

2. A two-stage threshold enrichment design. A two-stage enrichment threshold design with a continuous outcome is described in this section. The objectives of the design are to identify a subpopulation with clinical significance, if it exists, and to estimate the treatment effect for the target population derived. We assume that a single biomarker $x$ measured on a continuous scale has a linear relationship with the treatment outcome, or none. We also assume that a larger biomarker value implies a better outcome if a relationship exists. The sample size of the two stages $n_1$ and $n_2$ are fixed at the beginning of the study, although the hypothesis testing at the end of the study is based on asymptotic properties as $n_2 \to \infty$. At stage 1, patients are recruited from the full population. With an interim analysis, stage 1 data are used to choose the threshold that must be met for patients’ recruitment to stage 2. In this paper, the biomarker threshold is selected to be the point at which the standardized treatment effect $\lambda$ for the subpopulation is equal to some prespecified constant $c$; other criteria could be chosen. Because larger biomarker values are better, patients whose biomarker is above the threshold are enrolled into stage 2.

At the end of the study, the primary aim is to test whether there exists a threshold satisfying the criterion. When the target population defined by the threshold is determined to exist, the threshold for future use and the treatment effect for this target population will be estimated.

Assuming a bivariate normal model, the outcome $y$ and the biomarker $x$ are normally distributed, and the expectation of treatment effect, given $x$, is a linear function of $x$. The advantage of this model is discussed in the following section.

2.1. The threshold in the bivariate model. In this section, the bivariate model for the outcome and the biomarker is introduced. Then this model is contrasted to analogous ones in the literature, that is, with biomarkers assumed to be known [i.e., fixed as discussed by Rosset and Tibshirani (2020)]. The bivariate normal distribution has correlation $\rho$, mean vector $(\mu_Y, \mu_X)$, and variance-covariance matrix

$$
\begin{pmatrix}
\sigma_Y^2 & \rho \sigma_Y \sigma_X \\
\rho \sigma_Y \sigma_X & \sigma_X^2
\end{pmatrix}.
$$

It is well known that the conditional density of $Y$, given $X$ and denoted $f_{Y|X}(y|x)$, is normally distributed with mean $\mu_Y + \rho \sigma_Y (x - \mu_X)/\sigma_X$ and variance $(1 - \rho^2)\sigma_Y^2$.

Depending on the goals of a clinical trial, what constitutes a treatment effect may differ. In our treatment, the threshold and the benefiting population are derived on the basis of the definition of the treatment effect. When comparing to a historical treatment effect $\mu_h$, four different metrics can be described:

1. Minimum subpopulation treatment effect: $E(Y|x^*) - \mu_h$, where $E(Y|x^*)$ is the conditional expectation with respect to the conditional density $f_{Y|X}(y|x)$ and $x^*$ is a selected threshold.

2. Standardized minimum subpopulation treatment effect:

$$
\frac{E(Y|x^*) - \mu_h}{\sqrt{Var(Y|x^*)}}.
$$

where $Var(Y|x^*)$ is the conditional variance with respect to the conditional density $f_{Y|X}(y|x)$.
3. Average subpopulation treatment effect: \( E(Y|X > x^*) - \mu_h. \)

4. Standardized average subpopulation treatment effect:

\[
\frac{E(Y|X > x^*) - \mu_h}{\sqrt{\text{Var}(Y|X > x^*)}}. 
\]

Patient ethics may lead to adopting the first two metrics, which consider the treatment effect of the new treatment at the threshold. However, investigators or regulators are likely to be interested in the average treatment effect of the subpopulation. Definition (2.1) is considered in this paper, which is the most mathematical challenging.

The subpopulation, for which the standardized treatment effect (2.1) is equal to a prespecified clinical significance \( c \), is of interest. If such a threshold \( x^* \) exists and is used to screen patients for the treatment, then the joint distribution of \( X \) and \( Y \) has a truncated bivariate normal distribution with \( x > x^* \), \( -\infty < y < +\infty \). By (2.1), the threshold \( x^* \) satisfies

\[
\lambda(x^*) = \frac{E(Y|X > x^*) - \mu_h}{\sqrt{\text{Var}(Y|X > x^*)}} = \frac{\int_{x^*}^{+\infty} E(Y|x|x > x^*)dx - \mu_h}{\sqrt{\text{Var}(Y|X > x^*)}} = \frac{\mu_Y - \rho\sigma_Y \mu_X / \sigma_X + \rho \sigma_Y / \sigma_X \int_{x^*}^{+\infty} f(x|x > x^*)dx - \mu_h}{\sigma_Y \sqrt{1 - \rho^2 \left(1 - \int_{x^*}^{+\infty} f(x|x > x^*)dx/\sigma_Y^2 + \int_{x^*}^{+\infty} f(x|x > x^*)dx/\sigma_X^2\right)^2}} = c,
\]

where \( f(x|x > x^*) \) is the marginal probability density of the truncated bivariate normal distribution. Then enriched population consists of patients with \( x \geq x^* \).

The fixed biomarker model analogous to the one proposed here is

\[
Y|x = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2),
\]

where \( \beta_0 = \mu_Y - \rho \mu_X \sigma_Y / \sigma_X \) and \( \beta_1 = \rho \sigma_Y / \sigma_X \). For the same prespecified clinical significance \( c \), the threshold \( x^* \) satisfies

\[
\lambda(x^*) = \frac{\beta_0 + \beta_1 \mu_X > x^* - \mu_h}{\sigma_Y} = c,
\]

where \( \mu_{x > x^*} \) is the mean of biomarkers in the subpopulation. Based on Model (2.2), a homogeneous variance is assumed across the population, so the denominator of Equation (2.3) is \( \sigma_Y \). Note that, the fixed biomarker model overstates the variance of \( y \) by a factor of

\[
1 - \rho^2 \left(1 - \int_{x^*}^{+\infty} (x - \mu_X)^2 f(x|x > x^*)dx/\sigma_X^2 + \int_{x^*}^{+\infty} (x - \mu_X) f(x|x > x^*)dx/\sigma_X^2\right)^2.
\]

Consequently assuming fixed biomarkers, the enriched population identified by (2.3) is too conservative.

Figure 1 displays the average sample effect size, \( \lambda_{\text{sample mean/standard deviation}} \) for the subpopulation, using formulae for the thresholds for the random and fixed \( x \) models. As the correlation \( \rho \) increases, the average sample effect size based on the fixed \( x \) threshold increases and is greater than the target effect size \( c \), while the average effect size based on the random \( x \) threshold stays around the target. Moreover, the deviation from the target grows with larger target effect sizes using the fixed \( x \) model. This motivates basing the enrichment design on a bivariate model.
Fig 1: The average of the sample effect sizes $\hat{\lambda} = \text{sample mean/standard deviation}$ for the subpopulation based on 10,000 iterations. The data are generated from a truncated bivariate normal distribution with $\mu_X = 2$, $\mu_Y = 0.5$, $\sigma_X = \sigma_Y = 1$. The true effect size $c = \{0.2, 0.5, 0.8\}$, $\mu_h = 0.5$, $n = 500$, and thresholds are based on random and fixed $x$ models.

2.2. *Interim analysis.* As in Simon and Simon (2013) and Spencer et al. (2016), the biomarker threshold is defined in terms of a quantile of the biomarker distribution. Let $\Phi(\cdot)$ denote the cumulative standard normal distribution. The transformed biomarkers $U_j = \Phi \left( \left( X_j - \mu_X \right) / \sigma_X \right)$, $j = 1, \ldots, n$, are a sample from a $U(0, 1)$ distribution. Because the quantile and the biomarker values are one to one, the conditional distribution of $Y$ given $x$ is equivalent to that of $Y$ given $u$:

$$Y \mid u = \mu_Y + \rho \Phi^{-1}(u) \sigma_Y + \epsilon, \epsilon \sim N \left( 0, \sigma_Y^2 (1 - \rho^2) \right).$$

After the first stage, an interim analysis is carried out to estimate the threshold. As we assume that a larger biomarker implies a better outcome, the threshold $u^*$ is

$$u^* = \min \{ u \mid \lambda(u) \geq c \} = \min \left\{ u \left| \frac{\mu_Y + \rho \sigma_Y \Phi^{-1}(u) / (1 - \rho^2) - \mu_h}{\sqrt{\text{Var}(Y \mid U > u)}} \geq c \right. \right\},$$

where

$$\text{Var}(Y \mid U > u) = \sigma_Y \sqrt{1 - \rho^2 \left( 1 - \int_u^1 \Phi^{-1}(u)^2 / (1 - u) du + \left[ \int_u^1 \Phi^{-1}(u) / (1 - u) du \right]^2 \right)}.$$
Figure 2 shows distributions of the estimated threshold, $\hat{u}^*_1$, at the interim analysis for selected stage 1 sample sizes. As expected, the estimated threshold is more likely to be close to the true value as $n_1$ increases. When the first stage sample size $n_1$ is small, $\hat{u}^*_1 = 1$ may occur when $u^*_1$ is well below 1.0 because of the randomness, in which case the study would stop. To avoid stopping at stage 1, which decreases the power of tests and the precision of estimators, an upper bound on the threshold $\pi^*$ can be defined. If the first stage sample size $n_1$ is large enough the estimated threshold $\hat{u}^*_1$ is very close to the true value in which case the upper bound on the threshold $\pi^*$ can be set close or equal to 1. However, if $n_1$ is too small, the choice of the upper bound will be difficult. A bad choice will result in a poor estimate of the average treatment effect of the derived target population at the end of the design.

 Patients with the biomarker quantiles above

$$\hat{u}^* = \min\{\hat{u}^*_1, \pi^*\}$$

are enrolled in the second stage. Thus, second stage data depends on the first stage data through $\hat{u}^*$. The conditional density of second stage outcomes $f(y|u, \hat{u}^*)$ is given in Appendix A.1. The criterion can also be modified to consider early stopping, but this may sacrifice power.

Fig 2: Density of the estimated threshold, $\hat{u}^*_1$, at the interim analysis, based on 100,000 iterations. The true threshold is $u^* = 0.71$. The data are generated from a bivariate normal distribution with $\mu_X = 2$, $\mu_Y = 0.5$, $\sigma_X = \sigma_Y = 1$, $\rho = 0.4$, $\mu_h = 0.5$, $c = 0.5$.

3. Inference at the end of the study.

3.1. Parametric regions of interest. The important question is ‘does there exist a subpopulation defined by a threshold $u^*$ of the biomarker quantiles that has a clinically significant
standardized treatment effect?’. Specifically, let \( \lambda(u^*) \) denote the standardized treatment effect for patients with biomarker quantile above \( u^* \), and let \( c \) be a clinically significant effect size. Subhypotheses of interest are

\[
H_{01} : \mu_Y \leq c\sigma_Y + \mu_h \text{ vs. } H_{A1} : \mu_Y > c\sigma_Y + \mu_h,
\]

and

\[
H_{02} : \rho = 0 \text{ vs. } H_{A2} : \rho > 0.
\]

These are the parametric regions of interest:

1. Both \( H_{01} \) and \( H_{02} \) are true: the treatment effect for the whole population is not clinically significant and outcomes are independent of biomarker values. So there is no clinically significant subpopulation that can be defined by biomarker values.
2. \( H_{01} \) is false and \( H_{02} \) is true: the treatment effect for the whole population is clinically significant, but outcomes and biomarkers are independent. The target population is the whole population.
3. \( H_{01} \) is true and \( H_{02} \) is false: the treatment effect for the whole population is not clinically significant, but the outcome and the biomarker are correlated. So there exists a threshold \( u^* \) such that

\[
\lambda(u^*) = c.
\]

This is the most meaningful scenario where the enrichment design shows advantages. The MLE of the threshold at the end of the study defines the target population for future use. The treatment effect for this target population \( \delta(u^*_f) \) will be estimated, and the confidence interval will be given.
4. Both \( H_{01} \) and \( H_{02} \) are false: the treatment effect for the whole population is clinically significant and study outcomes are correlated with biomarker values. The target population is the whole population, and biomarker values are useful for predicting outcomes.

Scenario 1 is a negative study result, and scenarios 2-4 are positive. However, if only scenario 1 is tested in favor of scenarios 2-4, the target population would be narrowed while the whole population is clinically significant. To avoid restricting the target population too much, we propose testing \( H_{01} \) and \( H_{02} \) separately. To control the family-wise type I error rate at level \( \alpha \), a Bonferroni correction is applied in this paper. Other, less conservative, multiple testing adjustments might also be considered.

3.2. Asymptotic likelihood theory. To test \( H_{01} \) and \( H_{02} \) separately at the end of this proposed design, the asymptotic marginal distributions of the MLEs \( (\hat{\mu}_n, \hat{\rho}_n) \) of \( (\mu_Y, \rho) \) are derived. By employing the Cramér-Wold device, the asymptotic marginal distributions are derived from the asymptotic joint distribution. Let \( \theta = (\mu_Y, \rho); \hat{\theta} = (\hat{\mu}_n, \hat{\rho}_n) \) are their MLEs. Let \( n = n_1 + n_2 \), where \( n_2 \) is the number of patients screened to have \( u > \hat{u}^* \). In this design, \( F_n = \sigma(x_1, \ldots, x_{n_1}) \otimes \sigma(y_1, \ldots, y_{n_1}) \) is a sub \( \sigma \)-field of \( F_n = \sigma(x_1, \ldots, x_n) \otimes \sigma(y_1, \ldots, y_n) \), so \( \hat{u}^* \) is \( F_n \)-measurable.

3.2.1. The likelihood, score vector and its derivatives. For simplicity, we have omitted notation for patients who had biomarkers tested after stage 1 but whose values were below the threshold, but still the likelihood function must account for this patient selection as is done with the indicator function below:

\[
L_n(\theta|y_1, \ldots, y_n, u_1, \ldots, u_n) = \prod_{j=1}^{n_1} f(y_j|u_j)f(u_j) \prod_{j=n_1+1}^{n} f(y_j, u_j|y_1, \ldots, y_{n_1}, u_1, \ldots, u_{n_1})
\]
\[ \frac{1}{\sqrt{2\pi\sigma_Y}} \exp\left\{ -\frac{(y_j - \mu_Y - \rho\Phi^{-1}(u_j)\sigma_Y)^2}{2\sigma_Y^2(1 - \rho^2)} \right\} \]

\[ \frac{1}{\sqrt{2\pi\sigma_Y}} \exp\left\{ -\frac{(y_j - \mu_Y - \rho\Phi^{-1}(u_j)\sigma_Y)^2}{2\sigma_Y^2(1 - \rho^2)} \right\} \}

\[ I(u > \hat{\alpha}^*) \]

The score function vector \( S_n(\theta) = [\partial \log L_n(\theta)/\partial \mu_Y, \partial \log L_n(\theta)/\partial \rho] : [S^1_n(\theta), S^2_n(\theta)] \)

decomposes into score functions for the two stages separately; that is, \( S^k_n(\theta) = S^k_{n1}(\theta) + S^k_{n2}(\theta) \), \( k = 1, 2 \), where the superscript \( k \) = 1 denotes derivatives with respect to \( \mu_Y \) and \( k = 2 \) denotes derivatives with respect to \( \rho \).

The Hessian matrix is given by

\[ \hat{\mathbf{H}}_n(\theta) = \begin{pmatrix} \hat{S}^{(1,1)}_n(\theta) & \hat{S}^{(1,2)}_n(\theta) \\ \hat{S}^{(2,1)}_n(\theta) & \hat{S}^{(2,2)}_n(\theta) \end{pmatrix} + \begin{pmatrix} \hat{S}^{(1,1)}_{n2}(\theta) & \hat{S}^{(1,2)}_{n2}(\theta) \\ \hat{S}^{(2,1)}_{n2}(\theta) & \hat{S}^{(2,2)}_{n2}(\theta) \end{pmatrix}, \]

where for stages \( i = 1, 2 \),

\[ \hat{S}^{(1,1)}_n(\theta) = \frac{\partial}{\partial \mu_Y} S^1_n(\theta) = \frac{-n_i}{\sigma_Y^2(1 - \rho^2)}; \]

\[ \hat{S}^{(2,2)}_n(\theta) = \frac{\partial}{\partial \rho} S^2_n(\theta); \]

\[ \hat{S}^{(2,1)}_n(\theta) = \hat{S}^{(1,2)}_n(\theta) = \frac{\partial}{\partial \mu_Y} S^1_n(\theta); \]

and

\[ \frac{\partial}{\partial \rho} S^2_n(\theta) = \sum_{j=1}^{n_1} \left\{ 4\rho \Phi^{-1}(u_j)\sigma_Y [y_j - \mu_Y - \rho \Phi^{-1}(u_j)\sigma_Y] \right\} \sigma_Y^2(1 - \rho^2)^2 - \sum_{j=1}^{n_1} \left\{ \Phi^{-1}(u_j) \right\}^2 \sigma_Y^2(1 - \rho^2)^2; \]

\[ \frac{\partial}{\partial \rho} S^1_n(\theta) = \sum_{j=1}^{n_1} \left\{ 4\rho \Phi^{-1}(u_j)\sigma_Y [y_j - \mu_Y - \rho \Phi^{-1}(u_j)\sigma_Y] \right\} \sigma_Y^2(1 - \rho^2)^2 - \sum_{j=1}^{n_1} \left\{ \Phi^{-1}(u_j) \right\}^2 \sigma_Y^2(1 - \rho^2)^2; \]

3.2.2. Asymptotic distribution of \( \sqrt{n_1}(\hat{\theta} - \theta) \). Because the threshold for screening patients in the second stage is random, the MLEs jointly follow a randomly scaled mixture of normal distributions asymptotically (stably) as \( n_2 \to \infty \). A thorough description of stable and mixing convergence can be found in Rényi (1963) and Häusler and Luschgy (2015). A thorough description of stable and mixing convergence can be found in Rényi (1963) and Häusler and Luschgy (2015).

With \( V \) independent of \( Z \) as \( n_2 \to \infty \), where \( Z \) is a 2-dimensional standard normal random
is a random matrix with a.s. finite elements that depend on stage 1 data only through \( \hat{u}^* \).

3.2.3. Marginal asymptotic distributions of \( \hat{\mu}_n \) and \( \hat{\rho}_n \). Based on the limiting joint distribution, the asymptotic marginal distributions of the MLEs are also shown to be randomly scaled mixtures of normal distributions in Lemma 3.2.

**Lemma 3.2.** For the two-stage enrichment design,
\[
\sqrt{n}(\hat{\mu}_n - \mu_Y) \to W_1^{1/2} Z \mathcal{F}_{n_1} \text{ stably},
\]
\[
\sqrt{n}(\hat{\rho}_n - \rho) \to W_2^{1/2} Z \mathcal{F}_{n_1} \text{ stably},
\]
as \( n_2 \to \infty \), where \( Z \) is a standard normal random variable that is independent of
\[
W_1 = \frac{v_{22}}{v_{11}v_{22} - v_{12}^2},
\]
and
\[
W_2 = \frac{v_{11}}{v_{11}v_{22} - v_{12}^2}.
\]

3.2.4. The asymptotic distribution of the Fisher transformation of \( \rho \). When a hypothesis test of \( H_0 : \rho = 0 \) is performed, the Fisher transformation \( 0.5 \ln \left( \frac{1 + \rho}{1 - \rho} \right) \) may be applied to improve performance (e.g., Ferguson, 1996). The delta method yields the following asymptotic distribution in Lemma 3.3.

**Lemma 3.3.** For the two-stage enrichment design,
\[
\sqrt{n} \left( \frac{1}{2} \ln \frac{1 + \hat{\rho}_n}{1 - \hat{\rho}_n} - \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho} \right) \to \frac{1}{1 - \rho^2} W_2^{1/2} Z \mathcal{F}_{n_1} \text{ stably},
\]
as \( n_2 \to \infty \), where \( Z \) is a standard normal random variable that is independent of \( W_2 \).

3.2.5. Random normings. For the proposed enrichment design, it is shown that normalizing \( \hat{\mu}_n \) and \( \hat{\rho}_n \) with a function of the observed information matrix, one obtains asymptotically standard normal limiting distributions (mixing). To derive the test statistics for testing \( H_{01} \) and \( H_{02} \), the observed information matrix, given all data at the end of the study, is used. The observed information matrix evaluated at MLEs is
\[
j_n(\hat{\theta}) = \begin{pmatrix} j_{11}(\hat{\theta}) & j_{12}(\hat{\theta}) \\ j_{21}(\hat{\theta}) & j_{22}(\hat{\theta}) \end{pmatrix} = -\mathbf{S}_n(\hat{\theta}).
\]

We now state the main theorem.
THEOREM 3.4. For the two-stage enrichment design,

\[
\left[ \frac{j_{11}(\hat{\theta})j_{22}(\hat{\theta}) - j_{12}^2(\hat{\theta})}{j_{22}(\hat{\theta})} \right]^{1/2} (\hat{\mu}_n - \mu) \overset{d}{\rightarrow} N(0,1) \quad \text{(mixing)},
\]

\[
\left[ \frac{j_{11}(\hat{\theta})j_{22}(\hat{\theta}) - j_{12}^2(\hat{\theta})}{j_{11}(\hat{\theta})} \right]^{1/2} (\hat{\rho}_n - \rho) \overset{d}{\rightarrow} N(0,1) \quad \text{(mixing)},
\]

\[
(1 - \hat{\rho}_n^2) \left[ \frac{j_{11}(\hat{\theta})j_{22}(\hat{\theta}) - j_{12}^2(\hat{\theta})}{j_{11}(\hat{\theta})} \right]^{1/2} \left( \frac{1}{2} \ln \frac{1}{1 - \hat{\rho}_n} - \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho} \right) \overset{d}{\rightarrow} N(0,1)
\]

as \( n_2 \rightarrow \infty \).

3.3. Hypothesis testing. Using the asymptotic marginal distributions of the MLEs in Theorem 3.4, \( H_{01} \) is rejected at level \( \alpha/2 \), if

\[
(3.1) \quad \left[ \frac{j_{11}(\hat{\theta})j_{22}(\hat{\theta}) - j_{12}^2(\hat{\theta})}{j_{22}(\hat{\theta})} \right]^{1/2} (\hat{\mu}_n - c\sigma_Y - \mu_h) > z_{1-\alpha/2},
\]

where \( z_{1-\alpha/2} \) is the \( 1 - \alpha/2 \)th percentile of the standard normal distribution. We also test

\[
H_{02} : \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho} = 0
\]

at level \( \alpha/2 \). By Theorem 3.4, \( H_{02} \) is rejected if

\[
(3.2) \quad \left[ \frac{j_{11}(\hat{\theta})j_{22}(\hat{\theta}) - j_{12}^2(\hat{\theta})}{j_{11}(\hat{\theta})} \right]^{1/2} (1 - \hat{\rho}_n^2) \left( \frac{1}{2} \ln \frac{1 + \hat{\rho}_n}{1 - \hat{\rho}_n} - 0 \right) > z_{1-\alpha/2}.
\]

Because each hypothesis is tested at level \( \alpha/2 \), the family-wise type I error rate is controlled at level \( \alpha \).

3.4. Treatment effect estimation for the derived target population. When the result of the study is positive, the target population is defined, and the treatment effect for this derived target population is of interest.

3.4.1. The whole population as the target population. When \( H_{01} \) is rejected and \( H_{02} \) is not, the recommended target population is the whole population. Because the biomarker is independent of the outcome \( y \), the possible truncation of the distribution of the biomarker will not affect \( y \) in the second stage. Thus, the treatment effect for the target population \( \delta \) can be estimated by

\[
\hat{\delta} = \bar{y} - \mu_h = \frac{1}{n} \sum_{j=1}^{n} y_j - \mu_h.
\]

When both \( H_{01} \) and \( H_{02} \) are rejected, the recommended target population is again the whole population. In this case, the sample mean cannot be used to estimate the treatment effect for the whole population, since data from the second stage can still be from a truncated distribution, and this will have an effect on the distribution of the outcome. Therefore, it is better to estimate the treatment effect by \( \hat{\mu}_n - \mu_h \).
3.4.2. A subpopulation as the target population. When \(H_02\) is rejected, and \(H_{01}\) is not, a recommendation for the target population threshold would be a solution of the equation

\[
u_f^* = \min \{ u \mid \hat{\lambda}(u) \geq c \}
\]

\[
= \min \left\{ u \left| \frac{\hat{\mu}_n + \hat{\rho}_n \sigma_Y \int_{u}^{+\infty} \Phi^{-1}(u)/(1-u)du - \mu_h}{\text{Var}(Y_U > u)} \geq \frac{c}{\sigma_Y} \right. \right\},
\]

where

\[
\text{Var}(Y_U > u) = \sigma_Y \sqrt{1 - \rho_n^2 \left\{ 1 - \int_{u}^{+\infty} [\Phi^{-1}(u)]^2/(1-u)du + \left[ \int_{u}^{+\infty} \Phi^{-1}(u)/(1-u)du \right]^2 \right\}}.
\]

The treatment effect for this target population \(\delta(u_f^*)\) can be estimated by the following methods.

3.4.2.1. Naive estimation. Given \(u_f^*\), a natural way to estimate \(\delta(u_f^*)\) is the weighted average of treatment effect for patients with biomarker quantile above the threshold from both stages:

\[
\hat{\delta}_n(u_f^*) = \omega_1 \sum_{j=1}^{n} \frac{I(u_j \geq u_f^*) y_j}{I(u_j \geq u_f^*)} + \omega_2 \sum_{j=n+1}^{n} \frac{I(u_j \geq u_f^*) y_j}{I(u_j \geq u_f^*)} - \mu_h,
\]

where \(\omega_1\) and \(\omega_2\) are prespecified weights with the constraint \(\omega_1 + \omega_2 = 1\). The variance estimate is given by

\[
(3.3) \quad \sigma_Y^2 \left[ \sum_{j=1}^{n} \frac{\omega_1^2}{I(u_j \geq u_f^*)} + \sum_{j=n+1}^{n} \frac{\omega_2^2}{I(u_j \geq u_f^*)} \right].
\]

Of course, \(u_f^*\) is unknown and must be estimated, so the distribution of the estimate \(\hat{\delta}_n(u_f^*)\) of the target population’s effect size is a more complicated function of the observations than may seem apparent at first glance.

3.4.2.2. Model-based estimation. From the bivariate normal model, for a fixed threshold \(u_f^*\),

\[
\delta(u_f^*) = \int_{u_f^*}^{1} E(y|u) du - \mu_h = \int_{u_f^*}^{1} \mu_Y + \rho \Phi^{-1}(u) \sigma_Y du - \mu_h = \mu_Y + \rho \sigma_Y c(u_f^*) - \mu_h,
\]

which can be estimated by

\[
\hat{\delta}_m(u_f^*) = \hat{\mu}_n + \hat{\rho}_n \sigma_Y c(u_f^*),
\]

where

\[
c(u_f^*) = \int_{u_f^*}^{1} \Phi^{-1}(u) du.
\]

Since this estimator is based on all data from both stages, it is more efficient than the naive approach, which is only based on a subset of patients. The estimator \(\hat{\delta}_m(u_f^*)\) also can be shown to follow a randomly scaled mixture of normal distributions asymptotically as \(n_2 \to \infty\) [Equation (A.4) in the Appendix]. Theorem 3.5 shows that using a function of the observed information matrix to normalize \(\hat{\delta}_m(u_f^*)\) provides a standard normal limit, from which a confidence interval can be derived.
Theorem 3.5. For the two-stage enrichment design,

\[ T^{1/2}(\hat{\pi}_m(u_f^*) - \pi_m(u_f^*)) \overset{d}{\to} N(0, 1) \quad \text{(mixing)} \]

as \( n_2 \to \infty \), where

\[ T = \frac{|j_n(\hat{\theta})|}{[j_{11}(\hat{\theta})]^2 + 2|j_1(\hat{\theta})|} \]

and \( |j_n(\hat{\theta})| \) is the determinant of the observed information matrix evaluated at MLEs.

4. Simulation. In this section, the proposed enrichment design is evaluated under several possible scenarios by simulation studies, and it is compared with a single stage design with a guessed threshold.

4.1. Data generation. Stage 1 data are simulated assuming the biomarker \( x \) and the outcome \( y \) follow a bivariate normal distribution with the mean vector

\[
\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix},
\]

and the variance covariance matrix

\[
\begin{pmatrix} \sigma_X^2 & \rho \sigma_Y \sigma_X \\ \rho \sigma_Y \sigma_X & \sigma_Y^2 \end{pmatrix}.
\]

The parameters \( \mu_X = 2 \), \( \sigma_X = \sigma_Y = 1 \), and some design settings are also kept constant, with \( \alpha = 0.025 \), \( \mu_h = 0.5 \), \( c = 0.5 \), \( n = 500 \). For the various simulation scenarios, \( n_i \), the sample size in stage \( i \), the correlation \( \rho \), and mean \( \mu_Y \) are varied.

The biomarker threshold quantile is estimated by MLEs based on the first stage data. Because we consider the first stage sample size to be small, an upper bound on the estimate \( \pi^* \) is set at 0.9 to avoid obtaining too conservative a subpopulation in stage 2. Data in stage 2 are simulated from a bivariate normal density truncated at the derived threshold. This process is repeated 10,000 times for each scenario.

4.2. Type I error rate and power of hypothesis tests. For the two-stage adaptive design, \( \alpha \) levels are adjusted using Bonferroni’s adjustment for individual tests of \( \rho \) and \( \mu_Y \) as described in Section 3.3. Hypothesis testing scenario 1 (in Section 3.1) is rejected at level \( \alpha = 0.025 \) if at least one of \( H_{01} \) and \( H_{02} \) is rejected at level \( \alpha = 0.0125 \).

For comparison, consider a single stage design with a guessed threshold \( u_f^* = 0.5 \). Standard test statistics can be used to test scenario 1 against other scenarios. Under scenario 1, both \( H_{01} \) and \( H_{02} \) are true, that is, biomarkers and outcomes are uncorrelated and the population treatment effect is not significant; so the distribution of the outcome \( y \) is not changed by using the biomarker for patient screening. Thus,

\[
\frac{\bar{y} - c \sigma_Y - \mu_h}{\sigma_Y/\sqrt{n}} \sim N(0, 1),
\]

and scenario 1 is rejected at level \( \alpha = 0.025 \) if

\[
\frac{\bar{y} - c \sigma_Y - \mu_h}{\sigma_Y/\sqrt{n}} > 1.96.
\]

Table 1 shows, for various \( \rho \) and \( \mu_Y \), the power of detecting a target population, and the power of subhypotheses \( H_{01} \) and \( H_{02} \). Column 3 is the true effect size calculated using the whole population mean. True thresholds for different combinations of \( \rho \) and \( \sigma_Y \) are given in...
column 4. When both null hypotheses \( H_{01} \) and \( H_{02} \) are true (\( \rho = 0, \mu_Y = 1.0 \)), the type I error rate of rejecting scenario 1 in the single stage design is controlled, while Bonferroni’s adjustment is applied to separate tests of \( \rho \) and \( \mu_Y \), the type I error rate in the two-stage design is slightly inflated. When \( \rho = 0 \), the power of the single stage design in rejecting scenario 1 is greater than that of the two-stage design. However, when \( \rho > 0 \), the power of the two-stage design has much greater power than the single stage design. For example, for \( \rho = 0.15, \mu_Y = 0.85 \), the power of two stage design is 0.822 as compared with a power of 0.063 for the single stage design. In addition, for the single stage design, the power is greater if the true threshold \( u^* < u^*_0 \) as compared with the power if \( u^* > u^*_0 \) given the same \( \rho \). Thus, for the single stage design, the power depends on the choice of the fixed threshold \( u^*_0 \).

We repeated the simulations with unknown variance \( \sigma^2 \), using the estimated variance. The power results were identical, to within one percent (data not shown).

In this paper, asymptotic results were obtained by taking \( n_2 \to \infty \). In this simulation, the total sample size is fixed at 500 to compare the power of a single stage study (\( \omega_1 = 1.0 \)) to an enrichment study having a relatively small first stage (\( \omega_1 = 0.2 \)). The power of these designs could also be compared under local alternatives, where the magnitude of the treatment effect decreases with the total sample size at rate \( 1/\sqrt{n} \). Local alternatives are common practice where, given a desired power, large sample sizes are used to detect small treatment effects and smaller sample sizes are used to detect large effects. Additional comments regarding local asymptotics are made in Section 6.

**Table 1**

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( \mu_Y )</th>
<th>Effect size</th>
<th>True Threshold</th>
<th>Power (( \omega_1 = 1.0 ))</th>
<th>( H_{01} \cup H_{02} )</th>
<th>( H_{A1} )</th>
<th>( H_{A2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.50</td>
<td>NA</td>
<td>0.022</td>
<td>0.030</td>
<td>0.011</td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>0.60</td>
<td>NA</td>
<td>0.608</td>
<td>0.470</td>
<td>0.458</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.65</td>
<td>NA</td>
<td>0.921</td>
<td>0.837</td>
<td>0.831</td>
<td>0.014</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.35</td>
<td>0.83</td>
<td>&lt; 0.001</td>
<td>0.467</td>
<td>&lt; 0.001</td>
<td>0.467</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.40</td>
<td>0.61</td>
<td>0.008</td>
<td>0.483</td>
<td>&lt; 0.001</td>
<td>0.483</td>
<td></td>
</tr>
<tr>
<td>0.95</td>
<td>0.45</td>
<td>0.29</td>
<td>0.095</td>
<td>0.503</td>
<td>&lt; 0.001</td>
<td>0.503</td>
<td></td>
</tr>
<tr>
<td>0.08</td>
<td>0.30</td>
<td>0.76</td>
<td>&lt; 0.001</td>
<td>0.794</td>
<td>&lt; 0.001</td>
<td>0.794</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.35</td>
<td>0.60</td>
<td>0.004</td>
<td>0.807</td>
<td>&lt; 0.001</td>
<td>0.807</td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.40</td>
<td>0.40</td>
<td>0.063</td>
<td>0.822</td>
<td>&lt; 0.001</td>
<td>0.822</td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td>0.20</td>
<td>0.82</td>
<td>&lt; 0.001</td>
<td>0.962</td>
<td>&lt; 0.001</td>
<td>0.962</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.80</td>
<td>0.60</td>
<td>0.002</td>
<td>0.961</td>
<td>&lt; 0.001</td>
<td>0.961</td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.40</td>
<td>0.29</td>
<td>0.260</td>
<td>0.968</td>
<td>&lt; 0.001</td>
<td>0.968</td>
<td></td>
</tr>
</tbody>
</table>

**4.3. Threshold estimation.** In Figure 3, the bias in estimating \( u^* \) as a function of \( \rho \) is given under \( n_1 = 100, \mu_Y = 0.5 \). As expected, the estimated threshold with the single stage fixed \( X \) model is too conservative, so the bias is positive for all \( \rho \). For a given level of clinical significance \( c \), the bias becomes greater as \( \rho \) increases. However, under the same \( \rho \), the bias is not only affected by the clinical significance \( c \). For example, the curve for the single stage with \( c = 0.8 \) is below that for the single stage with \( c = 0.5 \) at the beginning, then becomes above the curve with \( c = 0.5 \).

The bias based on the bivariate model in the two-stage design is very close to 0 for different \( \rho \) and \( c \).
4.4. Treatment effect estimation for the derived target population. Table 2 presents the characteristics of the treatment effect estimator for the derived target population: empirical bias, root mean squared error (RMSE) and 95% coverage probabilities for estimating $\delta(u_1^*)$ with the naive and model-based approaches under $n_1 = 100$, $n = 500$, $\mu_Y = 0.4$, $\rho = 0.5$ and $\omega_i = n_i/n$, $i = 1, 2$. For each estimator, the empirical bias is the empirical mean of $\hat{\delta}(u_1^*) - \delta(u_1^*)$, and the RMSE is the square root of the average $[\hat{\delta}(u_1^*) - \delta(u_1^*)]^2$. The 95% coverage probability of the naive estimator is calculated based on the variance given by (3.3), and the coverage probability of the model-based estimator is obtained from its distribution in Theorem 3.5.

The bias tends to be larger for smaller $n_1$ using the naive approach, while it is less than 0.1% for all $n_1$ using the model-based approach. The RMSE of the naive estimator decreases from 0.116 to 0.059 as $n_1/n$ increases from 0.10 to 0.50; then the RMSE increases again as fewer outcomes with biomarkers above $u_1^*$ can be used to estimate the treatment effect. For the model-based estimator, there are more outcome values at biomarkers above the final threshold when $n_1$ is small and $n_2$ is large, and thus the RMSE increases from 0.040 to 0.044 as $n_1/n$ increases from 0.10 to 0.60. The coverage probabilities of the model-based estimator are close to 95% across all scenarios. Furthermore, the coverage probabilities of the naive estimator are not stable. For example, it is 71.3 as $n_1 = 50$, and it increases to 97.1 as $n_1 = 300$.

4.5. Model Misspecification. The robustness of the proposed design is evaluated when the true model is not the bivariate normal distribution. Assume

\[
\begin{pmatrix}
Y \\
X
\end{pmatrix} = \begin{pmatrix}
\mu_Y + Z_1 \\
\mu_X + Z_2
\end{pmatrix},
\]  
(4.2)
where \((Z_1, Z_2)\) follows a bivariate skew-normal distribution.

\[
f(z_1, z_2) = 2\phi(z_1, z_2; \Omega)\Phi(\alpha_1 z_1 + \alpha_2 z_2),
\]

where \(\phi(z_1, z_2; \Omega)\) is a bivariate normal density with zero mean and correlation matrix \(\Omega\), \(\Phi(\cdot)\) is the univariate \(N(0, 1)\) distribution function, and \((\alpha_1, \alpha_2)\) regulates departure from symmetry. In particular, when \((\alpha_1, \alpha_2) = (0, 0)\), \((Y, X)\) follows a bivariate normal distribution.

Table 3 gives the power of subhypotheses \(H_{01}\) and \(H_{02}\) under the bivariate skew-normal distribution. In the first part of the table, the type I error rate is evaluated under the misspecified model. When there is skewness on \(Y\), the type I error rate in the two-stage design is inflated. For example, it is 0.100 as \((\alpha_1, \alpha_2) = (0.05, 0)\). When the skewness is on \(X\) only, the type I error rate is around 0.025. Therefore, the distribution of the outcome \(Y\) is sensitive to the normality, while the distribution of the biomaker \(X\) is robust to the departure of the normality. In the second part of the table, the power of the proposed two-stage design is examined. Because of the sensitivity of the distribution of \(Y\) to normality, only effects of the skewness of \(X\) on the power is shown. The skewness on \(X\) reduces the power of the proposed two-stage design from 0.807 to 0.549 as \(\alpha_2\) increases from 0 to 3.

### Table 3

**Power of Rejecting Scenario 1 in Favor of a Existence of a Clinically Significant Subpopulation. With \(\omega_1 = 0.2\), tests with alternatives \(H_{A1}\) use (3.1) and tests with alternatives \(H_{A2}\) use (3.2) under misspecified Model (4.2).**

<table>
<thead>
<tr>
<th>(\rho)</th>
<th>(\mu_Y)</th>
<th>Effect size</th>
<th>Threshold</th>
<th>((\alpha_1, \alpha_2))</th>
<th>(H_{A1} \cup H_{A2})</th>
<th>(H_{A1})</th>
<th>(H_{A2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.00</td>
<td>0.50</td>
<td>NA</td>
<td>((0, -1))</td>
<td>0.033</td>
<td>0.015</td>
<td>0.021</td>
</tr>
<tr>
<td>0</td>
<td>1.00</td>
<td>0.50</td>
<td>NA</td>
<td>((0, -0.5))</td>
<td>0.033</td>
<td>0.013</td>
<td>0.020</td>
</tr>
<tr>
<td>0</td>
<td>1.00</td>
<td>0.50</td>
<td>NA</td>
<td>(0, 0)</td>
<td>0.030</td>
<td>0.011</td>
<td>0.018</td>
</tr>
<tr>
<td>0</td>
<td>1.00</td>
<td>0.50</td>
<td>NA</td>
<td>((0, 0.5))</td>
<td>0.026</td>
<td>0.006</td>
<td>0.020</td>
</tr>
<tr>
<td>0</td>
<td>1.00</td>
<td>0.50</td>
<td>NA</td>
<td>((0, 1))</td>
<td>0.024</td>
<td>0.001</td>
<td>0.023</td>
</tr>
<tr>
<td>0</td>
<td>1.00</td>
<td>0.50</td>
<td>NA</td>
<td>((0, 0.3))</td>
<td>0.024</td>
<td>&lt; 0.001</td>
<td>0.024</td>
</tr>
<tr>
<td>0</td>
<td>1.00</td>
<td>0.50</td>
<td>NA</td>
<td>((0, 0.05, 0))</td>
<td>0.100</td>
<td>0.080</td>
<td>0.022</td>
</tr>
<tr>
<td>0</td>
<td>1.00</td>
<td>0.50</td>
<td>NA</td>
<td>(0.1, -0.1)</td>
<td>0.336</td>
<td>0.316</td>
<td>0.027</td>
</tr>
<tr>
<td>0</td>
<td>1.00</td>
<td>0.50</td>
<td>NA</td>
<td>((0, 1))</td>
<td>0.317</td>
<td>0.301</td>
<td>0.018</td>
</tr>
</tbody>
</table>

| 0.15 | 0.85 | 0.35 | 0.60 | \((0, -1)\) | 0.753 | < 0.001 | 0.753 |
| 0.15 | 0.85 | 0.35 | 0.60 | \((0, -0.5)\) | 0.804 | < 0.001 | 0.804 |
| 0.15 | 0.85 | 0.35 | 0.60 | \((0, 0)\) | 0.807 | < 0.001 | 0.807 |
| 0.15 | 0.85 | 0.35 | 0.60 | \((0, 0.5)\) | 0.750 | < 0.001 | 0.750 |
| 0.15 | 0.85 | 0.35 | 0.60 | \((0, 1)\) | 0.671 | < 0.001 | 0.671 |
| 0.15 | 0.85 | 0.35 | 0.60 | \((0, 0.3)\) | 0.549 | < 0.001 | 0.549 |
5. Extension to a parallel group design. When patients having matching biomarker values are randomized into two groups, the proposed enrichment design can be easily extended to include a control arm by letting the outcome $y$ be the difference of the outcome between treatment and control groups with the same biomarker values.

In addition, there are (at least) two ways to generalize the proposal to a randomized parallel group design: one is to include an indicator variable of the treatment in the model, and the other is to model two arms separately. These two methods are compared and discussed by Holgersson, Nordström and Öner (2014); Schepers (2016). Because the proposed design is based on the joint distribution of outcomes and biomarkers, modelling by arms is recommended.

Let $X$ and $Y_l$ follow a bivariate normal distribution with correlation $\rho_l$, mean vector $(\mu_l, \mu_X)$, and variance-covariance matrix

$$
\begin{pmatrix}
\sigma_Y^2 & \rho_l \sigma_Y \sigma_X \\
\rho_l \sigma_Y \sigma_X & \sigma_X^2
\end{pmatrix},
$$

where $l = 0$ for the control arm and 1 for the treatment arm; and for illustration purpose, assume $\rho_1 \geq \rho_0 \geq 0$. A prognostic biomarker is one in which the likelihood of the patient outcome, independent of the treatment received, is provided. A predictive biomarker is one in which the treatment effect, as compared with the control, is indicated. If the biomarker is predictive, the variance of the outcome in the subpopulation is larger in the treatment group. At the interim analysis, the predictability of the biomarker is unknown, so the treatment effect used to define the threshold needs to be standardized appropriately. In this case, the threshold can be determined by the difference of the mean outcomes between the two groups in the subpopulation standardized by the square root of the sum of individual variances.

There are four scenarios depending on the effect of a biomarker, which may be prognostic and/or predictive.

1. When $\rho_1 = \rho_0 = 0$, the biomarker and the outcome are unrelated, so the biomarker is neither prognostic nor predictive.
2. When $\rho_1 = \rho_0 > 0$, the biomarker is prognostic, but not predictive. A higher biomarker value implies a better outcome regardless of the treatment received. Therefore, no clinically significant subpopulation can be defined based on biomarker values.
3. When $\rho_1 > \rho_0 = 0$, the biomarker is predictive, but not prognostic. A subpopulation with desired treatment effect can be targeted if the treatment effect for the whole population is not significant.
4. When $\rho_1 > \rho_0 > 0$, the biomarker is both predictive and prognostic. A subpopulation still can be obtained if the treatment effect for the whole population is not significant.

Note, the implications of scenarios 3 and 4 are the same. A useful subpopulation can be identified in scenarios 3 and 4, but not in scenarios 1 and 2.

Controlling the family-wise type I error rate also needs to be considered for such parallel group designs.

6. Conclusions. In this paper, a two-stage enrichment design is proposed based on a bivariate model. The proposed design increases the efficiency of estimators and the power of hypothesis test, and it protects patients from exposure to a futile treatment when compared with a single stage design in most cases studied. In addition, test statistics with normal limits for two important subhypotheses are derived using observed information measures. It is more powerful to test two subhypotheses than using statistics from a single stage design in detecting the existence of the benefiting subpopulation when the outcome and the biomarker are related.
In addition, the proposed design determines a target population with prespecified clinical significance for the future use at the end of the study. By testing two subhypotheses simultaneously under the bivariate model, a more accurate threshold estimate for the target population is obtained than using a single stage design. Further, for the derived target population, point and interval treatment effect estimates are given. Figure 2 illustrates that the choice of \( n_1 \) is very important. Issues associated with the choice of \( n_1 \) demand more attention in all enrichment designs.

Our design is predicated on the assumption that there is only one related biomarker, and both the outcome and the biomarker are normally distributed. The advantage of this model is that it provides closed-form test statistics and estimators. In this context, we have described a set of hypothesis tests, derived test statistics, and their asymptotic distributions. Additionally, the robustness of the bivariate normal model is evaluated, and the distribution of the biomarker is largely robust to departures from normality. Proposals for extending our design to parallel group experiments considering prognostic and predictive biomarkers were given.

This research should be thought of as a first step in developing bivariate, model-based enrichment designs. Further research will be done when continuous outcomes have other bivariate distributions, and when outcomes are binary, censored, or when there are multiple related biomarkers. Extensions of interest, including parallel group and other more complex designs, may benefit from sparse linear programming (e.g. Rosenblum, Fang and Liu, 2020).

Other important extensions include altering the way in which the threshold is defined. For example, in the parallel group situation, the threshold might be the minimum biomarker for which a significant treatment effect is detected in stage 1 data. Performing hypothesis tests against local alternatives are required to compare limiting distributions from stage 1 and 2 test statistics using different designs with a prescribed power (less than one). To test against local alternatives at the end of the study (after stage 1 dependent screening) requires a careful retracing of the limit theory in Lane and Flournoy (2012), Lin, Flournoy and Rosenberger (2020) and here, in comparison with standard local asymptotics [see van der Vaart (2000)].

Similar results to those obtained here might be obtained with effect sizes going to zero at rate \( 1/\sqrt{n_2} \). But if \( n_1 \) and \( n_2 \) both tend to infinity with \( n_1/n_2 \) converging to a constant, limiting mixture distributions likely cannot be simplified by random norming, as with Tarima and Flournoy (2019).

APPENDIX

A.1. Conditional distributions for stage 2 data. In the second stage, \( n_2 \) patients with \( u \geq \hat{u}^* \) are enrolled. So outcomes and biomarker quantiles follow the truncated distribution

\[
f(u, y|\hat{u}^*) = \begin{cases} 
\frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \exp \left\{ -\frac{[y-\mu_Y-\rho\Phi^{-1}(u)\sigma_Y]^2}{2\sigma_Y^2(1-\rho^2)} \right\} \left( \frac{1}{1-\hat{u}^*} \right), & -\infty < y < \infty, \hat{u}^* \leq u \leq 1, \\
0, & \text{otherwise.}
\end{cases}
\]

The conditional probability of biomarker quantiles is

\[
f(u|\hat{u}^*) = \begin{cases} 
\frac{1}{1-\hat{u}^*}, & \hat{u}^* \leq u \leq 1, \\
0, & \text{otherwise.}
\end{cases}
\]

So conditional probability of \( y \) given \( u \) and \( \hat{u}^* \) is

\[
f(y|\hat{u}^*, u) = \frac{f(y, u|\hat{u}^*)}{f(u|\hat{u}^*)} = \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \exp \left\{ -\frac{[y-\mu_Y-\rho\Phi^{-1}(u)\sigma_Y]^2}{2\sigma_Y^2(1-\rho^2)} \right\}, -\infty < y < \infty, \hat{u}^* \leq u \leq 1,
\]

and \( f(y|\hat{u}^*, u) = 0 \) otherwise.
A.2. Proof of Lemma 3.1. By classical large sample theory (Cramér, 1946; Lehmann, 1999; Ferguson, 1996),
\[
\sqrt{n}(\hat{\theta} - \theta) \approx [-\hat{S}_n(\theta)/n]^{-1}(1/\sqrt{n})S_n(\theta)
\]
(A.1)
where
\[
D_i(\theta) = S_i(\theta) - S_{i-1}(\theta)
\]
is the ith subject-wise increment. The numerator of the right hand side of (A.2) is a zero-mean, square integrable, martingale difference array with respect to the \(\sigma\)-field \(\mathcal{F}_n\) (as defined in Hall and Heyde (2014)). To show \(S_{n\cdot}(\theta)/\sqrt{n} \Rightarrow V \mathcal{F}_{n\cdot}\)-stably as \(n_2 \rightarrow \infty\), multivariate version of conditions 3.18-3.20 (Hall and Heyde, 2014, Theorem 3.2), need to be met.

In this enrichment design, condition 3.19 implies
\[
\frac{1}{n} \sum_{j=n_1+1}^{n} D_j(\theta)D_j^T(\theta) = \frac{n_2}{n} \left( \begin{array}{cc} d_{11} & d_{12} \\ d_{21} & d_{22} \end{array} \right) \Rightarrow V,
\]
where
\[
d_{11} = \frac{1}{n_2} \sum_{j=n_1+1}^{n} \frac{[y_j - \mu_Y - \rho \Phi^{-1}(u_j)\sigma_Y]^2}{\sigma_Y^4(1 - \rho^2)^2};
\]
\[
d_{12} = d_{21} = \frac{1}{n_2} \sum_{j=n_1+1}^{n} \frac{[y_j - \mu_Y - \rho \Phi^{-1}(u_j)\sigma_Y]^2\phi^{-1}(u_j)}{\sigma_Y^3(1 - \rho^2)^2}
\]
\[- 1 \frac{1}{n_2} \sum_{j=n_1+1}^{n} \frac{[y_j - \mu_Y - \rho \Phi^{-1}(u_j)\sigma_Y]^3}{\sigma_Y^4(1 - \rho^2)^3}\rho + 1 \frac{1}{n_2} \sum_{j=n_1+1}^{n} \frac{[y_j - \mu_Y - \rho \Phi^{-1}(u_j)\sigma_Y]}{\sigma_Y^3(1 - \rho^2)^2}\rho + \frac{\rho}{1 - \rho^2}\right)^2.
\]

To verify this condition, consider the limit of each \(d_{ij}\) separately, conditional on \(\hat{u}^*\); in this case second stage observations are independent and hence each \(d_{ij}\) converges to its conditional expectation (Rao, 2009, Theorem 7). This result is used repeatedly in obtaining their limits as follows.

\[
d_{11} \Rightarrow E \left\{ \frac{[y_j - \mu_Y - \rho \Phi^{-1}(u_j)\sigma_Y]^2}{\sigma_Y^4(1 - \rho^2)^2} \right| \hat{u}^* \right\}
\]
\[
= \frac{1}{\sigma_Y^4(1 - \rho^2)^2} E \left[ E \left\{ [y_j - \mu_Y - \rho \Phi^{-1}(u_j)\sigma_Y]^2 \right| u_j, \hat{u}^* \right] \right]\]
\[
= \frac{1}{\sigma_Y^4(1 - \rho^2)^2} \int_{-\infty}^{+\infty} [y_j - \mu_Y - \rho \Phi^{-1}(u_j)\sigma_Y]^2 f(y|u_j, \hat{u}^*) dy \left| \hat{u}^* \right]
\]
\[
= \frac{1}{\sigma_Y^4(1 - \rho^2)^2} E \left[ I(u_j > \hat{u}^*) \int_{-\infty}^{+\infty} [y_j - \mu_Y - \rho \Phi^{-1}(u_j)\sigma_Y]^2 f(y|u_j) dy \right| \hat{u}^* \right]
\]
\[ = \frac{1}{\sigma^2 (1-\rho^2)^2} E \left[ I(u_j > \tilde{u}^*) \mid \rho \right] V \left( y / y_j \right) dy / \tilde{u}^* = v_{11}; \]

\[
d_{12} \xrightarrow{p} E \left\{ \frac{[y_j - \mu_Y - \rho \Phi^{-1}(u_j) \sigma_Y \rho \Phi^{-1}(u_j)]^2}{\sigma^2 (1-\rho^2)^2} \mid \tilde{u}^* \right\} - E \left\{ \frac{[y_j - \mu_Y - \rho \Phi^{-1}(u_j) \sigma_Y \rho \Phi^{-1}(u_j)]^2}{\sigma^2 (1-\rho^2)^2} \mid \tilde{u}^* \right\} \\
+ E \left\{ \frac{[y_j - \mu_Y - \rho \Phi^{-1}(u_j) \sigma_Y \rho \Phi^{-1}(u_j)]^2}{\sigma^2 (1-\rho^2)^2} \mid \tilde{u}^* \right\} \\
= \frac{1}{\sigma^2 (1-\rho^2)^2} E \left\{ [y_j - \mu_Y - \rho \Phi^{-1}(u_j) \sigma_Y \rho \Phi^{-1}(u_j)]^2 \mid \tilde{u}^* \right\} \\
= \frac{1}{\sigma^2 (1-\rho^2)^2} E \left\{ I(u_j > \tilde{u}^*) \frac{\sigma^2 Y}{(1-\rho^2)} \Phi^{-1}(u_j) \mid \tilde{u}^* \right\} = v_{12}; \]

\[
d_{22} = \sum_{j=n+1}^{n} \frac{\rho y_j - \rho \Phi^{-1}(u_j) \sigma_Y \rho \Phi^{-1}(u_j)}{\sigma^2 (1-\rho^2)^2} + \sum_{j=n+1}^{n} \frac{2[y_j - \mu_Y - \rho \Phi^{-1}(u_j) \sigma_Y \rho \Phi^{-1}(u_j)]}{\sigma^2 (1-\rho^2)^2} \\
+ \sum_{j=n+1}^{n} \frac{2[y_j - \mu_Y - \rho \Phi^{-1}(u_j) \sigma_Y \rho \Phi^{-1}(u_j)]}{\sigma^2 (1-\rho^2)^2} \\
+ \sum_{j=n+1}^{n} \frac{2[y_j - \mu_Y - \rho \Phi^{-1}(u_j) \sigma_Y \rho \Phi^{-1}(u_j)]}{\sigma^2 (1-\rho^2)^2} \\
\xrightarrow{p} \frac{1}{1-\rho^2} E \left\{ I(u_j > \tilde{u}^*) \frac{\sigma^2 Y}{(1-\rho^2)} \Phi^{-1}(u_j) \mid \tilde{u}^* \right\} + \frac{3\rho^2}{(1-\rho^2)^2} + \frac{2\rho^2}{(1-\rho^2)^2} = v_{22}. \]

Therefore, as \( n_2 \to n \),

\[
\frac{1}{n n_2} \sum_{j=n+1}^{n} D_j(\theta) D_j^T(\theta) \xrightarrow{p} V, \]

and condition 3.19 is obtained.

Conditions 3.18 and 3.20 are easily verified, so by Theorem 3.2 [Hall and Heyde (2014)],
the numerator of (A.2) converges as

\[
\frac{S_{n_2}(\theta)}{\sqrt{n}} \xrightarrow{\text{stably}} \mathbf{V}^{1/2} \mathbf{Z} \mathbf{F}_{n_1}, \]

as \( n_2 \to \infty \). The denominator of equation (A.2) is

\[
-\frac{S_{n_2}(\theta)}{n} = \left( \frac{-\partial}{\partial \mu_Y} S^1_{n_2}(\theta) / n - \frac{-\partial}{\partial \rho} S^1_{n_2}(\theta) / n \right) \\
- \frac{-\partial}{\partial \rho} S^2_{n_2}(\theta) / n - \frac{-\partial}{\partial \rho} S^2_{n_2}(\theta) / n \]

and as \( n_2 \to \infty \), conditional on \( \tilde{u}^* \),

\[
-\frac{-\partial}{\partial \mu_Y} S^1_{n_2}(\theta) / n = \frac{n n_2}{n_2 \sigma^2 (1-\rho^2)^2} \xrightarrow{p} v_{11}; \]

\[
-\frac{-\partial}{\partial \rho} S^1_{n_2}(\theta) / n = \frac{n n_2}{n_2 \sigma^2 (1-\rho^2)^2} \xrightarrow{p} v_{12}; \]

\[
-\frac{-\partial}{\partial \rho} S^2_{n_2}(\theta) / n = \frac{n n_2}{n_2 \sigma^2 (1-\rho^2)^2} \xrightarrow{p} v_{12}; \]

\[
-\frac{-\partial}{\partial \rho} S^2_{n_2}(\theta) / n = \frac{n n_2}{n_2 \sigma^2 (1-\rho^2)^2} \xrightarrow{p} v_{12}; \]

\[
\frac{\sum_{j=n+1}^{n} \{ 4 \rho \Phi^{-1}(u_j) \sigma_Y \} \left( y_j - \mu_Y - \rho \Phi^{-1}(u_j) \sigma_Y \right)^2}{\sigma^2 (1-\rho^2)^2} + \frac{\sum_{j=n+1}^{n} \{ \Phi^{-1}(u_j) \}^2}{\sigma^2 (1-\rho^2)^2} \]

\[
+ \frac{\sum_{j=n+1}^{n} \{ (1+3\rho^2) \} \left( y_j - \mu_Y - \rho \Phi^{-1}(u_j) \sigma_Y \right)^2}{\sigma^2 (1-\rho^2)^2} - \frac{n n_2 (1+\rho^2) \sigma_Y}{n \sigma^2 (1-\rho^2)^2} \]

\[
\frac{n n_2 (1+\rho^2) \sigma_Y}{n \sigma^2 (1-\rho^2)^2} \]

\[
\frac{n n_2 (1+\rho^2) \sigma_Y}{n \sigma^2 (1-\rho^2)^2} \]

\[
\frac{n n_2 (1+\rho^2) \sigma_Y}{n \sigma^2 (1-\rho^2)^2} \]

\[
\frac{n n_2 (1+\rho^2) \sigma_Y}{n \sigma^2 (1-\rho^2)^2} \]

\[
\frac{n n_2 (1+\rho^2) \sigma_Y}{n \sigma^2 (1-\rho^2)^2} \]

\[
\frac{n n_2 (1+\rho^2) \sigma_Y}{n \sigma^2 (1-\rho^2)^2} \]

\[
\frac{n n_2 (1+\rho^2) \sigma_Y}{n \sigma^2 (1-\rho^2)^2} \]
\[
\frac{\Phi^{-1}(u)\sqrt{1 - \rho^2}}{1 - \rho^2} + \frac{1 + 3\rho^2}{(1 - \rho^2)^2} - \frac{1 + \rho^2}{(1 - \rho^2)^2} = v_{22}.
\]

In summary, \(-S_{n_2}(\theta)/n \overset{p}{\to} V\).

From the generalized Cramér-Slutsky theorem (Aldous and Eagleson, 1978),
\[
\left[\frac{-S_{n_2}(\theta)}{n}\right]^{-1} \frac{1}{\sqrt{n}}S_{n_2}(\theta) \to V^{-1/2}Z \mathcal{F}_{n_1} \text{ stably}
\]
as \(n_2 \to \infty\), and because \(\sqrt{n}(\hat{\theta}_n - \theta)\) and equation (A.2) are asymptotically equivalent,
\[
\sqrt{n}(\hat{\theta} - \theta) \to V^{-1/2}Z \mathcal{F}_{n_1} \text{ stably}
\]
as \(n_2 \to \infty\).

**A.3. Proof of Lemma 3.2.** By Lemma 3.1,
\[
\sqrt{n}(\hat{\theta} - \theta) \to V^{-1/2}Z \mathcal{F}_{n_1} \text{ stably}
\]
as \(n_2 \to \infty\). Using Cramer-Wold device,
\[
\sqrt{n}(\hat{\mu}_n - \mu) = \sqrt{n}C'_1(\hat{\theta} - \theta) \to C'_1V^{-1/2}Z \mathcal{F}_{n_1} \text{ stably},
\]
and
\[
\sqrt{n}(\hat{\rho}_n - \rho) = \sqrt{n}C'_2(\hat{\theta} - \theta) \to C'_2V^{-1/2}Z \mathcal{F}_{n_1} \text{ stably},
\]
as \(n_2 \to \infty\), where \(C_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \) and \(C_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \). Define the elements of the random matrix \(V^{-1/2}\) by
\[
(V^{-1/2}) = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix}
\]
and let
\[
Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix},
\]
where \(Z_1\) and \(Z_2\) are standard normal variables. Then
\[
C'_1V^{-1/2}Z = q_{11}Z_1 + q_{12}Z_2 \text{ and } C'_2V^{-1/2}Z = q_{21}Z_1 + q_{22}Z_2.
\]
The characteristic function of \(q_{11}Z_1 + q_{12}Z_2\) is
\[
E[e^{it(q_{11}Z_1 + q_{12}Z_2)}] = E\{E[e^{itq_{11}Z_1 + q_{12}Z_2} | \hat{u}^*]E[e^{itq_{12}Z_2} | \hat{u}^*]\}
\]
\[
= E\{e^{-q_{11}^2t^2/2 - q_{12}^2t^2/2}\} = E\{e^{-q_{11}^2t^2/2 - q_{12}^2t^2/2}\},
\]
which is also the characteristic function of \(\sqrt{q_{11}^2 + q_{12}^2}Z\), where \(Z\) is a standard normal variable. Therefore, \(q_{11}Z_1 + q_{12}Z_2\) and \(\sqrt{q_{11}^2 + q_{12}^2}Z\) are equivalent in distribution, and as \(n_2 \to \infty\),
\[
\sqrt{n}(\hat{\mu}_n - \mu) \to \sqrt{q_{11}^2 + q_{12}^2}Z \mathcal{F}_{n_1} \text{ stably and}
\]
\[
\sqrt{n}(\hat{\rho}_n - \rho) \to \sqrt{q_{21}^2 + q_{22}^2}Z \mathcal{F}_{n_1} \text{ stably}.
\]
It is well known that
\[
V^{1/2} = \frac{1}{\sqrt{q_{11} + q_{12}^2 + 2\sqrt{|V|}}} \begin{pmatrix} v_{11} + \sqrt{|V|} & v_{12} \\ v_{12} & v_{22} + \sqrt{|V|} \end{pmatrix},
\]
So
\[
V^{-1/2} = \frac{\sqrt{v_{11} + v_{22} + 2\sqrt{V}}}{(v_{11} + \sqrt{V})(v_{22} + \sqrt{V}) - v_{12}^2} \begin{pmatrix} v_{22} + \sqrt{V} & -v_{12} \\ -v_{12} & v_{11} + \sqrt{V} \end{pmatrix}.
\]

With some algebra, it can be shown that
\[
q_{11}^2 + q_{12}^2 = \frac{v_{22}}{v_{11}v_{22} - v_{12}^2},
\]
and
\[
q_{21}^2 + q_{22}^2 = \frac{v_{11}}{v_{11}v_{22} - v_{12}^2}.
\]

Thus,
\[
\sqrt{n}(\hat{\mu}_n - \mu) \rightarrow W_1^{1/2} Z F_{n_1} \text{ stably},
\]
\[
\sqrt{n}(\hat{\rho}_n - \rho) \rightarrow W_2^{1/2} Z F_{n_2} \text{ stably},
\]
as \(n_2 \rightarrow \infty\), where \(Z\) is independent of \(W_1\) and \(W_2\), respectively, as defined in Lemma 3.2.

A.4. Proof of Theorem 3.4. As \(n \rightarrow \infty\) and \(n_2/n \rightarrow 1\),
\[
\frac{j_{11}(\theta)}{n} = \frac{n_1}{n\sigma_Y^2(1-\rho^2)} + \frac{n_2}{n\sigma_Y^2(1-\rho^2)} = v_{11};
\]
\[
\frac{j_{12}(\theta)}{n} = -\frac{\sum_{i=1}^{n} 2p[y_i - \mu_Y - \rho \Phi^{-1}(u_i)\sigma_Y]}{n\sigma_Y^2(1-\rho^2)} + \frac{\sum_{j=n_1+1}^{n} 2p[y_j - \mu_Y - \rho \Phi^{-1}(u_j)\sigma_Y]}{n\sigma_Y^2(1-\rho^2)} + \frac{\sum_{j=n_1+1}^{n} \Phi^{-1}(u_j)}{n\sigma_Y^2(1-\rho^2)} \quad \overset{p}{\rightarrow} \quad v_{12};
\]
and similarly,
\[
\frac{j_{22}(\theta)}{n} \overset{p}{\rightarrow} v_{22}.
\]

Then
\[
\frac{n j_{22}(\theta)}{j_{11}(\theta) j_{22}(\theta) - j_{12}^2(\theta)} \overset{p}{\rightarrow} \frac{v_{22}}{v_{11}v_{22} - v_{12}^2},
\]
and
\[
\frac{n j_{11}(\theta)}{j_{11}(\theta) j_{22}(\theta) - j_{12}^2(\theta)} \overset{p}{\rightarrow} \frac{v_{11}}{v_{11}v_{22} - v_{12}^2}
\]
as \(n \rightarrow \infty\) and \(n_2/n \rightarrow 1\).

By Theorem 4.3 in Lin, Flourny and Rosenberger (2020), Lemma 3.2 and Lemma 3.3, it follows
\[
\left( \frac{j_{11}(\theta) j_{22}(\theta) - j_{12}^2(\theta)}{j_{22}(\theta)} \right)^{1/2} (\hat{\mu}_n - \mu) \overset{d}{\rightarrow} N(0,1) \quad \text{(mixing)},
\]
\[
\left( \frac{j_{11}(\theta) j_{22}(\theta) - j_{12}^2(\theta)}{j_{11}(\theta)} \right)^{1/2} (\hat{\rho}_n - \rho) \overset{d}{\rightarrow} N(0,1) \quad \text{(mixing)}.
\]
\[
(1 - \hat{\rho}_n^2) \left( \frac{j_{11}(\theta) j_{22}(\theta) - j_{12}^2(\theta)}{j_{11}(\theta)} \right)^{1/2} \left( \frac{1}{2} \ln \frac{1 + \hat{\mu}_n}{1 - \hat{\mu}_n} - \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho} \right) \overset{d}{\rightarrow} N(0,1) \quad \text{(mixing)}
\]
as \(n_2 \rightarrow \infty\).
A.5. Proof of Theorem 3.5. According to Lemma 3.1 in Section 2.2,
\[
\sqrt{n} \left( \hat{\mu}_n - \mu_Y \right) \to V^{-1/2} Z \mathcal{F}_{n_1} - \text{stably}
\]
with $V$ independent of $Z$ as $n_2 \to \infty$. Using Cramer-Wold device,
\[
\sqrt{n} [\hat{\delta}_m(u_j^*) - \delta_m(u_j^*)] = \sqrt{n} \left( \begin{array}{c} \sigma_Y c(u_j^*) \\ \hat{\rho}_n - \rho \end{array} \right) \to \left( \begin{array}{c} \sigma_Y c(u_j^*) \\ \hat{\rho}_n - \rho \end{array} \right) \to \left( \begin{array}{c} \sigma_Y c(u_j^*) \\ \hat{\rho}_n - \rho \end{array} \right)
\]
\[
\to \left( \begin{array}{c} \sigma_Y c(u_j^*) \\ \hat{\rho}_n - \rho \end{array} \right) V^{-1/2} Z \mathcal{F}_{n_1} - \text{stably}
\]
as $n_2 \to \infty$. Based on Equation (A.3),
\[
\sqrt{n} [\hat{\delta}_m(u_j^*) - \delta_m(u_j^*)] \to \left( \begin{array}{c} \sigma_Y c(u_j^*) \\ \hat{\rho}_n - \rho \end{array} \right) \to \left( \begin{array}{c} \sigma_Y c(u_j^*) \\ \hat{\rho}_n - \rho \end{array} \right)
\]
as $n_2 \to \infty$. Again, it is equivalent that
\[
\sqrt{n} [\hat{\delta}_m(u_j^*) - \delta_m(u_j^*)] \to \sqrt{[q_{11} + q_{21} \sigma_Y c(u_j^*)]^2 + [q_{12} + q_{22} \sigma_Y c(u_j^*)]^2} Z \mathcal{F}_{n_1} - \text{stably}
\]
as $n_2 \to \infty$. By substituting in each element in $V$,
\[
\sqrt{[q_{11} + q_{21} \sigma_Y c(u_j^*)]^2 + [q_{12} + q_{22} \sigma_Y c(u_j^*)]^2}
\]
\[
= \sqrt{\frac{v_{22} + \sqrt{v_{11} v_{22} - v_{11}^2}}{v_{11} + v_{22} + 2 \sqrt{v_{11} v_{22} - v_{11}^2}}},
\]
Similarly to Theorem 3.4 above, by Theorem 4.3 in Lin, Fournoy and Rosenberger (2020),
\[
T^{1/2} [\hat{\delta}_m(u_j^*) - \delta_m(u_j^*)] \overset{d}{\to} N(0, 1) \quad \text{(mixing)}
\]
as $n_2 \to \infty$, where
\[
T = \frac{|j_n(\hat{\theta})||j_{11}(\hat{\theta})| + |j_{22}(\hat{\theta})| + 2|j_n(\hat{\theta})|^{1/2}}{[j_{22}(\hat{\theta}) + |j_n(\hat{\theta})|^{1/2} - j_{11}(\hat{\theta})^{1/2} \sigma_Y c(u_j^*)]^{1/2} + [j_{11}(\hat{\theta}) + |j_n(\hat{\theta})|^{1/2} \sigma_Y c(u_j^*) - j_{12}(\hat{\theta})]^{1/2}}.
\]

Acknowledgements. The authors thank an anonymous associate editor and referees for helpful comments that greatly improved the manuscript.

REFERENCES


