Refined moderation analysis with categorical outcomes in precision medicine

Su, X.; Cho, Y.; Liqiang, N.; Liu, L.; Dusseldorp, E.

Received: 6 November 2021 | Revised: 9 November 2022 | Accepted: 30 November 2022

DOI: 10.1002/sim.9627

RESEARCH ARTICLE

Statistics in Medicine WILEY

Refined moderation analysis with categorical outcomes in precision medicine

Xiaogang Su[1] | **Youngjoo Cho**[2] | **Liqiang Ni**[3] | **Lei Liu**[4] | **Elise Dusseldorp**[5]

[1]Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas, USA

[2]Department of Applied Statistics, Konkuk University, Gwangjin-gu, Seoul, Republic of Korea

[3]Department of Statistics and Data Science, University of Central Florida, Orlando, Florida, USA

[4]Division of Biostatistics, Washington University in St. Louis, St. Louis, Missouri, USA

[5]Institute of Psychology, Leiden University, Leiden, Netherlands

Correspondence
Youngjoo Cho, Department of Applied Statistics, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea.
Email: yvc5154@konkuk.ac.kr; yvc5154@gmail.com

Moderation analysis is an integral part of precision medicine research. Concerning moderation analysis with categorical outcomes, we start with an interesting observation, which shows that heterogeneous treatment effects could be equivalently estimated via a role exchange between the outcome and the treatment variable in logistic regression models. Hence two estimators of moderating effects can be obtained. We then established the joint asymptotic normality for the two estimators, on which basis refined inference can be made for moderation analysis. The improved precision is helpful in addressing the lack-of-power problem that is common in search of moderators. The above-mentioned results hold for both experimental and observational data. We investigate the proposed method by simulation and provide an illustration with data from a randomized trial on wart treatment.

KEYWORDS

heterogeneous treatment effects, Logistic regression, Moderation analysis, Precision medicine

1 | INTRODUCTION

Precision medicine aims to tailor medical treatments or interventions to the individual patient profiles, as opposed to the "one-size-fits-all" approach in traditional medicine. This innovative approach involves a long-term endeavor and is a high priority in clinical research.[1] The same conception has been extended to comparative studies in many other fields. To advance precision medicine, it is crucial to identify important moderators and understand their influences on the treatment effects, which prescribes the general scope of moderation analysis. In the moderation analysis literature, closely related concepts such as subgroup analysis, predictive or prescriptive factors, effect modification, stratified/individualized treatment effects, and treatment-by-covariates interactions receive common referrals.

Moderation analysis faces common challenges including, but not limited to, multicollinearity and lack of power. Multicollinearity occurs since moderation analysis is routinely performed by including cross-product terms between treatment and covariates into the model. The correlations between the additive terms and the cross-product terms are naturally high. As a result of multicollinearity, the standard errors of the coefficient estimates are inflated, leading to statistical insignificance. Lack of power emerges since moderation analysis is rarely pre-planned as one of the study aims. Due to time and cost constraints, designed experiments or trials are often merely empowered to establish the efficacy or effectiveness of the treatment in terms of its main effect at some prespecified significance level. The planned sample size may

not be sufficient to allow for a conclusive moderation analysis. This is particularly the case in subgroup analysis where the treatment effects are re-assessed within several subpopulations. The reduced sample size within each subgroup can make it difficult to uncover any interesting findings.

This research is primarily concerned with moderation analysis with a categorical endpoint, where the effect of a treatment is measured by the relative risk (RR) or the odds ratio (OR). It is common practice to conduct moderation analysis via a logistic regression model with interactions. We refer to this model as the direct model. We first present an interesting observation, which shows that heterogeneous treatment effects on the endpoint are equivalent to the heterogeneous endpoint effects on the treatment. This implies that, the roles of the response and the treatment in the logistic model can be swapped. We refer to the interaction model after the role swap as the inverse model. The inverse model can also be meaningfully interpreted in terms of RR or OR. This role-swapping strategy can be useful in certain scenarios where direct modeling is difficult.

Though moderation analysis can be equivalently informed by two interaction logistic models, empirically two different estimators for the moderating effects of covariates can be obtained. We next establish their joint asymptotic normality and put forward a new optimally linearly combined estimator that has a smaller standard error. On this basis, more efficient inferences can be made in moderation analysis. The assessment of the treatment effect depends on the source of data, that is, whether the data are collected from a randomized experiment and an observational study. The above mentioned results hold for both experimental and observational data.

The remainder of the article is arranged as follows. With initial focus on the scenario where both response and treatment are binary, Section 2 describes the rationale for the role-swapping mechanism, which leads to two estimators of moderating effects. In Section 3, the asymptotic joint normality of two estimators is established. An optimally linearly combined estimator for the moderating effect is then developed. Section 4 presents simulation studies that are designed to verify the conception and compare the estimators. An illustrative example from a randomized wart treatment trial is provided in Section 5. In Section 6, we extend the results to the general scenario with categorical outcomes and treatments. Finally, Section 7 concludes with a brief discussion. All proofs and additional numerical results are deferred to the Supplementary Material.

2 | THE ROLE-SWAPPING MECHANISM

Consider data $\{(y_i, t_i, \mathbf{x}_i) : i = 1, \dots, n\}$ that consist of n IID copies of $(y, t, \mathbf{x})$, where y is a categorical response, t is the treatment assignment variable, and $\mathbf{x} \in \mathbb{R}^p$ is the associated P-dimensional covariate vector. The primary goal is to evaluate the efficacy or effectiveness of t on y. We assume both y and t are binary and 0/1-coded for the time being. The treatment variable t has a value of 1 for the treated and 0 for the untreated; the response y has a value 1 indicating the occurrence of an event of interest (such as death) and 0 indicating the absence of the event (eg, survival). Participants with the event are termed as *cases* while those without the event are called *controls* in clinical settings.

The treatment effect is commonly measured on either the relative risks (RR) scale or the odds ratio (OR) scale. Fixing covariates at $\mathbf{x}$, two RR measures are defined as

$$RR_{1,\mathbf{x}} = \frac{\Pr(y = 1|t = 1, \mathbf{x})}{\Pr(y = 1|t = 0, \mathbf{x})}$$

and

$$RR_{0,\mathbf{x}} = \frac{\Pr(y = 0|t = 1, \mathbf{x})}{\Pr(y = 0|t = 0, \mathbf{x})},$$

depending on whether the death rate or the survival rate is considered. The OR measure is given by

$$OR_{\mathbf{x}}^{(y)} = \frac{\Pr(y = 1|t = 1, \mathbf{x})/\Pr(y = 0|t = 1, \mathbf{x})}{\Pr(y = 1|t = 0, \mathbf{x})/\Pr(y = 0|t = 0, \mathbf{x})} = \frac{RR_{1,\mathbf{x}}}{RR_{0,\mathbf{x}}}. \quad (1)$$

It follows that $OR_{\mathbf{x}}^{(y)} \approx RR_{1,\mathbf{x}}$ when cases (with $y = 1$) are rare or $\Pr(y = 0|t, \mathbf{x}) \approx 1$ and $OR_{\mathbf{x}}^{(y)} \approx 1/\mathrm{RR}_{0,\mathbf{x}}$ when controls (with $y = 0$) are rare or $\Pr(y = 1|t, \mathbf{x}) \approx 1$.

Moderation analysis is concerned about differential treatment effects among patients with different characteristics. To assess the heterogeneity of treatment effects, natural measures are the ratio of relative risks (RRR) and the ratio of odds ratios (ROR). Comparing patients with covariates $\mathbf{x}$ and those with covariates $\mathbf{x}'$, define

$$RRR^{(y=1)}(\mathbf{x} : \mathbf{x}') = \frac{RR_{1,\mathbf{x}}}{RR_{1,\mathbf{x}'}} = \frac{\Pr(y=1|t=1,\mathbf{x})/\Pr(y=1|t=0,\mathbf{x})}{\Pr(y=1|t=1,\mathbf{x}')/\Pr(y=1|t=0,\mathbf{x}')} \tag{2}$$

to be the RRR for death or case ($y=1$) between individuals with covariates $\mathbf{x}$ and individuals with covariates $\mathbf{x}'$. Similarly, define

$$RRR^{(y=0)}(\mathbf{x} : \mathbf{x}') = \frac{RR_{0,\mathbf{x}}}{RR_{0,\mathbf{x}'}} = \frac{\Pr(y=0|t=1,\mathbf{x})/\Pr(y=0|t=0,\mathbf{x})}{\Pr(y=0|t=1,\mathbf{x}')/\Pr(y=0|t=0,\mathbf{x}')} \tag{3}$$

for survival or control ($y=0$). It is worth noting that, if $RR_{1,\mathbf{x}} \geq 1$, then we must have $RR_{0,\mathbf{x}} \leq 1$ concerning treatment effect assessment; however, $RRR_1(\mathbf{x} : \mathbf{x}') \geq 1$ does not necessarily imply that $RRR_0(\mathbf{x} : \mathbf{x}') \leq 1$ concerning moderation. On the odds ratio scale, the ratio of odds ratios (ROR) that compares the OR at $\mathbf{x}$ versus that at $\mathbf{x}'$ is given by

$$ROR^{(y)}(\mathbf{x} : \mathbf{x}') = \frac{OR^{(y)}_{\mathbf{x}}}{OR^{(y)}_{\mathbf{x}'}} = \frac{RR_{1,\mathbf{x}}/RR_{0,\mathbf{x}}}{RR_{1,\mathbf{x}'}/RR_{0,\mathbf{x}'}} = \frac{RRR^{(y=1)}(\mathbf{x} : \mathbf{x}')}{RRR^{(y=0)}(\mathbf{x} : \mathbf{x}')}. \tag{4}$$

By the invariance of the odds ratio, there is no need to define $ROR^{(y)}(\mathbf{x} : \mathbf{x}')$ for cases and for controls separately since they are simply inverses of each other.

If we swap the roles of t and y, define

$$OR^{(t)}_{y=1}(\mathbf{x} : \mathbf{x}') = \frac{\Pr(t=1|y=1,\mathbf{x})/\Pr(t=0|y=1,\mathbf{x})}{\Pr(t=1|y=1,\mathbf{x}')/\Pr(t=0|y=1,\mathbf{x}')} \tag{5}$$

to be the odds ratio for event $t=1$ that compares cases ($y=1$) with $\mathbf{x}$ and cases with $\mathbf{x}'$. Similarly, define

$$OR^{(t)}_{y=0}(\mathbf{x} : \mathbf{x}') = \frac{\Pr(t=1|y=0,\mathbf{x})/\Pr(t=0|y=0,\mathbf{x})}{\Pr(t=1|y=0,\mathbf{x}')/\Pr(t=0|y=0,\mathbf{x}')} \tag{6}$$

to be the odds ratio for event $t=1$ that compares controls ($y=0$) with $\mathbf{x}$ and controls with $\mathbf{x}'$. Let $\varphi(\mathbf{x}) = \Pr(t=1|\mathbf{x})$ denote the propensity score.[2] We have the following lemma.

Lemma 1. *Concerning moderation analysis on the RRR scale,*

$$RRR^{(y=1)}(\mathbf{x} : \mathbf{x}') = OR^{(t)}_{y=1}(\mathbf{x} : \mathbf{x}') \cdot \frac{\varphi(\mathbf{x}')/(1-\varphi(\mathbf{x}'))}{\varphi(\mathbf{x})/(1-\varphi(\mathbf{x}))}, \tag{7}$$

and

$$RRR^{(y=0)}(\mathbf{x} : \mathbf{x}') = OR^{(t)}_{y=0}(\mathbf{x} : \mathbf{x}') \cdot \frac{\varphi(\mathbf{x}')/(1-\varphi(\mathbf{x}'))}{\varphi(\mathbf{x})/(1-\varphi(\mathbf{x}))}. \tag{8}$$

Concerning moderation analysis on the ROR scale,

$$ROR^{(y)}(\mathbf{x} : \mathbf{x}') = ROR^{(t)}(\mathbf{x} : \mathbf{x}'), \tag{9}$$

where $ROR^{(t)}(\mathbf{x} : \mathbf{x}')$ is the ratio of odds ratios (ROR) after the role exchange of t and y.

If we further define the odds ratio on propensity as

$$OR^{(t)}(\mathbf{x} : \mathbf{x}') = \frac{\varphi(\mathbf{x})/(1-\varphi(\mathbf{x}))}{\varphi(\mathbf{x}')/(1-\varphi(\mathbf{x}'))}, \tag{10}$$

equations (7) and (8) in Lemma 1 can be rewritten as

$$RRR^{(y=1)}(\mathbf{x}:\mathbf{x}') = \frac{OR^{(t)}_{y=1}(\mathbf{x}:\mathbf{x}')}{OR^{(t)}(\mathbf{x}:\mathbf{x}')} \text{ and } RRR^{(y=0)}(\mathbf{x}:\mathbf{x}') = \frac{OR^{(t)}_{y=0}(\mathbf{x}:\mathbf{x}')}{OR^{(t)}(\mathbf{x}:\mathbf{x}')}.$$

The results of Lemma 1 hold no matter whether data are obtained from a randomized experiment or from an observational study. Moreover, the OR-based equation (9) is applicable to retrospective matched case-control studies.[3]

In the case of randomized experiments, the propensity score is a constant, that is, $\varphi(\mathbf{x}) = \varphi(\mathbf{x}') = \pi \in (0,1)$. Equation (7) reduces to $RRR^{(y=1)}(\mathbf{x}:\mathbf{x}') = OR^{(t)}_{y=1}(\mathbf{x}:\mathbf{x}')$, which is referred to as the case-only analysis in the literature.[4] With rare case events, $RRR^{(y=1)}(\mathbf{x}:\mathbf{x}') \approx ROR^{(y)}(\mathbf{x}:\mathbf{x}')$, implying that $RRR^{(y=1)}(\mathbf{x}:\mathbf{x}')$ can be roughly construed as $ROR^{(y)}(\mathbf{x}:\mathbf{x}')$. The applications and extensions of case-only analysis have been explored by many authors.[5-8] Clearly, a control-only analysis could be proceeded similarly by reducing (8) to $RRR^{(y=0)}(\mathbf{x}:\mathbf{x}') = OR^{(t)}_{y=0}(\mathbf{x}:\mathbf{x}')$. With rare control events, $1/RRR^{(y=0)}(\mathbf{x}:\mathbf{x}') \approx ROR^{(y)}(\mathbf{x}:\mathbf{x}')$.

The above RRR and ROR quantities can be naturally formulated via generalized linear models (GLM), as prescribed by the following theorem.

Theorem 1. *Consider the usual logistic regression model for moderation analysis on the odds ratio (OR) scale, where the conditional distribution of* $(Y \mid t, \mathbf{x})$ *is formulated by*

$$\log \frac{\Pr(y=1|t,\mathbf{x})}{\Pr(y=0|t,\mathbf{x})} = \beta_0 + \beta_1 t + \mathbf{x}^T\boldsymbol{\beta}_2 + t\cdot\mathbf{x}^T\boldsymbol{\beta}_3. \tag{11}$$

Suppose that the conditional distribution of $(t \mid y, \mathbf{x})$ *is formulated by the logistic regression model*

$$\log \frac{\Pr(t=1|y,\mathbf{x})}{\Pr(t=0|y,\mathbf{x})} = \alpha_0 + \alpha_1 y + \mathbf{x}^T\boldsymbol{\alpha}_2 + y\cdot\mathbf{x}^T\boldsymbol{\alpha}_3. \tag{12}$$

Then we must have

$$\boldsymbol{\beta}_3 = \boldsymbol{\alpha}_3,$$

regardless of whether data come from a randomized experiment or from an observational study.

Theorem 1 indicates that the moderation analysis can be conducted via the interaction logistic model (12) that regresses the treatment variable t on the outcome y and covariates $\mathbf{x}$. This amounts to a role exchange between the treatment variable t and the outcome variable y. Since model (12) involves an inverse regression by swapping the roles of t and y, we refer to it as the "inverse model" for simplicity while model (11) is referred to as the "direct model".

We have assumed that model (11) and model (12) are correctly specified as in the case-only analysis.[5-7] Similar scenarios can be found in other statistical approaches such as mediation analysis and structural equation modeling (SEM) where multiple model specifications are entailed. As pointed out by one referee, Theorem 1 still holds if the additive terms, that is, $\boldsymbol{x}^T\boldsymbol{\beta}_2$ in model (11) and $\boldsymbol{x}^T\boldsymbol{\alpha}_2$ in model (12), are left unspecified as unknown nonlinear functions, which won't affect the interpretation of $\boldsymbol{\beta}_3$ and $\boldsymbol{\alpha}_3$ as logarithms of ROR. This would greatly increase the flexibility of the direct and inverse models in approximating the true models even if misspecification occurs. To understand the inverse model (12), it is helpful to note that the distribution of $t|\mathbf{x}$ is different from that of $t|\mathbf{x}, y$. For example, $t \perp\!\!\!\perp \mathbf{x}$ when t is randomized, but $t \not\perp\!\!\!\perp \mathbf{x} \mid y$. model (11) cannot fully determine model (12) and vice versa; other quantities, such as the joint density $f(\mathbf{x}, y)$ or $f(\mathbf{x}, t)$, also play a role. This allows for some leeway to compatibility[9] between these two conditional models. For practical purposes, the two models can be viewed as ways of making approximations and extracting the ROR quantities. As we shall see in simulation studies, the estimates of $\boldsymbol{\beta}_3$ and $\boldsymbol{\alpha}_3$ are generally close to each other even if both models are wrongly specified.

For randomized experiments, additional interpretations concerning moderation analysis on the relative risk scale can be extracted from model (12), as stated in the following proposition.

Proposition 1. *For experiments where the assignment mechanism of treatment t is random, consider either of the following two log-linear regression models for moderation analysis on the relative risk (RR) scale, one for the control event and the other for the case event*,

$$\begin{cases} \log \Pr(y=0|t,\mathbf{x}) = \gamma_0^{(0)} + \gamma_1^{(0)} t + \mathbf{x}^T \boldsymbol{\gamma}_2^{(0)} + t \cdot \mathbf{x}^T \boldsymbol{\gamma}_3^{(0)}, \\ \log \Pr(y=1|t,\mathbf{x}) = \gamma_0^{(1)} + \gamma_1^{(1)} t + \mathbf{x}^T \boldsymbol{\gamma}_2^{(1)} + t \cdot \mathbf{x}^T \boldsymbol{\gamma}_3^{(1)}. \end{cases} \tag{13}$$

Furthermore assume that model (12) formulates the conditional distribution of $(t \mid y, \mathbf{x})$. *Then we must have*

$$\boldsymbol{\gamma}_3^{(0)} = \boldsymbol{\alpha}_2 \text{ or } \boldsymbol{\gamma}_3^{(1)} = \boldsymbol{\alpha}_2 + \boldsymbol{\alpha}_3.$$

Proposition 1 essentially breaks model (12) into two equations,

$$\log \frac{\Pr(t=1|y,\mathbf{x})}{\Pr(t=0|y,\mathbf{x})} = \begin{cases} \alpha_0 + \mathbf{x}^T \boldsymbol{\alpha}_2, & \text{if } y=0; \\ (\alpha_0 + \alpha_1) + \mathbf{x}^T (\boldsymbol{\alpha}_2 + \boldsymbol{\alpha}_3), & \text{if } y=1, \end{cases}$$

one for controls only and the other for cases only. It requires either, not necessarily both, of the two models in (13) hold, as $\Pr(y=0|t,\mathbf{x}) + \Pr(y=1|t,\mathbf{x}) = 1$. If both models in (13) hold, then the parameters can be reduced by introducing constraints. In this case, model (11) must hold with $\boldsymbol{\beta}_j = \boldsymbol{\gamma}_j^{(1)} - \boldsymbol{\gamma}_j^{(0)}$ for $j = 0, 1, 2, 3$, but not *vice versa*.

We defer all the proofs to the Supplementary Material. The proof of Lemma 1 essentially involves applications of Bayes' rule. Similar arguments have been used to derive the case-only analysis[4] and to show equivalence of the conditional odds ratio obtained from a prospective study and that from a retrospective study.[3] As a special case of moderation analysis with categorical outcomes, testing heterogeneity of odds ratios across levels of a categorical covariate such as different centers in a multicenter clinical trial has been previously considered by many authors including the classical Q-statistic.[10] In this simplified scenario, Theorem 1 and Proposition 1 hold trivially. However, our results cover more general cases that possibly involve higher-order interactions and covariate adjustment and allow for continuous covariates. In terms of inverse regression with role swapping, Efron[11] has shown that the logistic regression model for $(y|\mathbf{x})$ has the same coefficients as in the Gaussian linear discriminant analysis that models $(\mathbf{x}|y)$, under the assumption that $\mathbf{x}$ follows a multivariate normal distribution. Our results are derived in the same spirit of Bayes' rule, but are quite different in that the logistic model (11) for $(y|t,\mathbf{x})$ has the same set of coefficients as the logistic model (12) for $(t|y,\mathbf{x})$ only for interaction terms, regardless of the distribution of $\mathbf{x}$.

3 | REFINED MODERATION ANALYSIS

We have shown that moderation analysis could be equivalently conducted by swapping the roles of the response and the treatment in logistic regression models. This motivates possibly more efficient ways of making inference on the moderating effects. Note that a common estimator is not readily attainable by setting $\boldsymbol{\alpha}_3 = \boldsymbol{\beta}_3$ in model (11) and model (12) and simultaneously estimating both models. This is because the two models have their respective likelihood functions with the same data set, leading to a multi-objective optimization problem. Let $\widehat{\boldsymbol{\alpha}}_3$ and $\widehat{\boldsymbol{\beta}}_3$ denote the maximum likelihood estimator (MLE) of $\boldsymbol{\alpha}_3$ and $\boldsymbol{\beta}_3$ obtained separately from model (11) and model (12). In the following, we first establish the joint asymptotic normality for $\widehat{\boldsymbol{\alpha}}_3$ and $\widehat{\boldsymbol{\beta}}_3$ and then seek ways of making more efficient inferences.

Denote $\boldsymbol{\beta} = (\beta_0, \beta_1, \boldsymbol{\beta}_2^T, \boldsymbol{\beta}_3^T)^T = (\boldsymbol{\beta}_{(3)}^T, \boldsymbol{\beta}_3^T)^T$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \boldsymbol{\alpha}_2^T, \boldsymbol{\alpha}_3^T)^T = (\boldsymbol{\alpha}_{(3)}^T, \boldsymbol{\alpha}_3^T)^T$, where $\boldsymbol{\beta}_{(3)} = (\beta_0, \beta_1, \boldsymbol{\beta}_2^T)^T$ contains the parameters in model (11) excluding $\boldsymbol{\beta}_3$ and similarly for $\boldsymbol{\alpha}_{(3)} = (\alpha_0, \alpha_1, \boldsymbol{\alpha}_2^T)^T$. Since the focus of moderation analysis is on $\boldsymbol{\beta}_3$ and $\boldsymbol{\alpha}_3$, we treat $\boldsymbol{\beta}_{(3)}$ and $\boldsymbol{\alpha}_{(3)}$ as nuisance parameters in some sense. Let $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} L_y(\boldsymbol{\beta})$ and $\widehat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} L_t(\boldsymbol{\alpha})$ be their MLEs, where $L_y(\boldsymbol{\beta})$ and $L_t(\boldsymbol{\alpha})$ are the log-likelihood functions for model (11) and model (12), respectively. We shall derive the joint asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$, from which the joint asymptotic distribution of $\widehat{\boldsymbol{\beta}}_3$ and $\widehat{\boldsymbol{\alpha}}_3$ follows immediately.

To set up, consider the setting with stochastic regressors where observations $(y_i, t_i, \mathbf{x}_i^T)^T$ are IID copies of $(y, t, \mathbf{x})$ that has joint density $f(y, t, \mathbf{x}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to some measure. The joint density can be decomposed as

$$f(y, t, \mathbf{x}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(y|t, \mathbf{x}; \boldsymbol{\beta}) f(t, \mathbf{x}) = f(t|y, \mathbf{x}; \boldsymbol{\alpha}) f(y, \mathbf{x}),$$

where $f(y|t, \mathbf{x}; \boldsymbol{\beta})$ is determined by model (11) and $f(t|y, \mathbf{x}; \boldsymbol{\beta})$ is determined by model (12). Let $\boldsymbol{X}_y = (\mathbf{x}_{yi}^T)$ and $\boldsymbol{X}_t = \left(\mathbf{x}_{ti}^T\right)$ denote the design matrices associated with the two models, with $\mathbf{x}_{yi}^T = (1, t_i, \mathbf{x}_i^T, t\ \mathbf{x}_i^T)$ and $\mathbf{x}_{ti}^T = (1, y_i, \mathbf{x}_i^T, y_i\ \mathbf{x}_i^T)$ being their respective ith row vectors for $i = 1, \ldots, n$. We further assume both $\boldsymbol{X}_y$ and $\boldsymbol{X}_t$ are of full column rank almost surely (w.p. 1)

for any large enough n, that is, $\forall n \geq n_0$ for some $n_0 \in \mathbb{N}$. Denote the modeled probabilities by $\boldsymbol{\pi}_y = \left\{1 + \exp(-\boldsymbol{X}_y\boldsymbol{\beta})\right\}^{-1}$ and $\boldsymbol{\pi}_t = \{1 + \exp(-\boldsymbol{X}_t\boldsymbol{\alpha})\}^{-1}$.

Estimation of $\boldsymbol{\beta}$ can be made by maximizing the log-likelihood $L_y(\boldsymbol{\beta})$ of model (11) that is associated with $f(y|t, \mathbf{x}; \boldsymbol{\beta})$ alone, while estimation of $\boldsymbol{\alpha}$ can be made by maximizing the log-likelihood $L_t(\boldsymbol{\alpha})$ of model (12) that is associated with $f(t|y, \mathbf{x}; \boldsymbol{\alpha})$ alone. Since the resultant MLEs $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ are obtained separately, this multi-objective problem renders the usual arguments for M-estimators inapplicable. Nevertheless, their joint asymptotic normality can still be justified by viewing the associated score equations as estimating equations and treating $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ as Z-estimators.[12,13]

We assume the following technical conditions:

(i) $E(\mathbf{x}\mathbf{x}^T) \succ 0$, where '$\succ 0$' denotes "positive definite".
(ii) The conditional probabilities π_y and π_t are bounded away from 0 and 1 almost everywhere (a.e.), for example, $\pi_y \in (a, b)$ for some $0 < a < b < 1$; so are $\Pr(t = 1|\mathbf{x})$ and $\Pr(y = 1|\mathbf{x})$.

Theorem 2. *Under conditions (i) and (ii), $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ asymptotically exist and are strongly consistent. Furthermore, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ have a degenerate joint normal distribution asymptotically*

$$\boldsymbol{\Sigma}^{-1/2}\left[\begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\alpha}}\end{pmatrix} - \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\alpha}\end{pmatrix}\right] \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

The asymptotic variance-covariance (VCOV) matrix $\boldsymbol{\Sigma}$ has a "sandwich" form $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{M}\mathbf{B}^T$ with "bread" matrix

$$\mathbf{B} = \left[\begin{array}{c|cc} -E(\boldsymbol{X}_y^T\mathbf{W}_y\boldsymbol{X}_y) & -E(\boldsymbol{X}_y^T\mathbf{W}_y\boldsymbol{D}_y\mathbf{W}_t\boldsymbol{X}_t^{(A)}) & -E(\boldsymbol{X}_y^T\mathbf{W}_y\boldsymbol{X}_y^{(I)}) \\ \hline -E(\boldsymbol{X}_t^T\mathbf{W}_t\boldsymbol{D}_t\mathbf{W}_y\boldsymbol{X}_y^{(A)}) & -E(\boldsymbol{X}_t^T\mathbf{W}_t\boldsymbol{X}_t^{(I)}) & -E(\boldsymbol{X}_t^T\mathbf{W}_t\boldsymbol{X}_t) \end{array}\right]^{-1} \tag{14}$$

and 'meat' matrix

$$\mathbf{M} = \begin{bmatrix} E\left\{\boldsymbol{X}_y^T(\mathbf{y} - \boldsymbol{\pi}_y)(\mathbf{y} - \boldsymbol{\pi}_y)^T\boldsymbol{X}_y\right\} & E\left\{\boldsymbol{X}_y^T(\mathbf{y} - \boldsymbol{\pi}_y)(\mathbf{t} - \boldsymbol{\pi}_t)^T\boldsymbol{X}_t\right\} \\ E\left\{\boldsymbol{X}_t^T(\mathbf{t} - \boldsymbol{\pi}_t)(\mathbf{y} - \boldsymbol{\pi}_y)^T\boldsymbol{X}_y\right\} & E\left\{\boldsymbol{X}_t^T(\mathbf{t} - \boldsymbol{\pi}_t)(\mathbf{t} - \boldsymbol{\pi}_t)^T\boldsymbol{X}_t\right\} \end{bmatrix}, \tag{15}$$

where $\mathbf{W}_y = \operatorname{diag}\left(\boldsymbol{\pi}_y(\mathbf{1} - \boldsymbol{\pi}_y)\right)$, $\mathbf{W}_t = \operatorname{diag}\left(\boldsymbol{\pi}_t(\mathbf{1} - \boldsymbol{\pi}_t)\right)$, $\boldsymbol{D}_y = \operatorname{diag}\left(\beta_1 + \boldsymbol{\beta}_3^T\mathbf{x}_i\right)$ with diagonal elements $\beta_1 + \boldsymbol{\beta}_3^T\mathbf{x}_i$ for $i = 1, \ldots, n$, $\boldsymbol{D}_t = \operatorname{diag}\left(\alpha_1 + \boldsymbol{\alpha}_3^T\mathbf{x}_i\right)$, $\boldsymbol{X}_t = \left[\boldsymbol{X}_t^{(A)}, \quad \boldsymbol{X}_t^{(I)}\right]$ with $\boldsymbol{X}_t^{(A)} = \left(1, y_i, \boldsymbol{x}_i^T\right)$ being its sub-matrix corresponding to the additive terms in model (12) and $\boldsymbol{X}_t^{(I)} = \left(y_i\boldsymbol{x}_i^T\right)$ being the submatrix involved in the interaction terms, and similarly for $\boldsymbol{X}_y = \left[\boldsymbol{X}_y^{(A)}, \quad \boldsymbol{X}_y^{(I)}\right]$ with $\boldsymbol{X}_y^{(A)} = \left(1, t_i, \boldsymbol{x}_i^T\right)$ and $\boldsymbol{X}_y^{(I)} = \left(t_i\boldsymbol{x}_i^T\right)$.

Several comments are in order. First, the asymptotic joint normal distribution of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ is degenerate because $\boldsymbol{\beta}_3 = \boldsymbol{\alpha}_3$. For the same reason, the bread matrix $\mathbf{B}$ is asymmetric. As matrix $\boldsymbol{\Sigma}$ is not necessarily positive definite, the square root operation $\boldsymbol{\Sigma}^{-1/2}$ should be taken liberally as, for example, its Cholesky factorization. Here, $\boldsymbol{\Sigma}^{-1/2}$ is merely used as a normalization matrix in stating the asymptotic distribution. Secondly, the required assumptions for Theorem 2 are quite simple and mild; one main reason is because both y and t are bounded. Condition (i) is a common assumption in asymptotics for GLM. Combining condition (i) with condition (ii) leads to a sufficient condition as in Fahrmeir and Kaufmann[14] to ensure consistency and asymptotic normality of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$. Thirdly, besides $\boldsymbol{\alpha}_3 = \boldsymbol{\beta}_3$, $f(y|t, \mathbf{x}; \boldsymbol{\beta})$ may depend on $\boldsymbol{\alpha}_{(3)}$ while $f(t|y, \mathbf{x}; \boldsymbol{\beta})$ may depend on $\boldsymbol{\beta}_{(3)}$. As a result, parameters $\boldsymbol{\beta}_{(3)} = (\beta_0, \beta_1, \boldsymbol{\beta}_2^T)^T$ and $\boldsymbol{\alpha}_{(3)} = (\alpha_0, \alpha_1, \boldsymbol{\alpha}_2^T)^T$ are not orthogonal in the sense of Cox and Reid.[15] To evaluate the corresponding components of matrix $\boldsymbol{B}$ in (14), we have applied some approximations as detailed in Section I of the Supplementary Material. We would like to argue that these approximated components are only related to $\boldsymbol{\alpha}_{(3)}$ and $\boldsymbol{\beta}_{(3)}$, which may be deemed as nuisance parameters in moderation analysis. In numerical studies that follow, we shall see that the results based on this approximation are quite satisfactory. Fourthly, standard ML arguments can be used to establish the marginal asymptotic distributions of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$, which follow $\mathcal{N}(\boldsymbol{\beta}, \{E(\boldsymbol{X}_y^T\mathbf{W}_y\boldsymbol{X}_y)\}^{-1})$ and $\mathcal{N}(\boldsymbol{\alpha}, \{E(\boldsymbol{X}_t^T\mathbf{W}_t\boldsymbol{X}_t)\}^{-1})$, respectively. Numerically, the estimated (marginal) VCOV matrices for $\widehat{\boldsymbol{\beta}}$ or $\widehat{\boldsymbol{\alpha}}$ are very similar to those obtained using the sandwich formula in Theorem 2.

To estimate the asymptotic VCOV matrix $\boldsymbol{\Sigma}$, one replaces $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ with their MLEs $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ in quantities such as $\boldsymbol{\pi}_y$ and $\boldsymbol{\pi}_t$ and uses the empirical form of each expectation term. Specifically, the estimated "bread" matrix is

$$\widehat{\mathbf{B}} = \left[\begin{array}{cc|cc} & & & \\ & -\boldsymbol{X}_y^T\widehat{\mathbf{W}}_y\boldsymbol{X}_y & -\boldsymbol{X}_y^T\widehat{\mathbf{W}}_y\widehat{\boldsymbol{D}}_y\widehat{\mathbf{W}}_t\boldsymbol{X}_t^{(A)} & -\boldsymbol{X}_y^T\widehat{\mathbf{W}}_y\widehat{\boldsymbol{X}}_y^{(I)} \\ & & & \\ \hline -\boldsymbol{X}_t^T\widehat{\mathbf{W}}_t\widehat{\boldsymbol{D}}_t\widehat{\mathbf{W}}_y\boldsymbol{X}_y^{(A)} & -\boldsymbol{X}_t^T\widehat{\mathbf{W}}_t\widehat{\boldsymbol{X}}_t^{(I)} & -\boldsymbol{X}_t^T\widehat{\mathbf{W}}_t\boldsymbol{X}_t & \end{array}\right]^{-1} \tag{16}$$

where $\widehat{\mathbf{W}}_y = \text{diag}\left[\hat{\boldsymbol{\pi}}_y(\mathbf{1}-\hat{\boldsymbol{\pi}}_y)\right]$, $\widehat{\mathbf{W}}_t = \text{diag}\left[\hat{\boldsymbol{\pi}}_t(\mathbf{1}-\hat{\boldsymbol{\pi}}_t)\right]$, $\widehat{\boldsymbol{X}}_t^{(I)}$ is obtained from $\boldsymbol{X}_t^{(I)}$ by replacing $\mathbf{y}$ with $\hat{\boldsymbol{\pi}}_y$, and similarly for $\widehat{\boldsymbol{X}}_y^{(I)}$. The estimated meat matrix is

$$\begin{aligned}\widehat{\mathbf{M}} &= \begin{bmatrix} \boldsymbol{X}_y^T\boldsymbol{R}_{yy}\boldsymbol{X}_y & \boldsymbol{X}_y^T\boldsymbol{R}_{yt}\boldsymbol{X}_t \\ \boldsymbol{X}_t^T\boldsymbol{R}_{ty}\boldsymbol{X}_y & \boldsymbol{X}_t^T\boldsymbol{R}_{tt}\boldsymbol{X}_t \end{bmatrix} \\ &= \begin{bmatrix} \sum\limits_{i=1}\mathbf{x}_{yi}\mathbf{x}_{yi}^T(y_i-\hat{\pi}_{yi})^2 & \sum_{i=1}^n\mathbf{x}_{yi}\mathbf{x}_{ti}^T(y_i-\hat{\pi}_{yi})(t_i-\hat{\pi}_{ti}) \\ \sum_{i=1}^n\mathbf{x}_{yi}\mathbf{x}_{ti}^T(y_i-\hat{\pi}_{yi})(t_i-\hat{\pi}_{ti}) & \sum\limits_{i=1}\mathbf{x}_{ti}\mathbf{x}_{ti}^T(t_i-\hat{\pi}_{ti})^2 \end{bmatrix},\end{aligned} \tag{17}$$

where $\boldsymbol{R}_{yy} = \text{diag}\left[(y_i-\hat{\pi}_{yi})^2\right]$, $\boldsymbol{R}_{yt} = \text{diag}\left[(y_i-\hat{\pi}_{yi})(t_i-\hat{\pi}_{ti})\right]$, and $\boldsymbol{R}_{tt} = \text{diag}\left[(t_i-\hat{\pi}_{ti})^2\right]$ are diagonal matrices containing squares and cross-products of the residuals. For better finite-sample performance, an adjustment of $\widehat{\mathbf{M}} := n/(n-3p-4)\cdot\widehat{\mathbf{M}}$, where $(3p+4)$ is the total number of distinct parameters in both model (11) and model (12), is recommended in the sandwich VCOV matrix estimators.[16]

Given the distribution of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})$, the asymptotic joint normality of $(\widehat{\boldsymbol{\beta}}_3, \widehat{\boldsymbol{\alpha}}_3)$ can be obtained by extracting the corresponding components. We are interested in more efficient inference on the moderating effects. One natural approach is to combine the two estimators, $\widehat{\boldsymbol{\beta}}_3$ and $\widehat{\boldsymbol{\alpha}}_3$, into a more precise one. Optimally combining two unbiased estimators has been discussed in the literature. Graybill[17] considered combining two independent unbiased estimators; similar work can also be seen in meta-analysis. Samuel-Cahn[18] tackled the problem of how to combine two univariate correlated unbiased estimators. In the following, we present a general result on optimally combining K multivariate correlated unbiased estimators.

Proposition 2. *Suppose that $\widehat{\boldsymbol{\theta}}_k$ for $k = 1, \ldots, K$ are K unbiased estimators of parameter $\boldsymbol{\theta} \in \mathbb{R}^p$ with variance-covariance matrix $\boldsymbol{V} \in \mathbb{R}^{pK\times pK}$. Namely,*

$$\begin{pmatrix} \widehat{\boldsymbol{\theta}}_1 \\ \vdots \\ \widehat{\boldsymbol{\theta}}_K \end{pmatrix} \sim \left[\begin{pmatrix} \boldsymbol{\theta} \\ \vdots \\ \boldsymbol{\theta} \end{pmatrix}, \boldsymbol{V}\right].$$

Let $\boldsymbol{U} = \boldsymbol{V}^{-1}$ and partition $\boldsymbol{U}$ into K^2 matrices $\left\{\boldsymbol{U}_{kk'} \in \mathbb{R}^{p\times p} : k, k' = 1, \ldots, K\right\}$. Then the minimum-variance unbiased estimator of $\boldsymbol{\theta}$ that linearly combines $\widehat{\boldsymbol{\theta}}_k$'s is given by

$$\widetilde{\boldsymbol{\theta}} = \left[\sum_{k,k'=1}^K \boldsymbol{U}_{kk'}\right]^{-1}\sum_{k=1}^K\sum_{k'=1}^K\boldsymbol{U}_{k'k}\widehat{\boldsymbol{\theta}}_k, \tag{18}$$

with variance

$$\text{cov}(\widetilde{\boldsymbol{\theta}}) = \left[\sum_{k,k'=1}^K \boldsymbol{U}_{kk'}\right]^{-1}.$$

The key step of the proof is to reformulate the problem into a linear model with correlated errors, which can be estimated by the generalized least squares (GLS) approach. We have used the notation $\boldsymbol{x} \sim [\boldsymbol{\mu}, \boldsymbol{\Sigma}]$ to denote that the random vector $\boldsymbol{x}$ has mean $\boldsymbol{\mu}$ and VCOV $\boldsymbol{\Sigma}$. There is no need to assume normality with the GLS approach. The refinement or improvement from the combined estimator is guaranteed by the Gauss-Markov theorem, unless one individual estimator is the UMVUE (uniformly minimum-variance unbiased estimator). The difficult part of using this combining method for estimation, however, is the derivation of the (asymptotic) joint distributions of the individual estimators. Since these individual estimators are obtained from the same data, deriving their (asymptotic) joint distribution can be quite challenging.

When applied to $\widehat{\boldsymbol{\beta}}_3$ and $\widehat{\boldsymbol{\alpha}}_3$ with $K = 2$, where

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}_3 \\ \widehat{\boldsymbol{\alpha}}_3 \end{pmatrix} \sim \left[\begin{pmatrix} \boldsymbol{\beta}_3 \\ \boldsymbol{\beta}_3 \end{pmatrix}, \boldsymbol{\Sigma}' \right]$$

and $\mathbf{U} = \boldsymbol{\Sigma}'^{-1} = (\mathbf{U}_{kk'})$ for $k, k' = 1, 2$, the minimum-variance unbiased estimator of $\boldsymbol{\beta}_3$ that linearly combines $\widehat{\boldsymbol{\beta}}_3$ and $\widehat{\boldsymbol{\alpha}}_3$ is given by

$$\tilde{\boldsymbol{\beta}}_3 = (\mathbf{U}_{11} + \mathbf{U}_{21} + \mathbf{U}_{12} + \mathbf{U}_{22})^{-1} \left\{ (\mathbf{U}_{11} + \mathbf{U}_{21})\widehat{\boldsymbol{\beta}}_3 + (\mathbf{U}_{12} + \mathbf{U}_{22})\widehat{\boldsymbol{\alpha}}_3 \right\}, \tag{19}$$

with variance

$$\text{cov}(\tilde{\boldsymbol{\beta}}_3) = (\mathbf{U}_{11} + \mathbf{U}_{21} + \mathbf{U}_{12} + \mathbf{U}_{22})^{-1}.$$

One is often interested in inference on each component of $\boldsymbol{\beta}_3$. In the univariate case, let $\boldsymbol{\Sigma}' = (\sigma_{kk'})$ for $k, k' = 1, 2$. The above formula reduces to the result of Samuel-Cahn:[18]

$$\tilde{\beta}_3 = \lambda\hat{\beta}_3 + (1 - \lambda)\hat{\alpha}_3 \;\text{ with }\; \text{var}(\tilde{\beta}_3) = \lambda^2\sigma_{11} + (1 - \lambda)^2\sigma_{22} + 2\sigma_{12}, \tag{20}$$

where $\lambda = (\sigma_{22} - \sigma_{12})/(\sigma_{11} + \sigma_{22} - 2\sigma_{12})$. On the basis of Proposition 2, inference on $\boldsymbol{\beta}_3$ or its components can be made accordingly via Wald-type tests. As have been pointed out in the literature,[17,18] the optimally combined estimator may not be a weighted average of the two unbiased estimators since the coefficients λ and $(1 - \lambda)$, which are essentially GLS estimates of coefficients in a linear model, are not restricted to the range of $[0, 1]$.

4 | SIMULATION STUDIES

We validate the role-swapping mechanism and investigate different estimation methods by simulation. Data are generated from model (11), with $\mathbf{X} = (X_1, X_2, X_3)^T$, $\beta_0 = -1$, $\beta_1 = 0.5$, $\boldsymbol{\beta}_2 = (0.5, -1, 1)^T$, and $\boldsymbol{\beta}_3 = (-0.5, 1, 0)^T$. Three covariates are simulated from a multivariate uniform[0, 1] distribution with correlation matrix of form $\left(\rho^{|j-j'|}\right)$ for $j, j' = 1, 2, 3$; see the implementation in R[19] package **MultiRNG**.[20] Different choices of $\rho \in \{0.0, 0.2, 0.5, 0.8\}$ are tried out. To imitate observational studies, the binary treatment t is generated from the following logistic model

$$\log \frac{\Pr(t = 1|\mathbf{x})}{\Pr(t = 0|\mathbf{x})} = b_0 + \boldsymbol{b}_1^T\mathbf{x}, \tag{21}$$

where $b_0 = -0.5$ and $\boldsymbol{b}_1 = (1, -0.5, 1)^T$. To have data from randomized experiments, we set both b_0 and $\boldsymbol{b}_1$ as 0 so that t is simulated with $\Pr(t = 1) = 1/2$, which is independent of $\mathbf{x}$. With this setting, both the prevalence $\Pr(y = 1)$ and propensity $\Pr(t = 1)$ rates are about 50% in a generated data set.

For each model configuration, several sample sizes $n \in \{150{,}300, \ldots, 1500\}$ are examined. A total of 1000 simulation runs are made for each setting. With each simulated data set, we fit model (11) and model (12) and extract the estimates, $\widehat{\boldsymbol{\beta}}_3$ and $\widehat{\boldsymbol{\alpha}}_3$, of coefficients associated with the interaction terms, as well as their standard errors (SE). Then formula (20) is used to obtain the combined estimator $\tilde{\boldsymbol{\beta}}_3$ and the corresponding SEs. Thus $\widehat{\boldsymbol{\beta}}_3$, $\widehat{\boldsymbol{\alpha}}_3$, and $\tilde{\boldsymbol{\beta}}_3$ are all estimates of $\boldsymbol{\beta}_3 = (-0.5, 1, 0)^T$. While $\beta_{33} = 0$ helps evaluate the "size" issue in hypothesis testing, the two nonzero coefficients, β_{31} and β_{32}, help address the "power" issue. To this end, the P-values from the Wald z tests for each coefficient are also recorded. The empirical power and size are computed as proportions of P-values less than $\alpha = 0.05$.

Table 1 presents the aggregated results over 1000 simulation runs for the experimental data setting with $\rho \in \{0, 0.5\}$ and $n \in \{150{,}900\}$. The three estimation methods, that is, direct, inverse, and combined estimators, are respectively referred to as method I, II, and III in the table. The performance measures include the mean and standard deviation (SD) of the estimates and the averaged standard errors (ASE). If the standard error formula works effectively, ASE should be close to SD. Additionally, empirical size/power values are presented in the last three columns of the table.

TABLE 1 Simulation results with randomized experimental data based on 1000 simulation runs

			Estimate (mean)			SD			ASE			Size/power		
ρ	n		I	II	III	I	II	III	I	II	III	I	II	III
0	150	β_{31}	−0.569	−0.563	−0.524	1.329	1.312	1.280	1.274	1.264	1.203	0.081	0.068	0.076
		β_{32}	1.043	1.028	1.022	1.326	1.321	1.293	1.281	1.267	1.185	0.133	0.130	0.168
		β_{33}	0.001	0.015	0.053	1.307	1.292	1.292	1.288	1.276	1.209	0.043	0.044	0.059
	900	β_{31}	−0.502	−0.484	−0.478	0.480	0.473	0.467	0.491	0.485	0.479	0.168	0.159	0.160
		β_{32}	1.015	0.991	0.995	0.511	0.506	0.506	0.493	0.488	0.479	0.543	0.539	0.554
		β_{33}	−0.002	0.013	0.015	0.510	0.507	0.503	0.495	0.490	0.483	0.054	0.055	0.058
0.5	150	β_{31}	−0.506	−0.506	−0.477	1.523	1.521	1.468	1.459	1.452	1.391	0.059	0.065	0.065
		β_{32}	1.041	1.040	1.000	1.745	1.738	1.680	1.624	1.613	1.532	0.103	0.099	0.106
		β_{33}	−0.007	−0.011	0.022	1.509	1.499	1.438	1.464	1.455	1.389	0.050	0.047	0.053
	900	β_{31}	−0.526	−0.520	−0.513	0.575	0.571	0.566	0.558	0.555	0.549	0.171	0.171	0.171
		β_{32}	1.014	1.001	0.994	0.644	0.634	0.633	0.622	0.616	0.607	0.374	0.368	0.378
		β_{33}	0.016	0.016	0.024	0.590	0.582	0.579	0.562	0.558	0.551	0.059	0.060	0.061

Note: For each data set, the regression coefficient $\boldsymbol{\beta}_3$ associated with moderation analysis are estimated with three methods: (I) the direct estimator $\widehat{\boldsymbol{\beta}}_3$ with model (11); (II) the inverse estimator $\widehat{\boldsymbol{\alpha}}_3$ with model (12); and (III) the combined estimator $\tilde{\boldsymbol{\beta}}_3$.

First of all, it can be seen that the direct estimates $\widehat{\boldsymbol{\beta}}_3$ and the inverse estimates $\widehat{\boldsymbol{\alpha}}_3$ are quite close to each other by all measures. This empirically verifies the validity of the role-swapping mechanism. In this simulation setting, model (11) is an over-fitted model while model (12) is only an approximating model, compared to the true models for generating data; nevertheless, the conclusions in Theorem 1 show robustness regardlessly. We would like to point out that, although their similarity holds well at the aggregated level, a direct estimate could be quite different from its corresponding inverse estimate individually, especially with small samples.

On another note, we have skipped simulation studies for verifying Proposition 1. This is because Proposition 1, applicable only to experimental data with randomization, has been partly investigated elsewhere; see, for example, Dai, Li, and Gilbert,[6] Dai et al,[7] and Dai and LeBlanc[8] for simulation studies that are designed to verify the case-only analysis. The control-only analysis, albeit new, holds naturally by symmetry.

Among all three methods, the combined estimator (III) consistently yields the smallest variation as indicated by either SD or ASE. Compared to SD, the SE formulas from all three methods exhibit a degree of underestimation of the true variation when the sample size ($n = 150$) is small. This problem vanishes with larger samples ($n = 900$). In fact, SD and ASE match very well when $n = 300$ or larger. This is because the SE formulas are derived as asymptotic results. In terms of empirical sizes, we obtained two expected ranges in our setting based on concentration inequalities; see Section S.1.2 of the Supplementary Material. Applying the threshold significance level $\alpha = 0.05$, we would expect the empirical size to fall in the interval $(0.0192, 0.0808)$ with probability at least 95% by Chebyshev's inequality. In addition, Chernoff's bounds leads to an interval of $(0.0265, 0.0735)$. It can be seen that the empirical sizes from all three methods well stay within these nominal ranges. In terms of empirical powers, the combined estimator tends to have a higher empirical power than the other two methods in most cases, although the results are mixed occasionally. This can be explained by the higher precision achieved by the combined estimator.

The correlation ρ among covariates also affects the results vastly. Comparing the results when $\rho = 0$ and the results when $\rho = 0.5$ in Table 1, it can be seen that the SD and ASE values increase and the empirical power reduces for every estimation method when covariates are more correlated and multicollinearity becomes worse. However, the conclusions concerning the comparison of the three estimation methods remain. Namely, the combined estimator outperforms the other two by offering higher precision and increased empirical powers.

Figure 1 presents some additional graphical exploration of the results when $\rho = 0.5$. The upper panels present the boxplots of standard errors (SE) of $\widehat{\beta}_{3j}$ (method I in orange) and $\tilde{\beta}_{3j}$ (method III in blue) for $j = 1, 2, 3$ with varying sample sizes $n \in \{150, 300, \ldots, 1500\}$. The SDs of these estimates from 1000 simulation runs are also superimposed as lines on

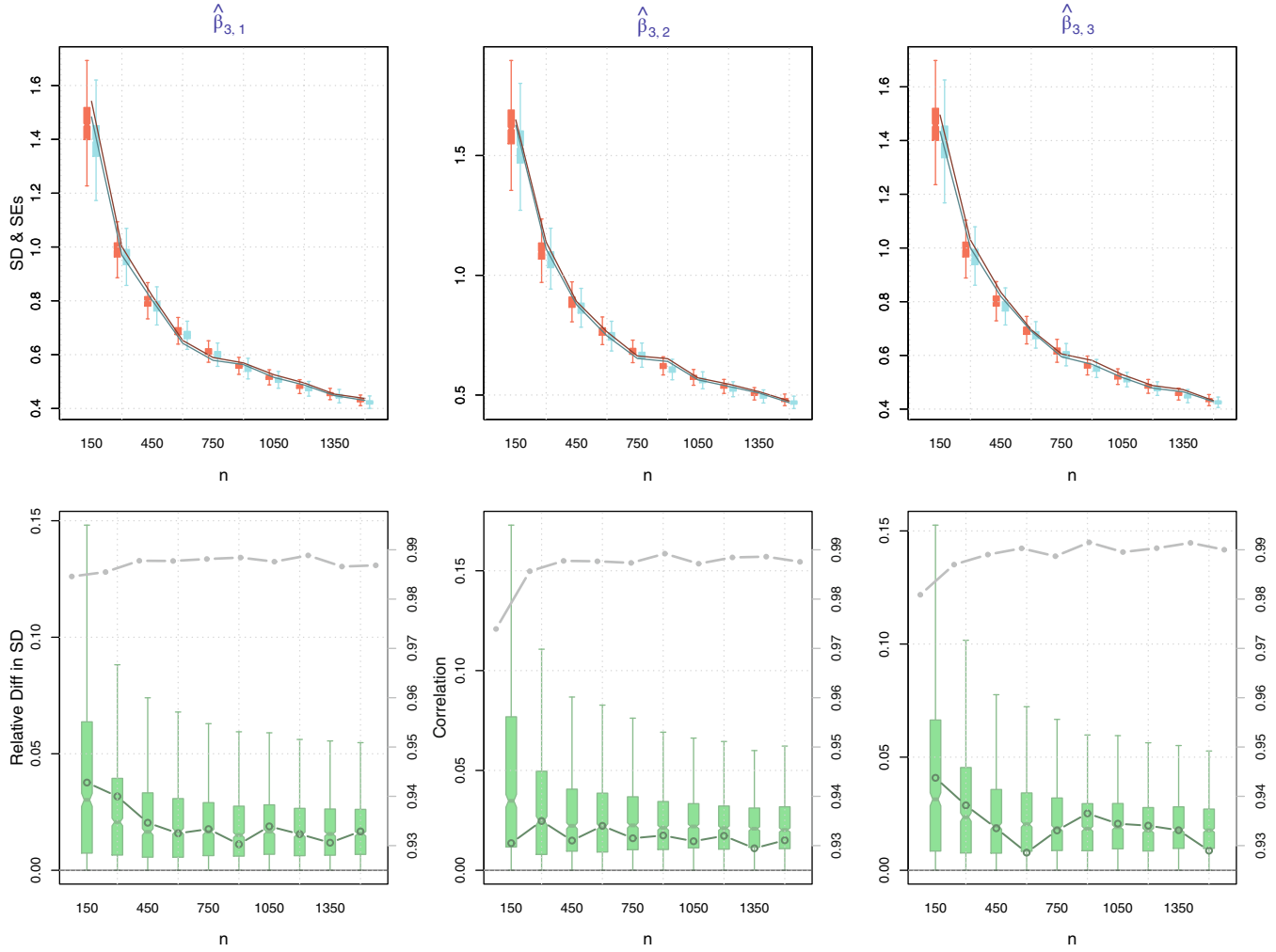

FIGURE 1 Simulation results with experimental data: sample size $n = 150, 300, \ldots, 1500$, `nrun`= 1000, and $\rho = 0.5$. The upper figures present the boxplots of SE values for the direct estimators (in orange) and the combined estimators (in lightblue). The superimposed curves are the mean SE values. The lower figures present the relative differences in SE between the direct and combined estimators. Each green curve plots their relative differences in SD. Each grey curve plots the averaged correlations between two estimators

the plots. We see that the SE values are close to SD largely. Again, the combined estimators $\tilde{\beta}_{3j}$ have lower variations and hence are more precise than the commonly used direct estimators $\hat{\beta}_{3j}$.

To gain more insight into the comparison between $\tilde{\boldsymbol{\beta}}_3$ and $\hat{\boldsymbol{\beta}}_3$, we have made some further examination in terms of relative efficiency. For each of the 1000 simulation runs, we compute the relative difference in SE, $\left\{ SE(\hat{\beta}_{3j}) - SE(\tilde{\beta}_{3j}) \right\} / SE(\hat{\beta}_{3j})$. The boxplots of these relative differences are presented in the lower panels of Figure 1. We also compute the relative difference in SD, $\left\{ SD(\hat{\beta}_{3j}) - SD(\tilde{\beta}_{3j}) \right\} / SD(\hat{\beta}_{3j})$, and plot them as line-connected circles. One striking observation is that the combined estimator $\tilde{\beta}_{3j}$ constantly has a smaller SE (as well as SD) than the direct estimator $\hat{\beta}_{3j}$, as indicated by the positive relative differences. Even though it is known that the MLE from the direct model is asymptotically efficient, in other words, optimal as $n \to \infty$, the combined estimator seems to steadily offer improvement in finite samples. The relative difference ranges from slightly over 0% up to 15% in the present model settings, showing a decreasing pattern as the sample size increases. One main reason accounting for this pattern is that the direct estimator $\hat{\beta}_{3j}$ and the inverse estimator $\hat{\alpha}_{3j}$ become close or highly correlated when n is large, as shown (gray dotted lines) in the lower panels of Figure 1. In this case, the strength that can be borrowed from the inverse estimator $\hat{\alpha}_{3j}$ becomes weaker, which reduces the room for improvement with the combined estimator $\tilde{\beta}_{3j}$.

TABLE 2 Simulation results with observational data based on 1000 simulation runs

			Estimate (mean)			SD			ASE			Size/power		
ρ	n		I	II	III	I	II	III	I	II	III	I	II	III
0	150	β_{31}	−0.529	−0.482	−0.466	1.407	1.400	1.470	1.322	1.323	1.242	0.072	0.080	0.096
		β_{32}	1.064	1.021	1.028	1.348	1.345	1.307	1.320	1.320	1.202	0.130	0.122	0.161
		β_{33}	0.012	0.055	0.049	1.363	1.358	1.306	1.329	1.334	1.250	0.047	0.044	0.066
	900	β_{31}	−0.484	−0.469	−0.475	0.513	0.513	0.504	0.505	0.505	0.498	0.171	0.167	0.166
		β_{32}	1.005	0.987	0.991	0.494	0.498	0.491	0.504	0.504	0.494	0.515	0.507	0.523
		β_{33}	−0.032	−0.025	−0.024	0.519	0.522	0.507	0.508	0.509	0.502	0.055	0.057	0.056
0.5	150	β_{31}	−0.571	−0.536	−0.552	1.539	1.531	1.485	1.494	1.497	1.417	0.076	0.076	0.088
		β_{32}	1.144	1.127	1.119	1.746	1.721	1.668	1.665	1.665	1.559	0.114	0.109	0.128
		β_{33}	−0.104	−0.076	−0.072	1.537	1.551	1.495	1.505	1.508	1.426	0.040	0.053	0.062
	900	β_{31}	−0.506	−0.494	−0.506	0.570	0.576	0.567	0.572	0.573	0.566	0.135	0.134	0.141
		β_{32}	1.047	1.024	1.031	0.612	0.610	0.599	0.635	0.636	0.626	0.357	0.350	0.360
		β_{33}	−0.026	−0.013	−0.013	0.592	0.589	0.577	0.575	0.576	0.568	0.062	0.061	0.053

Note: For each data set, the regression coefficient $\boldsymbol{\beta}_3$ associated with moderation analysis are estimated with three methods: (I) the direct estimator $\widehat{\boldsymbol{\beta}}_3$ with model (11); (II) the inverse estimator $\widehat{\boldsymbol{\alpha}}_3$ with model (12); and (III) the combined estimator $\tilde{\boldsymbol{\beta}}_3$.

Additional simulation results with observational data are presented in Table 2 and Figure 2, which are analogous to those with experimental data in Table 1 and Figure 1, respectively. The general conclusions are largely similar. This confirms that the role-swapping mechanism and the properties of the estimators hold regardless of the data source. In the Supplementary Material, Tables I and II present the results in the unbalanced scenarios where both the prevalence rate and the propensity are approximately 80%. We have expanded the simulation study extensively and examined a wide variety of scenarios and quantities. These include adjusting the correlation ρ among covariates, examining estimation of $\text{cov}(\tilde{\boldsymbol{\beta}}_3, \tilde{\boldsymbol{\alpha}}_3)$ via the formula in Theorem 2, sensitivity analysis with respect to model misspecification. We have found that the results are consistent with our general findings. Both the direct and inverse methods are comparable while the combined estimator stands out with higher precision.

5 | ANALYSIS OF WART TREATMENT TRIAL DATA

To illustrate, we consider data collected from a randomized wart treatment trial.[21] Warts are a type of skin disorders caused by human papillomavirus (HPV). A wart is a small fleshy bump on the skin or mucous membrane. Cryotherapy and immunotherapy are among the most common treatments for warts. The traditional cryotherapy freezes warts with liquid nitrogen while the relatively new immunotherapy treats warts by using the patient's own immune system.

To compare, $n = 180$ patients were randomized in this study to be treated with either the cryotherapy (`cryo` = 1) or immunotherapy (`cryo` = 0) method, with 90 patients in either treatment group. It is known that neither cryotherapy nor immunotherapy can heal all patients. Thus moderation analysis is crucial for designing more effective customized treatments.

In this data set, the outcome variable `response` is binary with 1 indicating a positive response to the treatment and 0 otherwise. Table 3 shows summary statistics on outcome and treatment. Also included are six covariates: patient gender (`sex`), age (`age`), self-reported time in months before treatment (`time`), the number of warts (`nwarts`), an indicator of whether or not patient has mixed types of warts (`type`), and surface area in mm^2 of the warts (`area`). Fitting the direct logistic regression models with variable selection and regularization,[22] two important covariates, `age` and `type`, are identified as potential moderators.

Table 4 presents the fitting results from three modeling analyses, including the parameter estimates, the standard errors, the Wald z test, and the resultant P-values. In the first approach, a logistic regression model was fit on `response`, examining the interactions between the treatment variable `cryo` and the covariates. For illustrative purposes, we also

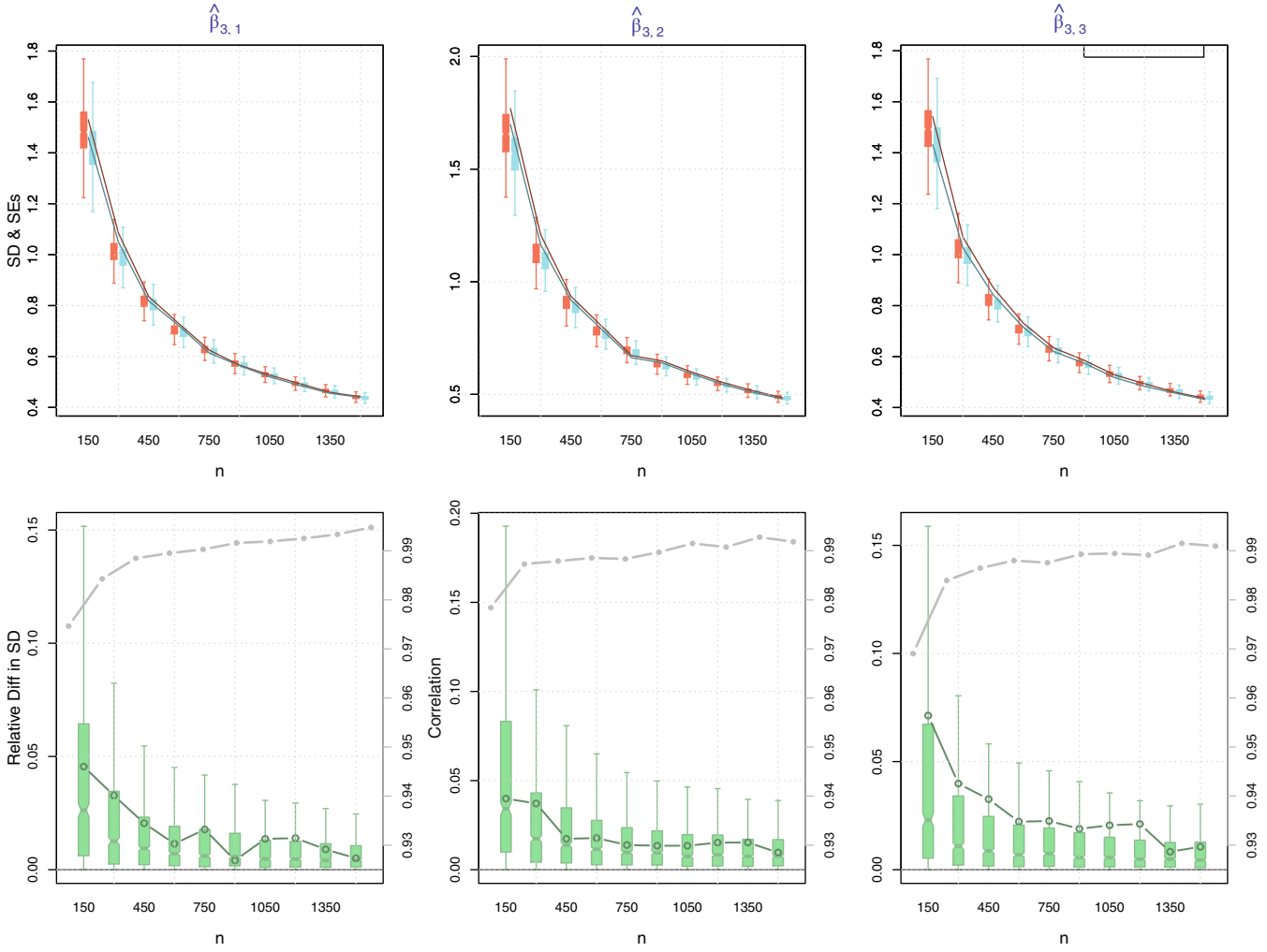

FIGURE 2 Simulation results with observational data: sample size $n = 150, 300, \ldots, 1500$, `nrun` $= 1000$, and $\rho = 0.5$. The upper figures present the boxplots of SE values for the direct estimators (in orange) and the combined estimators (in lightblue). The superimposed curves are the mean SE values. The lower figures present the relative differences in SE between the direct and combined estimators. Each green curve plots their relative differences in SD. Each grey curve plots the averaged correlations between two estimators

TABLE 3 Summary of `response` and `type` in wart data

	`type=0`		`type=1`	
	Immunotherapy	**Cryotherapy**	**Immunotherapy**	**Cryotherapy**
`response` = 0	16	19	3	23
`response` = 1	53	44	18	4

included an insignificant covariate `nwarts` in the model. In the second approach, the roles of `response` and `cryo` are swapped. These two approaches are labeled as the direct model and the inverse model in the table. The third analysis was based on the combined estimators, which are available for the coefficients associated with interaction terms only.

Comparing the direct model with the inverse model, we can see that their parameter estimates associated with the main effects are quite different, but the parameter estimates associated with the interaction terms are within the ballpark of each other. Both `age` and `type` show significant moderating effects, while `nwarts` does not. It is interesting to see that the SEs for the combined estimates are lower than those in the direct model and the inverse model. As a result, the interaction terms with `age` and `type` become more significant with much reduced P-values. On the other hand, `nwarts` remains insignificant.

TABLE 4 Analysis of wart trial data

	Direct model				Inverse model				Combined estimator			
	$\widehat{\beta}_j$	SE	z	P-value	$\widehat{\alpha}_j$	SE	z	P-value	$\tilde{\beta}_j$	SE	z	P-value
`Intercept`	2.594	0.898	2.889	0.004	0.966	0.967	0.999	0.318				
`cryo`	1.082	1.393	0.777	0.437	1.001	1.179	0.849	0.396				
`age`	−0.035	0.021	−1.649	0.099	−0.003	0.023	−0.131	0.896				
`type`	0.506	0.706	0.717	0.473	1.914	0.724	2.643	0.008				
`nwarts`	−0.035	0.064	−0.555	0.579	−0.120	0.089	−1.340	0.180				
`cryo:age`	−0.084	0.039	−2.137	0.033	−0.080	0.032	−2.475	0.013	−0.081	0.016	−5.080	0.000
`cryo:type`	−2.752	0.994	−2.769	0.006	−3.442	0.957	−3.598	0.000	−3.119	0.603	−5.176	0.000
`cryo:nwarts`	0.094	0.107	0.880	0.379	0.115	0.103	1.115	0.265	0.107	0.093	1.143	0.253

Note: The column z denotes the Wald z test statistics. All the P-values are two-sided.

The results of this moderation analysis can be meaningfully interpreted in terms of ROR. Take `type` for example. The odds ratio that compares cryotherapy versus immunotherapy in curing non-mixed types (`type` =0) of warts is $\exp(3.119) = 22.624$ of the odds ratio in treating mixed types (`type` =1) of warts, showing a great heterogeneity in the comparison. The following tabulated results provide a detailed look at such differential treatment effects without covariate adjustment, showing that immunotherapy does much better than cryotherapy in treating mixed types of warts while they have similar effects when treating non-mixed types.

An unadjusted estimate of ROR from the tabulated results can be obtained as $\{(16\times 44)/(53\times 19)\}/\{(3\times 4)/(23\times 18)\} = 24.119$, which is slightly different from the adjusted ROR estimate from the model. Similar way of interpreting can be carried out for `age`. To conclude, both `type` and `age` are important factors to consider in determining the optimal treatment for an individual patient.

6 | EXTENSION TO CATEGORICAL OUTCOMES AND TREATMENTS

All the foregoing results are readily extended to categorical outcomes and/or treatments that have two or more levels, though the notations are quite tedious. As seen from the proofs, the results in Lemma 1 are directly applicable to categorical outcomes and treatments with multiple levels. Suppose that $y_i \in \{0, 1, \ldots, K\}$ is categorical with $K+1$ levels while $t_i \in \{0, 1, \ldots, M\}$ has $M+1$ levels. Let $\mathbf{y}_i \in \mathbb{R}^{K\times 1}$ and $\mathbf{t}_i \in \mathbb{R}^{M\times 1}$ be the vectors of indicators induced by dummy variable coding with y_i and t_i, respectively, where category 0 is regarded as the reference or baseline level. We have the following result, which is analogous to those in Theorem 1.

Theorem 3. *Consider the following two multinomial logistic models with interaction terms*:

$$\log\frac{\Pr(y_i=k|\mathbf{t}_i,\mathbf{x}_i)}{\Pr(y_i=0|\mathbf{t}_i,\mathbf{x}_i)} = \beta_{k0} + \mathbf{t}_i^T\boldsymbol{\beta}_{k1} + \mathbf{x}_i^T\boldsymbol{\beta}_{k2} + (\mathbf{t}_i\otimes\mathbf{x}_i)^T\boldsymbol{\beta}_{k3} \quad \text{for } k = 1, \ldots, K, \tag{22}$$

and

$$\log\frac{\Pr(t_i=m|\mathbf{y}_i,\mathbf{x}_i)}{\Pr(t_i=0|\mathbf{y}_i,\mathbf{x}_i)} = \alpha_{m0} + \mathbf{y}_i^T\boldsymbol{\alpha}_{m1} + \mathbf{x}_i^T\boldsymbol{\alpha}_{m2} + (\mathbf{y}_i\otimes\mathbf{x}_i)^T\boldsymbol{\alpha}_{m3} \quad \text{for } m = 1, \ldots, M, \tag{23}$$

where $\otimes$ *denotes the outer product for vectors or the Kronecker product operator for matrices. Hence* $\boldsymbol{\beta}_{k3} = \left(\beta_{k3mj}\right) \in \mathbf{R}^{Mp\times 1}$ *and* $\boldsymbol{\alpha}_{m3} = \left(\alpha_{m3kj}\right) \in \mathbf{R}^{Kp\times 1}$. *We must have*

$$\beta_{k3mj} = \alpha_{m3kj},$$

for $k = 1, \ldots, K$, $m = 1, \ldots, M$, *and* $j = 1, \ldots, p$.

For model (22), let $\mathbf{X}_{y(k)}$ denote the common design matrix shared by each of its K model equations; namely, $\mathbf{X}_{y(k)}$ is of dimension $n\times(1+M+p+Mp)$ with i-th row vector $\mathbf{x}_{y(i)} = \left(1, \mathbf{t}_i^T, \mathbf{x}_i^T, (\mathbf{t}_i\otimes\mathbf{x}_i)^T\right)^T$. The entire design matrix for model

(22) can be written as $\mathbf{X}_y = \mathbf{X}_{y(k)} \otimes \mathbf{I}_K$, which is of dimension $nK \times (1 + M + p + Mp)K$. Similarly, let $\mathbf{X}_{t(m)}$ denote the $n \times (1 + K + p + Kp)$ common design matrix for each of the M model equations in (23), with its ith row vector being $\mathbf{x}_{t(i)} = \left(1, \mathbf{y}_i^T, \mathbf{x}_i^T, (\mathbf{y}_i \otimes \mathbf{x}_i)^T\right)^T$. The entire design matrix for model (22) can be written as $\mathbf{X}_t = \mathbf{X}_{t(m)} \otimes \mathbf{I}_M$, which is of dimension $nM \times (1 + K + p + Kp)M$.

Let $\boldsymbol{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_n^T)^T$ be the $nK \times 1$ response vector and $\boldsymbol{\pi}_y = \left(\pi_{y(ik)}\right)$ denote its corresponding probability vector with components $\pi_{y(ik)} = \Pr(y_i = k|\mathbf{t}_i, \mathbf{x}_i)$. Similarly, let $\boldsymbol{t} = (\mathbf{t}_1^T, \mathbf{t}_2^T, \ldots, \mathbf{t}_n^T)^T$ be the $nM \times 1$ treatment vector and $\boldsymbol{\pi}_t = \left(\pi_{t(im)}\right)$ denote its corresponding probability vector with components $\pi_{t(im)} = \Pr(t_i = m|\mathbf{y}_i, \mathbf{x}_i)$.

Furthermore, let $\boldsymbol{\beta}_k = (\beta_{k0}, \boldsymbol{\beta}_{k1}^T, \boldsymbol{\beta}_{k2}^T, \boldsymbol{\beta}_{k3}^T)^T$ and $\boldsymbol{\alpha}_m = (\alpha_{m0}, \boldsymbol{\alpha}_{m1}^T, \boldsymbol{\alpha}_{m2}^T, \boldsymbol{\alpha}_{m3}^T)^T$. Denote $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_K^T)^T$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \ldots, \boldsymbol{\alpha}_M^T)^T$, which include all the regression parameters involved in model (22) and model (23), respectively. Let $\hat{\boldsymbol{\beta}}$ be the MLE of $\boldsymbol{\beta}$ estimated from model (22) and $\hat{\boldsymbol{\alpha}}$ be the MLE of $\boldsymbol{\alpha}$ estiamted from model (23). The joint asymptotic normality of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ is established by the following theorem.

Theorem 4. *Assume that the condition probabilities* $\Pr(y = k|t, \mathbf{x})$, $\Pr(t = m|y, \mathbf{x})$, $\Pr(y = k|\mathbf{x})$, $\Pr(t = m|\mathbf{x})$ *are all bounded away from 0 and 1 (a.e.). Further assume that* $E(\mathbf{x}\mathbf{x}^T) \succ 0$. *Then* $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{\boldsymbol{\alpha}}$ *asymptotically exist and are strongly consistent. Their asymptotic joint normality is given by the same form as in Theorem 2 except for the definitions of the terms* $\boldsymbol{X}_y$, $\boldsymbol{X}_t$, $\boldsymbol{W}_y$ *and* $\boldsymbol{W}_t$, *which are given by*

$$\boldsymbol{W}_y = \operatorname{diag}\left(\boldsymbol{W}_i^{(y)}\right) \in \mathbb{R}^{nK \times nK} \quad \text{and} \quad \boldsymbol{W}_t = \operatorname{diag}\left(\boldsymbol{W}_i^{(t)}\right) \in \mathbb{R}^{nM \times nM},$$

where

$$\boldsymbol{W}_i^{(y)} = \begin{bmatrix} \pi_{y(i1)}(1 - \pi_{y(i1)}) & -\pi_{y(i1)}\pi_{y(i2)} & \cdots & -\pi_{y(i1)}\pi_{y(iK)} \\ -\pi_{y(i2)}\pi_{y(i1)} & \pi_{y(i2)}(1 - \pi_{y(i2)}) & \cdots & -\pi_{y(i2)}\pi_{y(iK)} \\ \vdots & \vdots & & \vdots \\ -\pi_{y(iK)}\pi_{y(i1)} & -\pi_{y(iK)}\pi_{y(i2)} & \cdots & \pi_{y(iK)}(1 - \pi_{y(iK)}) \end{bmatrix} \in \mathbb{R}^{K \times K}$$

and

$$\boldsymbol{W}_i^{(t)} = \begin{bmatrix} \pi_{t(i1)}(1 - \pi_{t(i1)}) & -\pi_{t(i1)}\pi_{t(i2)} & \cdots & -\pi_{t(i1)}\pi_{t(iM)} \\ -\pi_{t(i2)}\pi_{t(i1)} & \pi_{t(i2)}(1 - \pi_{t(i2)}) & \cdots & -\pi_{t(i2)}\pi_{t(iM)} \\ \vdots & \vdots & & \vdots \\ -\pi_{t(iM)}\pi_{t(i1)} & -\pi_{t(iM)}\pi_{t(i2)} & \cdots & \pi_{t(iM)}(1 - \pi_{t(iM)}) \end{bmatrix} \in \mathbb{R}^{M \times M}.$$

In the above expression, $\boldsymbol{W}_y$ is a block diagonal matrix with n block matrices $\boldsymbol{W}_i^{(y)}$ for $i = 1, \ldots, n$ on the main diagonal, and similarly for $\boldsymbol{W}_t$. The asymptotic VCOV matrix $\boldsymbol{\Sigma}$ can be estimated by replacing $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ with their MLEs. For moderation analysis, the asymptotic joint normal distribution of $\left(\widehat{\boldsymbol{\beta}}_{k3}\right)_{k=1}^K$ and $\left(\widehat{\boldsymbol{\alpha}}_{m3}\right)_{m=1}^M$ can be established by extracting components. Their combined estimators, as well as the associated SEs, can be obtained in the same way as in Section 3.

We have designed simulation studies to verify the findings and compare the three estimators with categorical outcomes and treatments. The results are deferred to the Supplementary Material (see Section S.3) due to page limit. An analysis of data collected from an adolescent placement study is also provided. This example illustrates an interesting scenario that involves a categorical outcome and a binary treatment. By switching the roles, it turns out that moderation analysis in the direct multinomial logistic model can be equivalently conducted through the corresponding inverse binary logistic model. Since the implementation of multinomial logistic regression related methods is not as widely available as that for binary logistic regression, working with the inverse model brings convenience.

7 | DISCUSSION

An interesting observation is made concerning moderation analysis with categorical outcomes. We have shown that, the role swap between the outcome variable and the treatment variable does not alter the ratio of odds ratios (ROR)

as a heterogeneity measure of treatment effects, which corresponds to the regression coefficients associated with the interaction terms in a logistic regression model. The conclusion holds regardless of whether data are experimental or observational. By role swap, we mean that the direct model and the inverse model have the same set of covariates and differ only in the outcome and treatment roles, though more flexible functional forms are allowed for the additive terms in the two models. To select variables, either the direct model or the inverse model can be used, depending on which way is more convenient. On this basis, we have developed a combined estimator of ROR with higher precision. Our numerical studies show that this strategy works well under a variety of model configurations and misspecifications. The results from our numerical studies suggest that the use of this combined estimator could be advocated for testing heterogeneous treatment effects in practical moderation analysis.

In terms of limitations of our approach, we have used both the direct and inverse models and assumed that the two logistic models with linear interaction terms both approximately hold well. We would like to comment that fitting both models can be seen in applications where the causal association between the response and the treatment is mutual and hence more correlational. Take the association study between obesity and depression for example. Obesity is associated with an increased risk of developing depression over time and, at the same time, depressed people tend to have an increased risk of obesity.[23] There is a mixed literature in which either obesity or depression is considered as the outcome and the other as the exposure or treatment. Both obesity and depression are often defined as binary variables and logistic regression is a natural choice for modeling them. As a result, both the direct model and the inverse model can be found in the literature; see Luppino et al[24] and references therein. In addition, our approach essentially requires that the linear interaction terms $t \cdot \boldsymbol{x}^T\boldsymbol{\beta}_3$ and $t \cdot \boldsymbol{x}^T\boldsymbol{\alpha}_3$ approximate the true moderation effects well in both models. In reality, it is possible that linear approximation performs poorly when the true models have strong curvilinear interactions. One promising remedial measure is to construct doubly robust[25,26] estimators that remain valid when either the direct or inverse model is misspecified. Such work has been done for the case-only analysis.[27]

Another issue is that we have focused on evaluation of the heterogeneity in treatment effects and estimation of the corresponding ratio of odds ratios. When the treatment effects are found significantly heterogeneous, the next step in common practice is subgroup analysis, in which data are first stratified according to one or few important effect-moderators and the treatment effect is assessed within each stratum by confidence intervals. While the combined estimator can be useful in tree-structured post-hoc subgroup identification,[28,29] our present results are not directly applicable for interval estimation of the treatment effect within each subgroup, which entails the asymptotic covariance between $\widehat{\boldsymbol{\beta}}_{(3)}$ and $\widetilde{\boldsymbol{\beta}}_3$. One possible solution that warrants future research is to jointly estimate the direct and inverse models where we set $\boldsymbol{\beta}_3 = \boldsymbol{\alpha}_3$ explicitly. The independent estimating equation (IEE)[30] method may be used for the joint model estimation. Another way of estimating the covariance is bootstrapping.

Other issues include, but not limited to, variable selection in both models, robustness against model misspecification, and model fitting difficulties in case of small sample sizes or complete separation. Regularization[31] has been used to select moderators in binary logistic regression models.[32] It is of future research interest to investigate how our results extend to regularized logistic models and how moderator selection with the direct model compares to that with the inverse model. We have seen that the proposed method shows considerable robustness against model misspecification under certain scenarios. Nevertheless, a more extensive study may be conducted to further investigate this issue under broader circumstances. For example, if the direct model and the inverse model are developed and selected separately, would the moderating effect estimates agree well? To handle complete separation, Firth's[33] regularization approach is commonly used. Investigating whether our results extend well to Firth's estimators of moderating effects can be another future research avenue.

The proposed methods can be useful and extended in other scenarios and there are remaining issues that may warrant future research. As one immediate potential application, consider the drug-drug interaction[34-36] problem which occurs when one drug or treatment modifies the efficacy of another. In this case with two treatments variables, two inverse models can be formed by regressing each treatment variable and hence lead to two inverse estimators of the drug-drug interaction effect. Plus the direct estimator, all three estimators may be combined into one with the general formula in Proposition 2 to facilitate a refined assessment of the drug-drug interaction. The mechanism of "role swapping" is particularly useful in other scenarios where a direct study of moderation is inconvenient owing to modeling complexity, numerical difficulty, or unavailability of implementation. The study of gene-environment interactions[37] presents a scenario with multiple treatments which are important genetic biomarkers. Common approaches are mostly hypothesis testing based. The main challenge stems from multiplicity of inferences. If we swap the roles and treat the genetic variables as clustered binary outcomes, a generalized linear mixed model[38] (GLMM) may facilitate an overall interaction

test for each environmental variable conveniently. In addition, our approach is applicable to retrospective case-control studies,[39] where estimates of OR and ROR remain valid, and its use in this regard may be further explored.

ACKNOWLEDGMENTS

The authors wish to thank the editor (Professor Joel Greenhouse), the associate editor, and two anonymous referees for their helpful comments and suggestions, which have greatly improved a preliminary version of the article.

DATA AVAILABILITY STATEMENT

The wart treatment data set[21,40] and the R code[19] for implementing the proposed method, as well as some raw simulation results, are available at GitHub (https://github.com/xgsu/rModAna).

ORCID

Xiaogang Su https://orcid.org/0000-0002-9642-9412
Youngjoo Cho https://orcid.org/0000-0001-5667-5654
Lei Liu https://orcid.org/0000-0003-1844-338X

REFERENCES

1. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med*. 2015;372(9):793-795.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
3. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. New York, NY: John Wiley & Sons; 1989.
4. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med*. 1994;13(2):153-162.
5. Vittinghoff E, Bauer DC. Case-Only Analysis of Treatment–Covariate Interactions in Clinical Trials. *Biometrics*. 2006;62(3):769-776.
6. Dai JY, Li SS, Gilbert PB. Case-only method for cause-specific hazards models with application to assessing differential vaccine efficacy by viral and host genetics. *Biostatistics*. 2014;15(1):196-203.
7. Dai JY, Liang CJ, LeBlanc M, Prentice RL, Janes H. Case-only approach to identifying markers predicting treatment effects on the relative risk scale. *Biometrics*. 2018;74(2):753-763.
8. Dai JY, LeBlanc M. Case-only trees and random forests for exploring genotype-specific treatment effects in randomized clinical trials with dichotomous end points. *J Royal Stat Soc Ser C-Appl Stat*. 2019;68(5):1371-1391.
9. Arnold BC, Press SJ. Compatible conditional distributions. *J Am Stat Assoc*. 1989;84(405):152-156.
10. Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954;10(1):101-129.
11. Efron B. The efficiency of logistic regression compared to normal discriminant analysis. *J Am Stat Assoc*. 1975;70(352):892-898.
12. White H. *Estimation, inference and specification analysis*. Cambridge University Press; 1994.
13. van der Vaart AW. *Asymptotic Statistics*. Cambridge University Press; 2000.
14. Fahrmeir L, Kaufmann H. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann Stat*. 1985;13(1):342-368.
15. Cox DR, Reid N. Parameter orthogonality and approximate conditional inference. *J Royal Stat Soc Ser B (Methodol)*. 1987;49(1):1-18.
16. Zeileis A. Object-oriented computation of sandwich estimators. *J Stat Softw*. 2006;16(1):1-16.
17. Graybill FA, Deal RB. Combining unbiased estimators. *Biometrics*. 1959;15(4):543-550.
18. Samuel-Cahn E. Combining unbiased estimators. *Am Stat*. 1994;48(1):34-36.
19. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2022.
20. Falk M. A simple approach to the generation of uniformly distributed random variables with prescribed correlations. *Commun Stat Simul Comput*. 1999;28(3):785-791.
21. Khozeimeh F, Alizadehsani R, Roshanzamir M, Khosravi A, Layegh P, Nahavandi S. An expert system for selecting wart treatment method. *Comput Biol Med*. 2017;81:167-175.
22. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22.
23. Blasco BV, Garcia-Jimenez J, Bodoano I, Gutierrez-Rojas L. Obesity and depression: its prevalence and influence as a prognostic factor: a systematic review. *Psychiatry Investig*. 2020;17(8):715-724.
24. Luppino FS, Wit LM, Bouvy PF, et al. Overweight, obesity, and depression: A systematic review and meta-analysis of longitudinal studies. *Arch Gen Psychiatry*. 2010;67(3):220-229.
25. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *J Am Stat Assoc*. 1995;90(429):122-129.
26. Kang JDY, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22(4):523-539.
27. Tchetgen Tchetgen EJ, Robins J. The semiparametric case-only estimator. *Biometrics*. 2010;66(4):1138-1144.
28. Su X, Tsai C-L, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. *J Mach Learn Res*. 2009;10(2):141-158.

29. Su X, Kang J, Fan J, Levine RA, Yan X. Facilitating score and causal inference trees for large observational studies. *J Mach Learn Res*. 2012;13:2955-2994.
30. Fitzmaurice GM. A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*. 1995;51(1):309-317.
31. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc Ser B (Methodol)*. 1996;58(1):267-288.
32. Lim M, Hastie T. Learning interactions via hierarchical group-lasso regularization. *J Comput Graph Stat*. 2015;24(3):627-654.
33. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27-38.
34. Goldberg RM, Mabee J, Chan L, Wong S. Drug-drug and drug-disease interactions in the ED: analysis of a high-risk population. *Am J Emerg Med*. 1996;14(5):447-450.
35. Solberg LI, Hurley JS, Roberts MH, et al. Measuring patient safety in ambulatory care: potential for identifying medical group drug-drug interaction rates using claims data. *Am J Manag Care*. 2004;10(11):753-759.
36. Polasek TM, Lin FP, Miners JO, Doogue MP. Perpetrators of pharmacokinetic drug-drug interactions arising from altered cytochrome P450 activity: a criteria-based assessment. *Br J Clin Pharmacol*. 2011;71(5):727-736.
37. Ottman R. Gene–environment interaction: definitions and study design. *Prev Med*. 1996;25(6):764-770.
38. Stroup WW. *Generalized Linear Mixed Models: Modern Concepts: Methods and Applications*. Boca Raton, FL: CRC Press; 2012.
39. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979;66(3):403-411.
40. Khozeimeh F, Azad FJ, Oskouei YM, et al. Intralesional immunotherapy compared to cryotherapy in the treatment of warts. *Int J Dermatol*. 2017;56(4):474-478.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Su X, Cho Y, Ni L, Liu L, Dusseldorp E. Refined moderation analysis with categorical outcomes in precision medicine. *Statistics in Medicine*. 2023;42(4):470-486. doi: 10.1002/sim.9627