



Statistical Methods for Selective Biomarker Testing

A. Adam Ding¹ · Natalie DelRocco² · Samuel S. Wu²

Received: 21 March 2023 / Revised: 7 August 2023 / Accepted: 6 October 2023
© The Author(s) 2024

Abstract

Biomarkers are critically important tools in modern clinical diagnosis, prognosis, and classification/prediction. However, there are fiscal and analytical barriers to biomarker research. *Selective Genotyping* is an approach to increasing study power and efficiency where individuals with the most extreme phenotype (response) are chosen for genotyping (exposure) in order to maximize the information in the sample. In this article, we describe an analogous procedure in the biomarker testing landscape where both response and biomarker (exposure) are continuous. We propose an intuitive reverse-regression least squares estimator for the parameters relating biomarker value to response. An expression for robust standard error and corresponding confidence interval are derived. A simulation study is used to demonstrate that this method is unbiased and efficient relative to estimates from random sampling when the joint normal distribution assumption is met, and to compare the estimator to an alternative under a related sampling design. We illustrate application of proposed methods on data from a chronic pain clinical trial.

Keywords Extreme sampling · Outcome-dependent sampling · Selective genotyping · Simple linear regression

✉ A. Adam Ding
a.ding@northeastern.edu

Natalie DelRocco
ndelrocco95@ufl.edu

Samuel S. Wu
samwu@biostat.ufl.edu

¹ Department of Mathematics, Northeastern University, Boston, MA, USA

² Department of Biostatistics, University of Florida, Gainesville, Florida, USA

1 Background

A biological marker, or biomarker, is an objective measurement which indicates a biologic process or response [1]. This umbrella definition captures a range of measurements that may be representative of a disease course, from simple indicators like blood pressure to complex laboratory tests [2]. In this paper, we focus specifically on physical markers which are detectable via serum assays. In the past 20 years, the explosive use of biomarkers in medical research has coined the “biomarker revolution” [3]. Clinically relevant biomarkers can provide information on both disease mechanisms and subsequent outcomes [4]. In practice, a biomarker can serve as both a risk indicator and a surrogate for disease status. As such, the biomarker is a critically important tool in modern clinical diagnosis, prognosis, and classification/prediction.

Despite the clinical utility and popularity of biomarkers and continual advancements in collection technology, there remain fiscal and analytical barriers to biomarker research. The cost of conducting biomarker assays for a sufficiently powered study is a major limitation. For example, a luminex assay with 39 sample slots and the capacity to detect up to 20 biomarkers can cost between \$300 and \$600 before fees for consultation, lab materials, and labor [5]. This means that collecting biomarker samples for each person in a study of 800 participants could cost more than \$12,600 when serum samples for each individual are collected at a single time point. Resource limitations have thus inspired the development of cost-effective experimental designs and corresponding statistical methodology [3, 6].

In genetics literature, one such approach concentrates sampling to the most informative observation units [7]. Selective genotyping [8] traditionally entails sampling individuals for genotyping based on extreme phenotypic values where genotype (exposure) is discrete (presence/absence; aa/Aa/AA) and phenotype (response) is continuous, most often assuming an ANOVA-style model [9]. From a statistical perspective, this is distinct from, but perhaps inspired by, the information perspective [10] for which the optimal design in regression analysis (i.e. sample) is the one that minimizes the variance of coefficient estimates [11].

Under selective genotyping, the response no longer follows a normal distribution and missing data in the middle of the phenotypic distribution must be accounted for [12]. Appropriate maximum likelihood methods have been developed for inference under such designs for QTL studies [7, 9], which have been shown to increase statistical power relative to random sampling using a fraction of the original sample [7, 13, 14]. However, the developed methods are highly specific to the field of genetics, modeling the exposure as discrete and accounting for elements such as backcross in models. Additionally, in genotyping studies it is advantageous to take a multistage approach wherein promising genetic markers are identified early out of a pool of candidates to meet study constraints [15, 16].

In this paper we study a selective biomarker-testing scheme, where similar to selective genotyping, individuals with extreme response values are selected for biomarker-testing. In contrast to typing the discrete genotypes in the selective

genotyping, here the individuals' continuous biomarker is measured. Selective genotyping and our selective biomarker-testing are both special cases of Outcome-Dependent Sampling (ODS) [10]. ODS (also known as response-dependent sampling) is a form of biased sampling, which has been studied in the context of likelihood-based estimation [17]. Zhou et al. [18] formally defined an ODS design for a linearly related continuous exposure and outcome where all response values are observed and covariate values are observed, as (1) a simple random sample (SRS) from the full available cohort and (2) additional SRS's from regions of the response that are of particular interest. They proposed a semiparametric empirical likelihood estimation method for this ODS design which was shown to increase efficiency relative to simple random sampling [18]. Weaver and Zhou [19] extended this design to incorporate all available information from the full sample including those with unknown exposure values taking an estimated likelihood approach. Zhou's original estimator has been expanded to accommodate different functional forms [20], auxiliary covariate information [21–23], multi-stage designs [23–25], mixed effects models [25], and more recently survival [26] and longitudinal endpoints [6, 27, 28].

Here we propose statistical analysis methods to analyze selective biomarker-testing data utilizing regression estimations available from standard statistical software. Notice that our selective biomarker-testing scheme can be considered as Extreme Outcome-Dependent Sampling (EODS) since we only sample individuals with extreme response values without the SRS from the full cohort (thus no biomarker-testing is conducted for any individual in the mid range of response values). The benefit, relative to the cost, of the incorporation of the primary SRS in traditional ODS designs is unknown. Selective genotyping literature suggests that there is no information to be gained by genotyping individuals outside the tails of the response variable distribution [7]. Also, while the ODS estimators in above literature are efficient, unbiased, and flexible under their statistical assumptions, the complexity of the semiparametric likelihood-based approach is a key barrier to their widespread use in practice. The least squares approach to estimation in regression analysis is standard, and therefore most accessible to researchers conducting biomarker-testing studies. However, least squares has been shown to be most susceptible to the bias induced by extreme sampling based designs [11]. To solve this issue, we propose a reverse-regression least squares estimation method for EODS designs with jointly normal distributed biomarker and response.

The organization of this article is as follows: In Sect. 2 we describe a general methodology for biomarker studies, including effect estimation, power/sample size considerations, and model checking methods as well as the BISP2 clinical trial dataset which motivated this research. In Sect. 3 we present numerical results of our estimator. We first study the finite sample properties of our estimator via simulation, then apply our estimator to the BISP2 Biomarker Study. We conclude with a discussion of strengths and limitations in Sect. 4 and a brief conclusion in Sect. 5.

2 Methods

In a clinical study, we are interested in the relationship between a response variable Y and a biomarker variable X . If both variables are measured on all subjects enrolled in the trial, we have a full data set consisting of $(X_1, Y_1), \dots, (X_{n_F}, Y_{n_F})$. Here n_F denotes the size of the full data set. For the EODS biomarker analysis, we observe the response variable for the full data set, Y_1, \dots, Y_{n_F} , but only observe the biomarker variable X on a subset corresponding to extreme values of Y . Say, we select γ proportion of the full data set and measure the biomarker X on the selected subset of size $n_S = \lceil \gamma n_F \rceil$, where $\lceil d \rceil$ denotes the closest integer to d . Without loss of generality, we can denote the first n_S subjects as the selected ones. That is, in the selective biomarker analysis, we observe $(X_1, Y_1), \dots, (X_{n_S}, Y_{n_S})$ and $Y_{n_S+1}, \dots, Y_{n_F}$, while Y_1, \dots, Y_{n_S} corresponds to the top $\gamma/2$ proportion and the bottom $\gamma/2$ proportion of Y_1, \dots, Y_{n_F} .

The aim of the statistical regression analysis is to study the dependence of response variable Y on the biomarker X :

$$E(Y|X) = \alpha_Y + \beta_Y X, \tag{1}$$

where $E(Y|X)$ denotes the conditional expectation of Y given X . When X and Y jointly follow a bivariate normal distribution, the full data set satisfies the standard regression model assumption:

$$Y_i = \alpha_Y + \beta_Y X_i + \varepsilon_{Y,i}, \quad \varepsilon_{Y,i} \sim N(0, \sigma_{\varepsilon_Y}^2), \tag{2}$$

for $i = 1, \dots, n_F$.

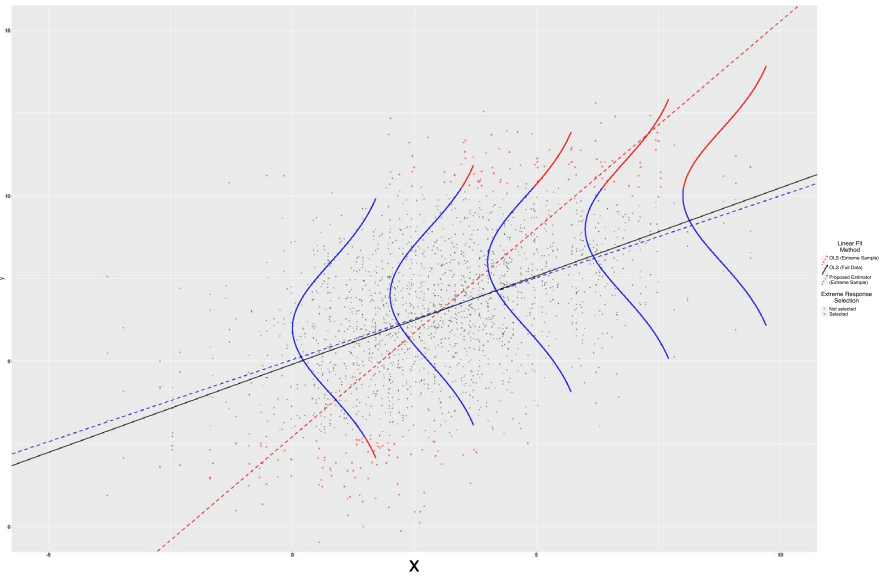
However, for the EODS biomarker analysis, naively conducting linear regression analysis directly on the subset $(X_1, Y_1), \dots, (X_{n_S}, Y_{n_S})$ does not work. The reason is that, the regression model assumption (2) does not hold on the selectively sampled subset because the response follows a truncated distribution due to the selection of extreme values for Y as shown by the red part of the curves in Fig. 1a. The ordinary least squares (OLS) fit on the selectively sampled subset will lead to highly biased slope estimation as shown by the dotted red line. Therefore, specific analysis methods are needed for the selective biomarker analysis. In this paper, we propose a parametric analysis method which utilizes standard regression formulas on reverse-regression: regressing X on Y for the selectively sampled subset. The main insight for the proposal is that, when X and Y jointly follow a bivariate normal distribution, the standard regression model assumption holds for the reverse-regression on both full data set and the selectively sampled subset:

$$X_i = \alpha_X + \beta_X Y_i + \varepsilon_{X,i}, \quad \varepsilon_{X,i} \sim N(0, \sigma_{\varepsilon_X}^2). \tag{3}$$

To see why the reverse-regression model (3) holds also on selectively sampled subset, we notice that, for jointly normally distributed X and Y , the distribution of X conditional on any $Y = y$ value is normal with

$$E(X|Y = y) = \alpha_X + \beta_X y, \quad \text{Var}(X|Y = y) = \sigma_{\varepsilon_X}^2,$$

(a) Truncation of the conditional distribution of Y given X under extreme response selection with the linear model approach



(b) Preservation of the conditional distribution of X given Y under extreme response selection

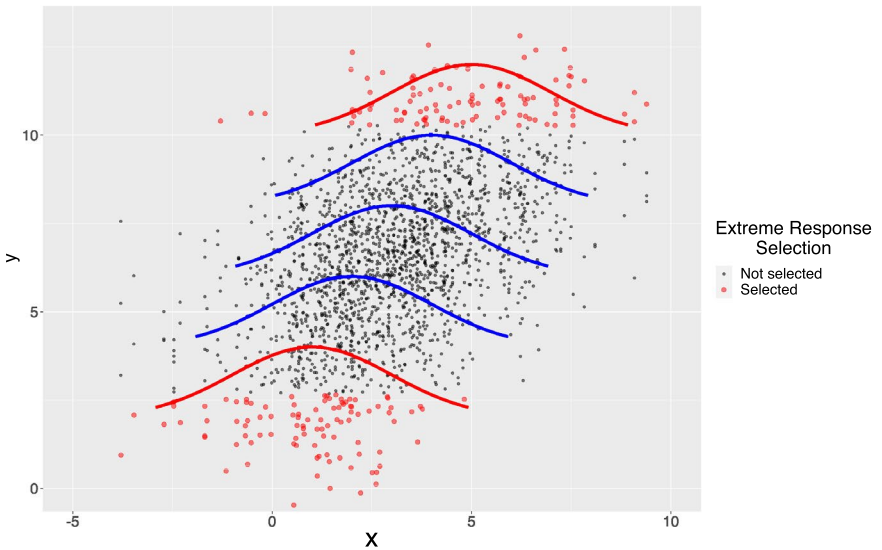


Fig. 1 Illustration of challenge posed to linear regression under extreme sampling

where σ_{ϵ_X} is a constant not affected by the y values. The relationship between the reverse-regression parameters $(\alpha_X, \beta_X, \sigma_{\epsilon_X})$ and the bivariate normal distribution parameters $(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$ can be found in derivations in Appendix A. Therefore, for any sampling scheme that only uses the value of random variable Y , conditional on selected $(Y_1 = y_1, \dots, Y_{n_S} = y_{n_S})$, the corresponding X_i follows the normal distribution satisfying equation (3). That is, the reverse-regression model (3) holds for any selective sampling scheme based only on response values, of which the EODS is a special case. Figure 1b illustrates this phenomenon for bivariate normal distribution: conditional on selected Y values, X is still normally distributed with common variance and a linearly changing mean.

Armed with this insight, we can conduct the reverse-regression $E(X|Y) = \alpha_X + \beta_X Y$ using standard statistical software on the selectively sampled subset. Then we convert the reverse-regression fit results into regression inferences, with a little help of the additional response variable observations $Y_{n_S+1}, \dots, Y_{n_F}$. In Appendix A, we provide detailed mathematical derivation of formulas converting the reverse-regression coefficients to regression coefficients. We describe the application of these conversion formulas for analyzing the EODS biomarker data in the rest of this section.

2.1 Hypothesis Test

One main focus of the biomarker analysis is to test whether a biomarker X affects the response variable Y . Statistically, this is usually done by the linear regression hypothesis testing

$$H_0 : \beta_Y = 0 \quad \text{versus} \quad H_A : \beta_Y \neq 0. \tag{4}$$

Mathematically, the null hypothesis $H_0 : \beta_Y = 0$ is equivalent to the reverse-regression null hypothesis $H_0 : \beta_X = 0$. Thus we can simply carry out the reverse-regression hypothesis test on the selectively sampled subset for

$$H_0 : \beta_X = 0 \quad \text{versus} \quad H_A : \beta_X \neq 0. \tag{5}$$

The p-value for the test of (5) is also valid for testing (4).

2.2 Effect Estimation

The effect of the biomarker X on response variable Y is measured by the slope β_Y in the regression function. Naively fitting the regression equation (1) on the selectively sampled subset would result in an overestimation of the biomarker effect. As illustrated in Fig. 1a, the conditional (truncated) distribution of Y changes with the given X values, leading to OLS estimate (shown as the dotted red line) with much bigger slope magnitude than the true β_Y value when $\beta_Y > 0$. For consistent estimation of β_Y , we use instead the parameter estimates $\hat{\beta}_X$ and $\hat{\sigma}_{\epsilon_X}^2$ from fitting the reverse-regression (3) on the selectively sampled subset. These estimates can be obtained from conducting the reverse-regression using any standard software. Furthermore, from the

full response variable observations of Y_1, \dots, Y_{n_F} , we have the sample mean $\tilde{\mu}_Y = \frac{1}{n_F} \sum_{i=1}^{n_F} Y_i$ and the sample variance $\tilde{\sigma}_Y^2 = \frac{1}{n_F-1} \sum_{i=1}^{n_F} (Y_i - \tilde{\mu}_Y)^2$.

Using equation (18) in the Appendix, we have a point estimator for β_Y as

$$\hat{\beta}_Y = \frac{1}{\frac{\hat{\sigma}_{\epsilon_X}^2}{\hat{\sigma}_Y^2} + \hat{\beta}_X^2} \hat{\beta}_X. \tag{6}$$

Then the standard error for $\hat{\beta}_Y$ can be calculated as in equation (23),

$$s.e.\{\hat{\beta}_Y\} = \sqrt{\frac{\left(\frac{\hat{\sigma}_{\epsilon_X}^2}{\hat{\sigma}_Y^2} - \hat{\beta}_X^2\right)^2 (s.e.\{\hat{\beta}_X\})^2 + \left(\frac{2\hat{\beta}_X^2 \hat{\sigma}_{\epsilon_X}^4}{\hat{\sigma}_Y^4}\right) \left(\frac{1}{n_X-2} + \frac{1}{n_F-1}\right)}{\left(\frac{\hat{\sigma}_{\epsilon_X}^2}{\hat{\sigma}_Y^2} + \hat{\beta}_X^2\right)^4}},$$

whose detailed derivation is provided in the Appendix. Therefore, the $(1 - \alpha)$ confidence interval for β_X can be calculated as

$$\hat{\beta}_Y \pm z_{\alpha/2} s.e.\{\hat{\beta}_Y\}.$$

Here $z_{\alpha/2}$ is the $\alpha/2$ -upper quantile for the standard normal distribution.

2.3 Power/Sample Size Calculation

As derived in the Appendix, equation (27) gives the power formula of the hypothesis test in section :

$$power = P(W > FQ_{\alpha, df_1, df_2}), \tag{7}$$

where W denotes a random variable following a non-central F-distribution with degrees of freedoms of $df_1 = 1$ and $df_2 = \gamma n_F - 2$ and noncentral parameter $n_F f^2 2 \int_{z_{\gamma/2}}^{\infty} x^2 \phi(x) dx$, and FQ_{α, df_1, df_2} denotes the α -upper quantile of a central F-distribution with degrees of freedoms of df_1 and df_2 . Here z_{α} denotes the α -upper quantile of a standard normal distribution $N(0, 1)$ whose density is denoted as $\phi(x)$. And f is the Cohen's effect size defined as

$$f^2 = \frac{R^2}{1 - R^2} = \frac{\rho^2}{1 - \rho^2}, \tag{8}$$

where R^2 is the proportion of variation in the data explained by the regression equation.

Based on this, we can choose the proportion γ . We illustrate the sample size calculation with a simple example here. Assume that we have a full data set of sample size $n_F = 200$, and we wish to detect a Cohen's effect size $f = 0.3$ with 90% power at the significance level $\alpha = 0.05$. When we select 10% individuals in EODS

for biomarker-testing (i.e., $n_S = 200 * 0.1 = 20$, selecting 10 persons with largest Y values and 10 persons with smallest Y values), (7) gives $power = 0.9150$. When we select 9% individuals in EODS for biomarker-testing (i.e., selecting 9 persons with largest Y values and 9 persons with smallest Y values), (7) gives $power = 0.8985$. Thus 20 individuals will be biomarker-tested to achieve the design goal of 90% power.

In contrast, the standard t-test in simple linear regression for testing (4) has $power = 0.8983$ when $n = 118$ and $power = 0.9007$ when $n = 119$ as given by equation (24). Thus a clinical trial without the EODS will require 119 individuals with both X and Y measurements, needing an almost sixfold increase in cost of biomarker-testing for X than the EODS method.

2.4 Model Checking

Our methodology is based on the assumption that X and Y are jointly normally distributed. Mathematically, that is equivalent to the following two assumptions holding simultaneously: (A) the response variable Y is normally distributed and (B) conditional on Y , the biomarker variable X is normally distributed. We now consider how to check these two assumptions on observed data.

Assumption (A) can be checked with standard methods such as the normal probability plot on the fully observed Y_1, \dots, Y_{n_F} . For assumption (B), we do not observe X for the whole range of Y . But we can still check whether (B) holds for the selected Y values by applying standard model checking methods on the reverse-regression, e.g., the normal probability plot of the reverse-regression residuals.

2.5 Motivating Data Study Design

Biopsychosocial Influence on Shoulder Pain Phase I (BISP) was a single-center, pre-clinical “proof of concept” study of 190 adults to identify genetic and psychological characteristics related to chronic musculoskeletal pain [29]. Musculoskeletal pain is the general pain affecting the muscles, ligaments, tendons, or bones with chronic indicating that the pain is long-lasting or consistently recurring. In general, musculoskeletal pain is a large contributor to the \$635 billion yearly healthcare cost of chronic pain in the United States [29]. Though this makes chronic pain a high priority research area, there are limited accepted treatment models due to the complexity of disease etiology. Treatment components must be personalized on the basis of genetic, psychological, environmental, and social risk factors, which all contribute to the individual variation in how people experience chronic pain.

Specifically, BISP targeted chronic musculoskeletal pain affecting the shoulder region by comparing predictors of pain level among healthy individuals pre- and post-induced shoulder pain. The target population was healthy adults. Participants were followed over the course of five days. Baseline covariates and DNA samples were collected on the first day of the study, before inducing shoulder pain. An exercise-based protocol was then used to create controlled shoulder pain through inflammation and muscular fatigue [29]. Four follow-up visits were conducted every 24 h

post-injury induction. Shoulder impairment, genetic testing, and other covariates related to pain level defined a priori based on clinical expertise were measured at baseline and each follow-up visit. BISP identified multiple prognostic factors which were associated with increased shoulder pain, including promising genes which showed evidence of predicting shoulder impairment.

BISP Phase II (BISP2) was a single-center, randomized follow-up trial to BISP which aimed to test whether interventions personalized on the basis of genetic and psychological characteristics are effective for induced shoulder pain (ClinicalTrials.gov Identifier: NCT02620579) [30]. The two-factor factorial design randomized 261 individuals to either propranolol or placebo and psychological education or general education. Propranolol is a drug chosen to target Catechol-O-methyl-transferase (COMT), which metabolizes adrenal hormones and is associated with pain sensitivity. The psychological education was designed to target pain rumination, which is magnification of pain by focusing on the pain with a pessimistic attitude. Pre-randomization, shoulder injury was induced using the same protocol.

BISP2 participants provided daily report on pain intensity and disability over the 5-day onsite observation period. Pain level was measured using the Brief Pain Inventory (BPI) [31], which is an 11-point scale ranging from 0 (no pain) to 10 (worst pain intensity imaginable). Participants rated the intensity of current pain and pain intensity at its worst, best, and average over the past 24 h. Clinically relevant covariates and saliva samples for genetic testing were again taken at each follow up visit.

The BISP2 investigators targeted 14 genetic markers associated with pain in the study's exploratory biomarker analysis. These biomarkers were chosen a priori based on clinical knowledge of propensity to release pro-inflammatory cytokines. All genetic samples were saliva-based. The primary question of interest for the exploratory analysis was: which biomarkers are associated with recorded level of shoulder pain at 48-hours post-randomization? As the response, pain, is measured via the BPI to represent a continuum pain, linear regression with pain level as the response is a typical analysis approach. However, the investigators were limited by the budget and time constraints of the BISP2 clinical trial. It was infeasible to biomarker-test every individual in the study. Therefore, the investigators began by biomarker-testing 31 individuals in the high and low tails of worst pain experienced at 48-hours post shoulder pain induction.

3 Results

3.1 Simulation Studies

In this section, we use simulations to check the finite sample performance of our proposed Outcome-Dependent Extreme biomarker-testing (ODEB) estimator. All simulations were conducted using R Statistical Software Version 3.6.0 [32].

3.1.1 Simulation Comparison with Ordinary Least Squares

We compare our ODEB estimation with the ordinary least squares (OLS) estimation in various parameter settings here. Data were simulated according to the probability model described in Equation (2). Specifically, Y_i is generated such that

$$Y_i = 5 + \beta_Y X_i + \varepsilon_{Y,i}, \quad X \sim N(20, 5), \quad \varepsilon_{Y,i} \sim N(0, 5). \quad (9)$$

Individual data sets consisting of $(X_1, Y_1), \dots, (X_{n_F}, Y_{n_F})$ were randomly generated for each of $B = 20,000$ Monte Carlo simulations. We studied combinations of the following sets of parameter values: $n \in \{100, 200, 400, 800\}$, $\beta_Y \in \{0, 0.2, 0.26, 0.4, 0.8, 1\}$, $\gamma \in \{0.1, 0.2, 0.4\}$. For each iteration, both the extreme-outcome-dependent sampling and random sampling were conducted. β_Y was then estimated for each sampling method using both OLS ($\hat{\beta}_{Y,OLS}$) and ODEB ($\hat{\beta}_{Y,ODEB}$) estimation.

To illustrate the impact of the proportion of selected extreme subsample, Fig. 2 plots the ODEB estimation for various proportion selected from a fixed dataset where $n = 800$ and $\beta_Y = 0.4$. As the proportion sampled increases, $\hat{\beta}_{Y,ODEB}$ trends closer to the true parameter value, and the 95% confidence interval contains β_Y for all proportions sampled.

Theoretically, both OLS and our ODEB provide valid inferences under random sampling, but only ODEB estimation is consistent under extreme sampling. The simulation results confirmed this.

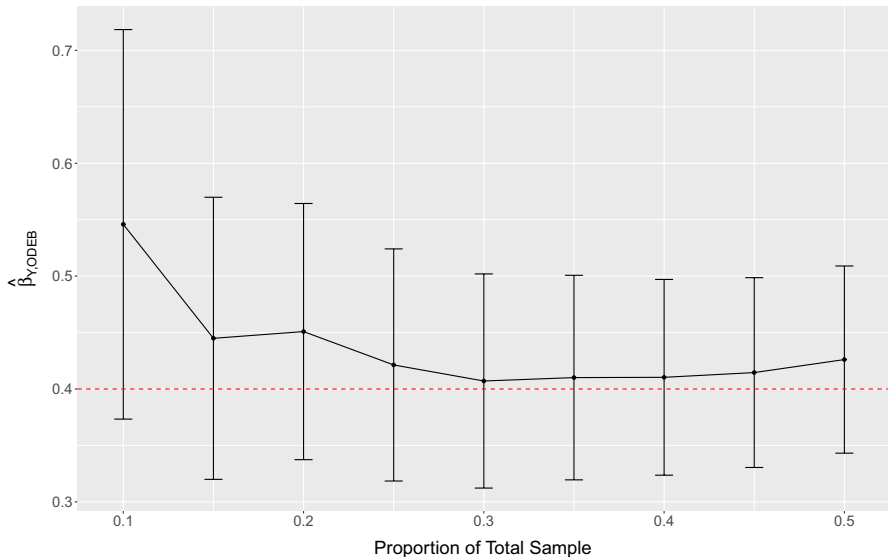


Fig. 2 ODEB point and interval estimation for $\hat{\beta}_Y$ under various proportions of the total study extreme sampled when $n = 800$ and $\beta_Y = 0.4$

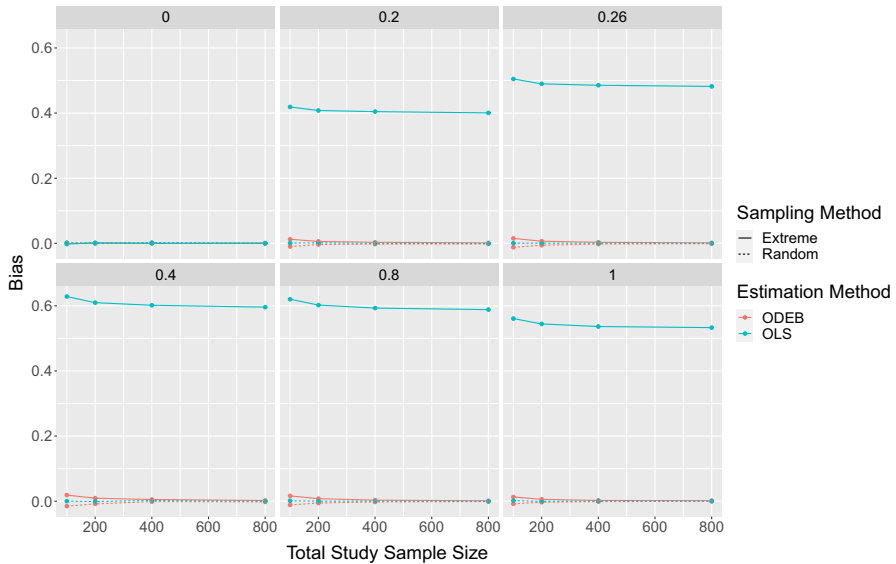


Fig. 3 Bias of $\hat{\beta}_Y$ under 20% of the total study sampled for various β_Y 's

Figure 3 displays the bias of $\hat{\beta}_Y$ when an extreme sample of 10% is taken from each tail of the response distribution (thus resulting in a total of 20% extreme sampling). When there is truly no effect of the biomarker X , both OLS and ODEB estimation are unbiased regardless of whether the sample is extreme or random. When there is a biomarker effect present, $\hat{\beta}_{Y,ODEB}$ under extreme or random sampling and $\hat{\beta}_{Y,OLS}$ under random sampling remain mostly unbiased, with only a small appreciable bias when $n = 100$ which dissipates as the sample size reaches $n = 400$. However, $\hat{\beta}_{Y,OLS}$ under extreme sampling is severely biased in the positive direction. The bias initially increases, then tapers, with the size of β_Y . For saving space, we omit the detailed numbers for simulations from other sampling proportions (10% and 40%) but summarize their bias patterns. The bias of $\hat{\beta}_{Y,OLS}$ under extreme sampling is exacerbated when $\gamma = 0.1$ (i.e., 5% per tail), with estimation bias exceeding twice the magnitude of β_Y . The bias of $\hat{\beta}_{Y,OLS}$ decreases slightly when $\gamma = 0.4$, with more moderate observations included, but the overall bias magnitude is still unacceptably large.

The root mean square error (RMSE) displays similar trends, as shown in Fig. 4. Under extreme sampling, $\hat{\beta}_{Y,ODEB}$ has the smallest RMSE of the four considered combinations across both study sample sizes and true β_Y 's. This reflects the efficient use of information by $\hat{\beta}_{Y,ODEB}$. In contrast, $\hat{\beta}_{Y,OLS}$ has the largest RMSE across study sample sizes and β_Y 's when applied to an extreme sample. Of note, under random sampling, $\hat{\beta}_{Y,OLS}$ and $\hat{\beta}_{Y,ODEB}$ perform similarly in terms of RMSE, but $\hat{\beta}_{Y,ODEB}$ still beats $\hat{\beta}_{Y,OLS}$ as β_Y becomes bigger.

In addition, $\hat{\beta}_{Y,ODEB}$ provides a practical increase in power to detect biomarker effect when combined with extreme sampling. First, for ODEB test under

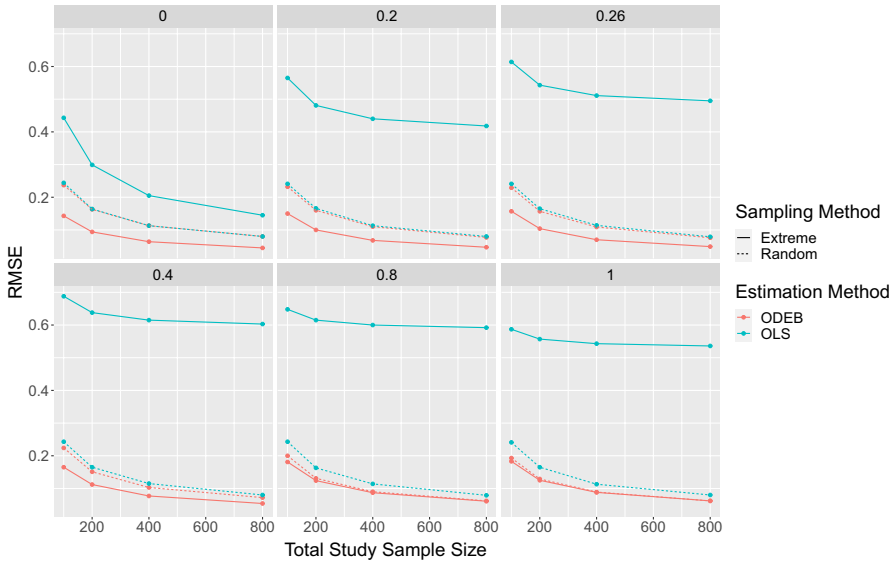


Fig. 4 Root mean square error (RMSE) of $\hat{\beta}_Y$ under 20% of the total study sampled for various β_Y 's

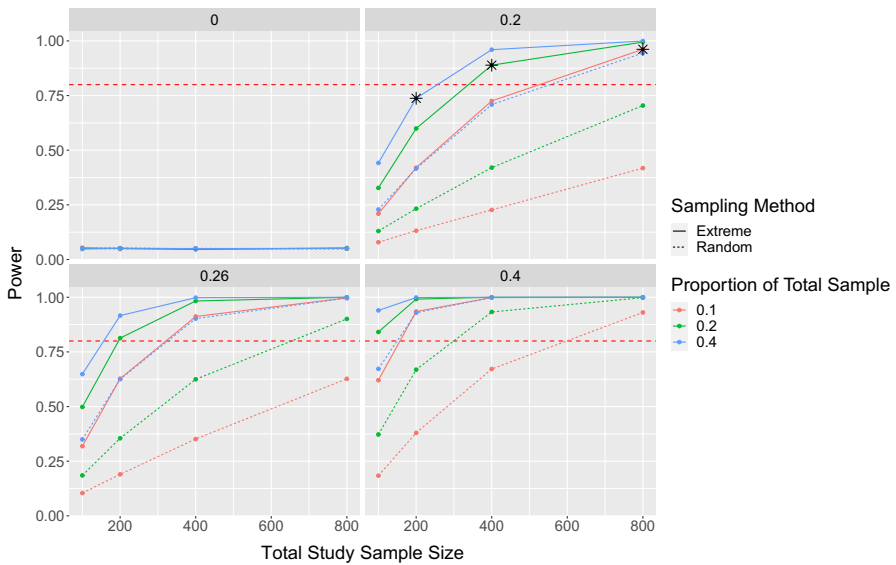


Fig. 5 Power to detect a nonzero effect of biomarker level on pain for various β_Y 's

all sampling methods, the type I error rate for the test of the null hypothesis $H_0 : \beta_Y = 0$ is controlled at the nominal level of 5% as illustrated in the upper-left panel of Fig. 5. When $\beta_Y \neq 0$, random sampling generally has less power to detect a biomarker effect than extreme sampling, with the dashed lines (random

sampling) falling below the solid lines (extreme sampling) in each panel. This highlights the utility of focusing on those observations with extreme response values. Taking $\beta_Y = 0.2$ as an illustrative example of this notion, the stars placed on the upper-right panel emphasize the increase in power afforded by successively more extreme samples. For the three starred cases, each sample contains 80 biomarker-tested observations but a more extreme sample from a larger population has the advantage in terms of power to detect the effect. A 96.2% power is achieved by the 10% extreme sample from 800 observations. In contrast, 20% and 40% extreme samples respectively from 400 and 200 observations results in 88.9% and 73.8% power only.

Figure 6 studies the performance of the 95% confidence interval for β_Y for OLS and ODEB. Only the OLS-based confidence interval under extreme sampling does not achieve correct nominal 95% coverage. Its poor coverage displayed is worsened both with larger total study sample size and larger size of β_Y . This is particularly severe when $\beta_Y = 1$, with coverage probabilities of effectively 0% for $n = 200, 400,$ and 800 . The OLS-based confidence interval under extreme sampling also displays the longest confidence interval length (Fig. 7). ODEB estimation affords the most precise confidence interval compared to OLS, particularly when applied to an extreme sample.

The conclusions of comparison between OLS and our ODEB estimation under example sampling are also summarized in Fig. 8 by fixing β_Y and examining the resulting RMSE, bias, average confidence interval coverage, and average confidence interval length for various sampling proportions. RMSE, bias, and confidence interval length are all higher for OLS estimation. Also, only ODEB-based confidence intervals maintain nominal coverage.

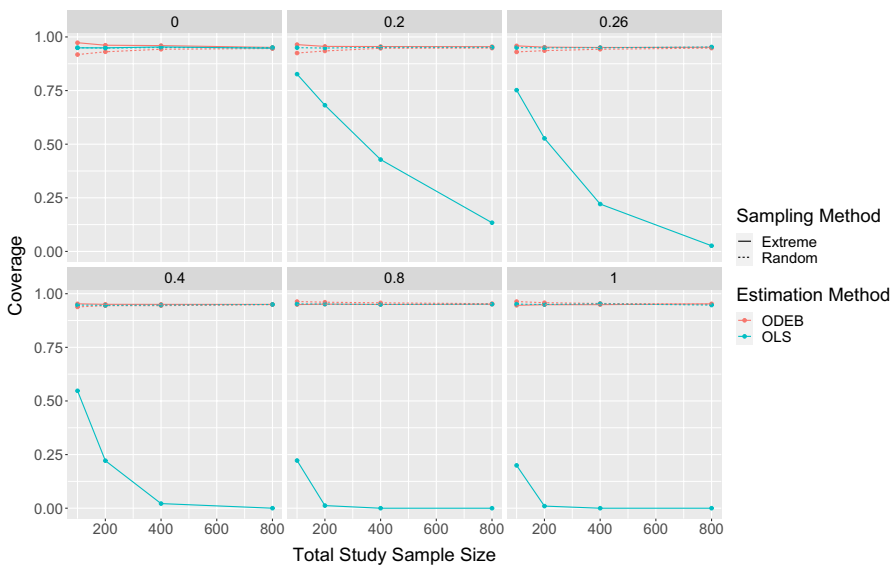


Fig. 6 95% β_Y confidence interval coverage under 20% of the total study sampled for various β_Y 's

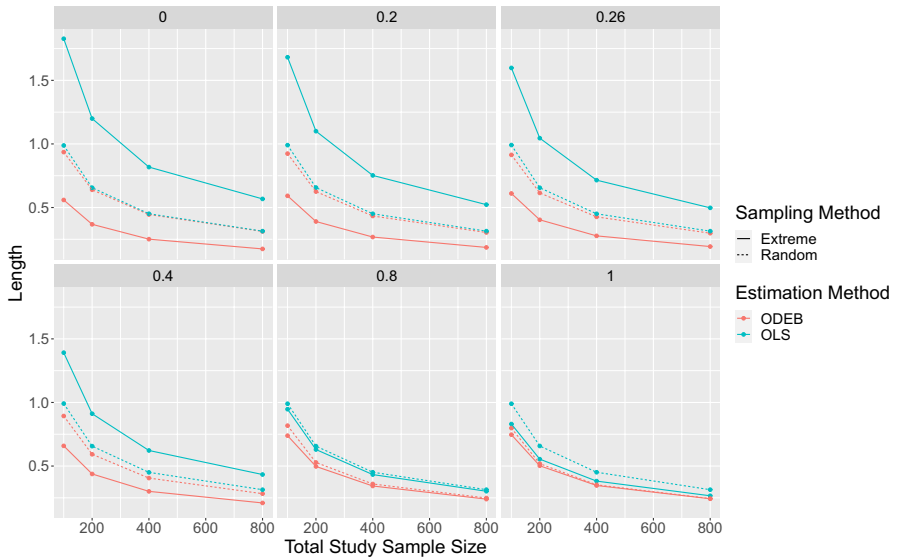


Fig. 7 95% β_Y confidence interval length under 20% of the total study sampled for various β_Y 's

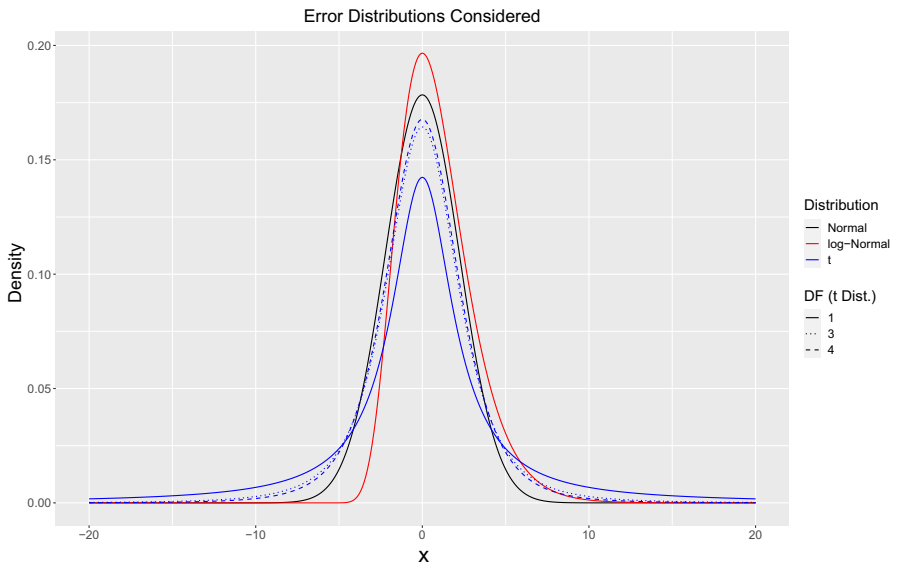


Fig. 8 Operating characteristics of ODEB and OLS estimation for $\beta_Y = 0.2$ under extreme sampling

3.1.2 Simulations with Non-normal Residuals and Non-normal Inputs

To explore the degree to which the ODEB estimation method is affected by violations to the assumption of joint normality, the simulation above was repeated for $\epsilon_{Y,i}$

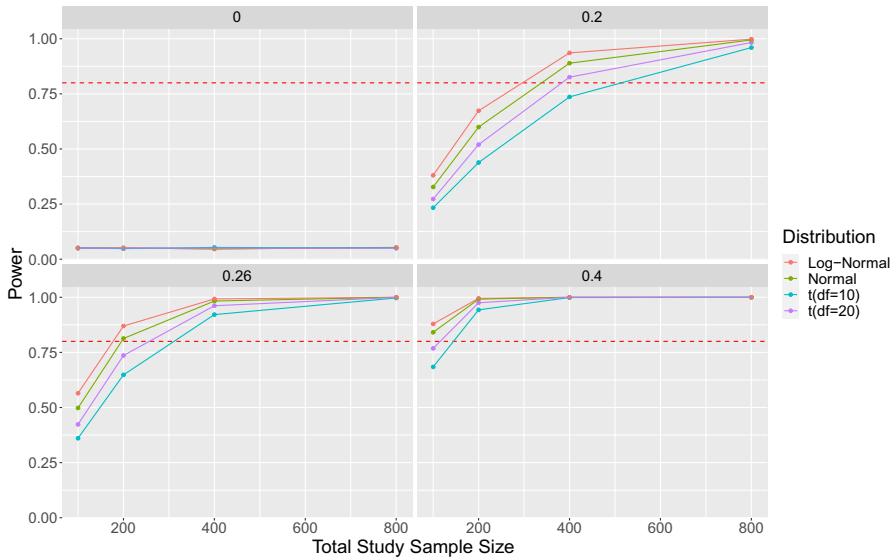


Fig. 9 Power of $\hat{\beta}_{Y,ODEB}$ under skewed (log-normal) or heavy-tailed (t) residual distributions and an extreme sample of 20% (Normal residual distribution also plotted for reference)

following a scaled t distribution (scale $\sqrt{5}$) with $DF \in \{10, 20\}$ and a shifted log-Normal distribution with a mode of zero and variance of 5.

From the upper-left panel (the case of no biomarker effect with $\beta_Y = 0$) of Fig. 9, the hypothesis test to detect the effect always has the correct level of $\alpha = 0.05$ even if the normality assumption is violated. This is expected because, whatever the residual distribution is, there is no effect in the reverse-regression when the regression effect $\beta_Y = 0$. The power of the test improves when the residuals are skewed (shifted log-normal) and deteriorates when the residual distribution becomes more heavily tailed (smaller degrees of freedom for t-distribution). In all simulated settings, the power of detection exceeds 80% (shown as the horizontal dotted red line) when the full sample size is $n_F = 800$ (with $n_S = 160$ selected for biomarker-testing).

The bias of $\hat{\beta}_{Y,ODEB}$ is affected by the residual distributions as seen in Fig. 10. Correspondingly, the coverage probability of the 95% confidence interval is no longer valid for non-normal residuals in Fig. 11. The coverage deteriorates when the effect sizes increases, when the sample size increases, and when the tails of residual distribution become heavier.

We also conduct similar simulations to check the effect of non-normal inputs X . Instead of non-normal residuals, we repeat the simulation above for covariate X following a scaled t distribution (scale $\sqrt{5}$, mean of 20) with $DF \in \{10, 20\}$. The effect patterns are the same as those observed above when the residual distribution is heavy-tailed. That is, ODEB estimation with an extreme sampling scheme still provides good power to detect an effect, but the estimation and confidence interval coverage deteriorate as β_Y increases.

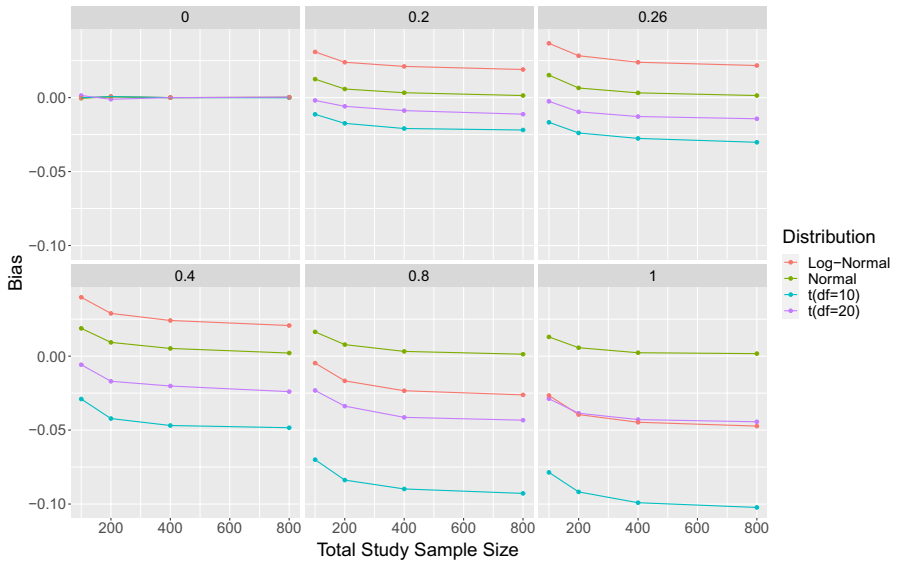


Fig. 10 Bias of $\hat{\beta}_{Y,ODEB}$ under skewed (log-normal) or heavy-tailed (t) residual distributions and an extreme sample of 20% (Normal residual distribution also plotted for reference)

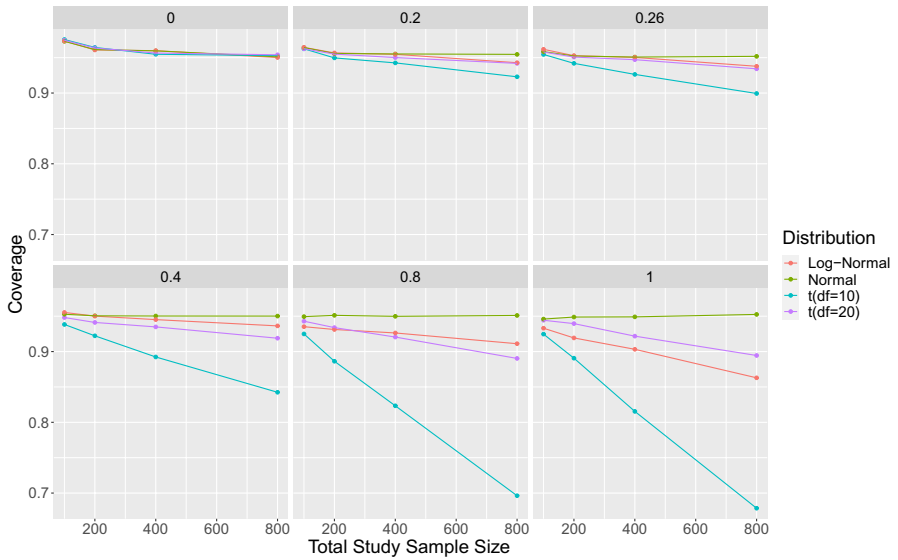


Fig. 11 Coverage of 95% confidence interval for β_Y under skewed (log-normal) or heavy-tailed (t) residual distributions and an extreme sample of 20% (Normal residual distribution also plotted for reference)

Overall, when the normality assumption is violated, the effect detection test is still valid and powerful but the effect estimation is no longer reliable.

3.1.3 Comparison with MSELE Estimator

Additionally, we compared the ODEB estimator to the MSELE estimator [18]. Notice that the MSELE estimator cannot be applied directly for data from extreme outcome-dependent sampling. It requires an additional simple random sample (SRS) so that some of the individuals with medial Y values are biomarker-tested also. Thus the comparison is conducted on extreme outcome-dependent sampling data plus some SRS data, mostly as a sanity check for the new ODED estimator against a well-established estimator in a setting close to extreme outcome-dependent sampling. We generate data following the ODS experimental design outlined by Zhou et al [18].

Again, for each of $B = 20,000$ iterations, individual data sets consisting of $(X_1, Y_1), \dots, (X_{n_F}, Y_{n_F})$ were randomly generated. This data was eligible for extreme outcome-dependent sampling. An additional SRS of size $n_{SRS} = 80$ was generated. For each combination of parameter values described above, an extreme sample of size γn_F was selected. Both $\hat{\beta}_{Y,ODEB}$ and $\hat{\beta}_{Y,MSELE}$ were calculated on the merged sample consisting of the extreme sample and the additional SRS for a total sample size of $n = \gamma n_F + n_{SRS}$. The *ODS* package [33] in R was used to calculate $\hat{\beta}_{Y,MSELE}$ and its corresponding standard error.

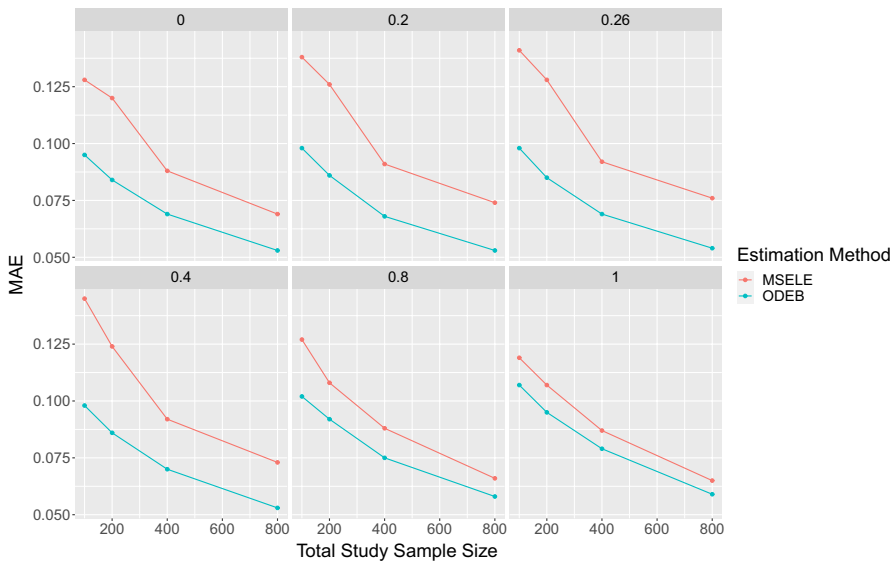


Fig. 12 Root mean square error (RMSE) of $\hat{\beta}_{Y,MSELE}$ compared to $\hat{\beta}_{Y,ODEB}$ under 40% of the total study sampled for various β_Y 's

Figure 12 compares the RMSE of $\hat{\beta}_{Y,MSELE}$ and $\hat{\beta}_{Y,ODEB}$. The $\hat{\beta}_{Y,MSELE}$ is obtained from an iterative numerical optimization algorithm, and it does depend on the starting values which poses a challenge in convergence. We also plotted proportion of convergence of $\hat{\beta}_{Y,MSELE}$ out of $B = 20,000$ simulation runs on the figure. The convergence rate of $\hat{\beta}_{Y,MSELE}$ is around 50% or lower when the effect size is small. As the effect size increases, the divergence issue alleviates but is still significant even when $\beta_Y = 1$. The RMSE of $\hat{\beta}_{Y,MSELE}$ from the convergent runs are plotted, and are much higher than RMSE of $\hat{\beta}_{Y,ODEB}$.

The RMSE is heavily affected by outliers. Sometimes, $\hat{\beta}_{Y,MSELE}$ converges to a value that is far from the true value, indicating failure of the iterative algorithm to find the true root. These wrong convergent cases inflate the RMSE of $\hat{\beta}_{Y,MSELE}$ a lot. To be fairer for $\hat{\beta}_{Y,MSELE}$, Fig. 13 presents the median absolute error (MAE) comparison instead. The MAE of $\hat{\beta}_{Y,MSELE}$ is now closer to, but still clearly exceeds, the MAE of $\hat{\beta}_{Y,ODEB}$. This accuracy improvement by $\hat{\beta}_{Y,ODEB}$ is expected here as it utilizes the correct parametric assumption which $\hat{\beta}_{Y,MSELE}$ does not make. As β grows larger, this gap in performance between $\hat{\beta}_{Y,MSELE}$ and $\hat{\beta}_{Y,ODEB}$ shrinks.

These results confirm that, in a setting close to extreme outcome-dependent sampling, the proposed ODEB estimator provides correct answer. When the parametric assumptions hold, the ODEB estimator has smaller estimation error than MSELE as expected. For extreme outcome-dependent sampling, MSELE cannot be used, and ODEB should work well.

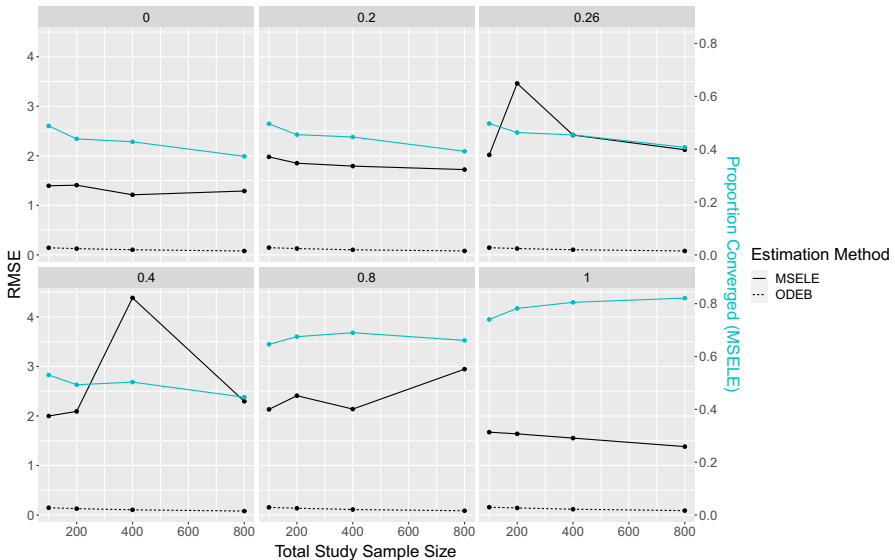


Fig. 13 Median absolute error (MAE) of $\hat{\beta}_{Y,MSELE}$ compared to $\hat{\beta}_{Y,ODEB}$ under 40% of the total study sampled for various β_Y 's

3.2 Data Analysis/Application to BISP2 Trial

The dependent variable of interest in the BISP2 Trial, pain intensity, is measured via the BPI to represent an underlying pain continuum for all participants. A pilot sample of 31 participants in the high and low tails of worst pain experienced at 48-hours post shoulder pain induction was selected for initial biomarker-testing. As the BPI is a discrete index represented on a continuous scale, there were ties for the highest and lowest pain scores. To determine which of the equal scores were included in the pilot sample, those who responded vs. did not respond to the treatment were selected in a balanced manner. Based on promising results from the initial pilot sample, a follow-up sample of 57 participants was selected in the same manner. All biomarkers were logarithm (base 10, for interpretability) transformed based on prior knowledge of the typical biomarker distribution. We model the relationship between pain and log-transformed biomarker linearly as:

$$y_{\text{Pain}} = \alpha + \beta \log(x_{\text{Biomarker}}) + \epsilon$$

Two pain-related outcomes of interest were explored using the above methodology: worst pain experienced at 48-hours post shoulder pain induction and change in worst pain from 48- to 96-hours post induction. Additionally, a self-reported disability score (Quick-DASH) at 48-hours post induction and change in Quick-DASH score from 48- to 96-hours post induction was investigated as outcomes to assess endpoint specificity. The results of modeling the association between pain-related outcomes and log-transformed biomarkers are discussed below. The distribution of worst pain at 48-hours in each stage of sampling is given in Fig. 14. For each pain-related

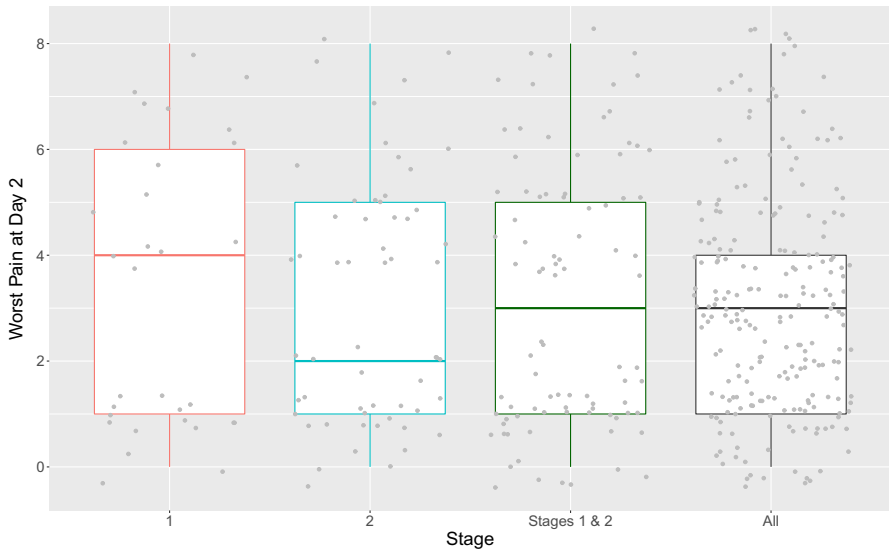


Fig. 14 Worst pain experienced at 48-hours shoulder pain induction for each sampling stage, combined stages, and the full BISP2 study population

Table 1 Summaries from univariable reverse-regression based extreme sampling estimation and inference: worst pain at 48 h regressed on log-scale biomarkers

	Estimate	Std. Error	LCL	UCL	P-Value
log(CCL2)	1.923	0.847	0.238	3.607	0.026
log(CRP)	0.369	0.296	-0.220	0.958	0.216
log(CXCL6)	-0.775	0.973	-2.710	1.161	0.428
log(IL17/IL17A)	0.682	1.059	-1.423	2.787	0.521
log(IL10)	0.742	1.290	-1.823	3.306	0.567
log(TNF Alpha)	0.690	1.403	-2.099	3.479	0.624
log(BDNF)	-0.139	0.313	-0.760	0.483	0.658
log(IL6)	0.376	1.475	-2.556	3.307	0.799
log(Beta NGF)	0.323	1.823	-3.301	3.947	0.860
log(TNF R1)	0.269	1.641	-2.993	3.531	0.870
log(Oncostatin M OSM)	0.176	1.737	-3.278	3.629	0.920
log(Substance P)	0.107	1.085	-2.05	2.265	0.921
log(Cortisol)	0.020	0.491	-0.955	0.995	0.968

Table 2 Summaries from univariable reverse-regression based extreme sampling estimation and inference: change in worst pain from 48 h to 96 h regressed on log-scale biomarkers

	Estimate	Std. Error	LCL	UCL	P-Value
log(CRP)	0.710	0.332	0.050	1.370	0.039
log(IL10)	2.752	1.452	-0.135	5.639	0.065
log(IL6)	3.078	1.661	-0.225	6.380	0.071
log(TNF Alpha)	2.269	1.599	-0.910	5.448	0.163
log(Oncostatin M OSM)	2.262	1.991	-1.697	6.22	0.262
log(TNFR1)	2.018	1.883	-1.726	5.761	0.289
log(Beta NGF)	1.456	2.104	-2.727	5.639	0.492
log(CCL2)	0.659	0.988	-1.305	2.623	0.507
log(Cortisol)	-0.376	0.566	-1.502	0.750	0.509
log(Substance P)	0.411	1.256	-2.086	2.909	0.744
log(BDNF)	-0.083	0.362	-0.803	0.637	0.819
log(IL17/IL17A)	0.174	1.228	-2.267	2.615	0.888
log(CXCL6)	-0.062	1.130	-2.308	2.183	0.956

outcome, reverse-regression based estimates of the effect of each log-transformed biomarker on pain (β) are reported with standard errors, 95% confidence intervals, and p-values. The the reverse-regression model assumption was assessed using diagnostic plots of residuals.

The results of the analysis of the combined stages 1 and 2 data for worst pain and change in pain are shown in Tables 1 and 2. Table 1 shows evidence of association between CCL2 and worst pain at 48 h, with an effect estimate of 1.92 (95% CI: 0.24, 3.61). This indicates that for a tenfold increase in CCL2, we expect the worst

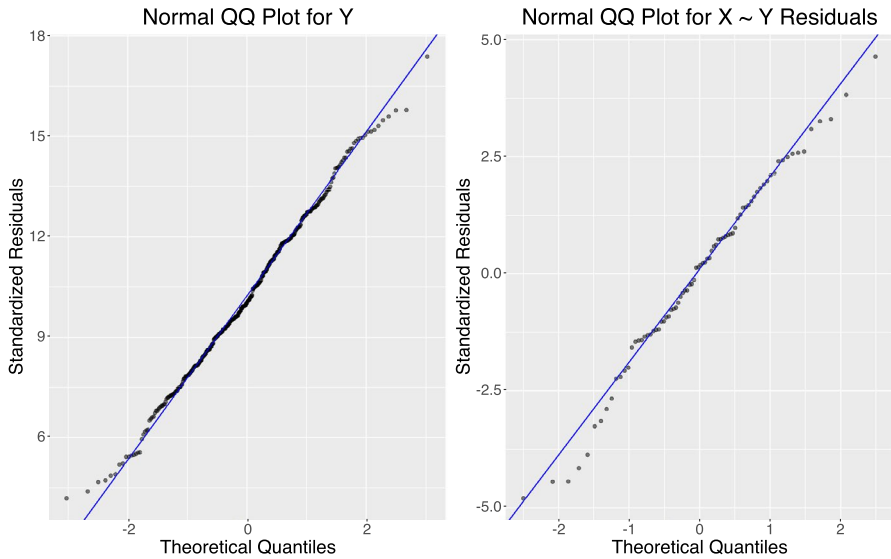


Fig. 15 Model checking: normal QQ plot example for the change in pain outcome and the reverse-regression of $Y_{Pain\ Change}$ on $\log(X_{CRP})$

pain at 48 h to increase by 1.92 points on the BPI scale. For change in pain, there is evidence of association from CRP ($\hat{\beta} = 0.71$ (95% CI 0.05, 1.37)). We also see potential from IL10 and IL6, however the evidence in this sample is insufficient to conclude such. In additional analyses for self-reported disability outcomes (whose full result tables are omitted for succinctness), CRP shows evidence of association with Quick-DASH score. For a tenfold increase in CRP, we expect an increase of 3.92 points in the Quick-DASH score.

Figure 15 provides the model checking diagnostic plots for the change in pain outcome and the residuals of the reverse-regression of log-transformed CRP on change in pain. Our joint normality assumption appears to be reasonable for this example.

4 Discussion

We proposed a new least squares approach to analysis for data generated by extreme outcome-dependent sampling based on the joint normality assumption. Prior methods developed for outcome-dependent sampling data are analyzed through likelihood-based methods [18] with an additional SRS to provide some observations in the middle range of response values. The ODS sampling schemes described in such works are the closest to EODS in the statistical literature. However, due to the presence of the supplemental SRS across the entire range of the response in the traditional ODS designs, the prior analysis approaches proposed are not applicable to data generated from an EODS design. Despite the differences in analysis needs, it is

helpful in defining the future development of this novel method to contrast ODEB with the existing MSELE methodology.

The MSELE estimator is a semi-parametric empirical likelihood-based analysis method. The main advantage of this likelihood approach is that no parametric distributional assumption is needed, as the supplemental SRS provides knowledge of the range of the response outside of the typically extreme regions of interest. In contrast, through the joint normality assumption, our method provides two significant practical advantages. First, our method allows concentration of all sampling to the more informative individuals with the most extreme response values, resulting in great cost reduction for the expensive biomarker-testing. Taking on a parametric model allows us to assume the shape of the relationship in the medial region of the response range where no data were collected.

However, a primary limitation of this estimation is that in practice the joint normality assumption cannot be taken for granted. As in many model-based analyses, thorough knowledge of the relationship between the response and biomarkers to be studied is recommended to plan whether a reasonable transformation can be applied before the sampling design is enacted. Additionally, the diagnostic plots for model checking described in Sect. 2 must be thoroughly inspected to ensure validity of estimation. When the normality assumption is violated, the confidence interval coverage deteriorates only for very heavy-tailed residuals, but the hypothesis testing for effect detection always remains valid.

The second practical advantage is that our method is simple and can be conducted with existing standard statistical software. As seen in the simulation, often the likelihood method diverges or it converges to wrong parameter values, thus its data analysis often requires much care from an expert statistician. On the other hand, our method only requires applications of simple formulas on standard least-square estimates from the reverse-regression, and can be handled by any practitioner with a minimum statistical training. Furthermore, the model checking techniques used for the joint normality assumption are standard, with familiarity among any who have had fundamental training in linear regression.

This approach to analysis of EODS data is promising, but future development of the estimation method is key. In particular, extension to multiple regression analysis, which accounts for the correlation between biomarkers, is needed. In principle, derivation of formulas assuming joint normal distribution of a response variable and multiple biomarker variables may follow the same approach, but is non-trivial and hence the topic of future study. If the biomarkers and response variable Y jointly follow the multivariate normal distribution, then each biomarker and Y follows the bivariate normal distribution, thus the proposed inference methods are valid to apply separately on each biomarker even when multiple biomarkers are correlated amongst themselves. However, the test statistics of (5) for multiple biomarkers are correlated, thus an appropriate multiple testing adjustment (beyond the conservative Bonferroni correction or the step-down testing procedure) for variable selection needs to be carefully developed in the future. Additionally, exploration of whether the joint normality assumption can be relaxed is another future step.

5 Conclusion

We proposed an intuitive least squares-based estimator appropriate for EODS biomarker-testing studies with response and biomarker following a bivariate normal distribution, and showed via simulation that the estimation method is efficient and unbiased when compared to OLS estimation. To our knowledge, this is the first approach proposed for EODS which was not specifically developed for the genetics literature, where the variable X is discrete.

Our new ODEB estimator is easy to understand and implement, and so may be a bridge between practitioners who may benefit from ODS designs but do not have the background to immediately understand the likelihood-based methods by serving as an accessible introduction to ODS analysis. In the future, multivariate extension to our ODEB estimator will be particularly suited for exploratory biomarker studies due to its cost savings and ease of usage.

Appendix A Mathematical Derivation of Formulas Relating Parameters of Regression and Reverse-regression

We assume that (X, Y) follows the bivariate normal distribution

$$N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right).$$

Under this assumption, we consider how to get regression parameters α_Y, β_Y and σ_{ε_Y} from the reverse-regression parameters α_X, β_X and σ_{ε_X} .

Conditional on X , we have the regression equation

$$Y = \alpha_Y + \beta_Y X + \varepsilon_Y, \tag{10}$$

where $\varepsilon_Y \sim N(0, \sigma_{\varepsilon_Y}^2)$ with $\sigma_{\varepsilon_Y}^2 = (1 - \rho^2)\sigma_Y^2$. Similarly, conditional on Y , we have the reverse-regression equation

$$X = \alpha_X + \beta_X Y + \varepsilon_X, \tag{11}$$

where $\varepsilon_X \sim N(0, \sigma_{\varepsilon_X}^2)$ with $\sigma_{\varepsilon_X}^2 = (1 - \rho^2)\sigma_X^2$.

Firstly we derive how the regression parameters of equation (11) relate to the bivariate normal distribution parameters. To do this, we consider the standardized versions of X and Y as $X^* = \frac{X - \mu_X}{\sigma_X}$ and $Y^* = \frac{Y - \mu_Y}{\sigma_Y}$. Then clearly, (X^*, Y^*) follows the bivariate normal distribution

$$N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Hence we have

$$X^* = 0 + \rho Y^* + \sqrt{1 - \rho^2} Z, \tag{12}$$

where $Z \sim N(0, 1)$ is independent of Y . Plug $X^* = \frac{X-\mu_X}{\sigma_X}$ and $Y^* = \frac{Y-\mu_Y}{\sigma_Y}$ into equation (12) and compare with equation (11), we can express the reverse-regression parameters in terms of the bivariate normal distribution parameters as

$$\begin{aligned} \alpha_X &= \mu_X - \beta_X \mu_Y &= \mu_X - \rho \frac{\sigma_X}{\sigma_Y} \mu_Y, \\ \beta_X &= \rho \frac{\sigma_X}{\sigma_Y}, \\ \sigma_{\varepsilon_X}^2 &= (1 - \rho^2) \sigma_X^2. \end{aligned} \tag{13}$$

Secondly, by symmetry, similar derivations give the regression parameters' formulas in terms of the bivariate normal distribution parameters as

$$\begin{aligned} \alpha_Y &= \mu_Y - \beta_Y \mu_X &= \mu_Y - \rho \frac{\sigma_Y}{\sigma_X} \mu_X, \\ \beta_Y &= \rho \frac{\sigma_Y}{\sigma_X}, \\ \sigma_{\varepsilon_Y}^2 &= (1 - \rho^2) \sigma_Y^2. \end{aligned} \tag{14}$$

We now find formulas to express regression parameters α_Y , β_Y and σ_{ε_Y} in terms of α_X , β_X , σ_{ε_X} , μ_Y and σ_Y .

We start from the parameter of most interest β_Y . Combine the middle equations in (13) and (14), we get

$$\beta_Y = \rho \frac{\sigma_Y}{\sigma_X} = \left(\rho \frac{\sigma_X}{\sigma_Y} \right) \frac{\sigma_Y^2}{\sigma_X^2} = \beta_X \frac{\sigma_Y^2}{\sigma_X^2}. \tag{15}$$

Now, to get the desired expression of β_Y , we only need to express σ_X^2 in terms of α_X , β_X , σ_{ε_X} , μ_Y and σ_Y .

From the middle equation in (13),

$$\beta_X = \rho \frac{\sigma_X}{\sigma_Y} \Rightarrow \beta_X \sigma_Y = \rho \sigma_X. \tag{16}$$

Square both sides of equation (16) and divides the last equation in (13), we get

$$\frac{\beta_X^2 \sigma_Y^2}{\sigma_{\varepsilon_X}^2} = \frac{\rho^2}{1 - \rho^2}.$$

Therefore, we solve ρ^2 in this expression to get

$$\rho^2 = \frac{1}{1 + \frac{\sigma_{\varepsilon_X}^2}{\beta_X^2 \sigma_Y^2}}, \quad \text{and} \quad 1 - \rho^2 = 1 - \frac{1}{1 + \frac{\sigma_{\varepsilon_X}^2}{\beta_X^2 \sigma_Y^2}} = \frac{\frac{\sigma_{\varepsilon_X}^2}{\beta_X^2 \sigma_Y^2}}{1 + \frac{\sigma_{\varepsilon_X}^2}{\beta_X^2 \sigma_Y^2}} = \frac{1}{1 + \frac{\beta_X^2 \sigma_Y^2}{\sigma_{\varepsilon_X}^2}}. \tag{17}$$

Plug this back into the last equation in (13) to solve for σ_X^2 , we get

$$\sigma_X^2 = \frac{\sigma_{\epsilon_X}^2}{1 - \rho^2} = \sigma_{\epsilon_X}^2 \left[1 + \frac{\beta_X^2 \sigma_Y^2}{\sigma_{\epsilon_X}^2} \right] = \sigma_{\epsilon_X}^2 + \beta_X^2 \sigma_Y^2.$$

Put this into equation (15), we have the desired expression for β_Y as

$$\beta_Y = \beta_X \frac{\sigma_Y^2}{\sigma_{\epsilon_X}^2 + \beta_X^2 \sigma_Y^2} = \frac{1}{\frac{\sigma_{\epsilon_X}^2}{\sigma_Y^2} + \beta_X^2} \beta_X = \frac{\beta_X \sigma_Y^2}{\sigma_{\epsilon_X}^2 + \beta_X^2 \sigma_Y^2}. \tag{18}$$

Next, we express α_Y in terms of α_X , β_X , σ_{ϵ_X} , μ_Y and σ_Y . From the first equation in (13), $\mu_X = \alpha_X + \beta_X \mu_Y$. Plug this and (18) both into the first equation of (14), we get

$$\alpha_Y = \mu_Y - \beta_Y \mu_X = \mu_Y - \left(\frac{\beta_X \sigma_Y^2}{\sigma_{\epsilon_X}^2 + \beta_X^2 \sigma_Y^2} \right) (\alpha_X + \beta_X \mu_Y) = \frac{\sigma_{\epsilon_X}^2 \mu_Y - \alpha_X \beta_X \sigma_Y^2}{\sigma_{\epsilon_X}^2 + \beta_X^2 \sigma_Y^2}. \tag{19}$$

Finally, for the expression of $\sigma_{\epsilon_Y}^2$, we plug (17) into the last equation of (14) to get

$$\sigma_{\epsilon_Y}^2 = \frac{1}{1 + \frac{\beta_X^2 \sigma_Y^2}{\sigma_{\epsilon_X}^2}} \sigma_Y^2 = \frac{\sigma_Y^2 \sigma_{\epsilon_X}^2}{\sigma_{\epsilon_X}^2 + \beta_X^2 \sigma_Y^2}. \tag{20}$$

Equations (18), (19) and (20) give the relationship between the theoretical parameters, where the quantities β_Y , α_Y and $\sigma_{\epsilon_Y}^2$ on the left-hand side are the desired parameters for regression analysis but cannot be directly estimated from data (their OLS estimates are inconsistent). Since the quantities on the right-hand side can be consistently estimated directly on the extremely sampled data through reverse-regression, as explained in the main text, plugging-in those estimators to equations (18), (19) and (20) result in consistent estimators $\hat{\beta}_Y$, $\hat{\alpha}_Y$ and $\hat{\sigma}_{\epsilon_Y}^2$ for the desired parameters.

Furthermore, we can use the Delta Method [34, p. 61] to derive the standard errors of estimators for the plug-in estimators since the standard errors for the right-hand side quantities estimators are available directly from the reverse-regression. Particularly, for $\hat{\beta}_Y$, we first find the following partial derivatives from the equation (18):

$$\begin{aligned} \frac{\partial}{\partial \beta_X} \beta_Y &= \frac{1}{\frac{\sigma_{\epsilon_X}^2}{\sigma_Y^2} + \beta_X^2} + \beta_X \frac{-1}{\left(\frac{\sigma_{\epsilon_X}^2}{\sigma_Y^2} + \beta_X^2\right)^2} (2\beta_X) = \frac{\frac{\sigma_{\epsilon_X}^2}{\sigma_Y^2} - \beta_X^2}{\left(\frac{\sigma_{\epsilon_X}^2}{\sigma_Y^2} + \beta_X^2\right)^2}, \\ \frac{\partial}{\partial (\sigma_{\epsilon_X}^2)} \beta_Y &= \frac{-\beta_X}{\left(\frac{\sigma_{\epsilon_X}^2}{\sigma_Y^2} + \beta_X^2\right)^2} \frac{1}{\sigma_Y^2} = \frac{\frac{-\beta_X}{\sigma_Y^2}}{\left(\frac{\sigma_{\epsilon_X}^2}{\sigma_Y^2} + \beta_X^2\right)^2}, \\ \frac{\partial}{\partial (\sigma_Y^2)} \beta_Y &= \frac{-\beta_X}{\left(\frac{\sigma_{\epsilon_X}^2}{\sigma_Y^2} + \beta_X^2\right)^2} \frac{-\sigma_{\epsilon_X}^2}{(\sigma_Y^2)^2} = \frac{\frac{\beta_X \sigma_{\epsilon_X}^2}{\sigma_Y^4}}{\left(\frac{\sigma_{\epsilon_X}^2}{\sigma_Y^2} + \beta_X^2\right)^2}. \end{aligned} \tag{21}$$

Since $\hat{\beta}_X$, $\hat{\sigma}_{\epsilon_X}^2$ and $\hat{\sigma}_Y^2$ are uncorrelated with each other, using the Delta Method, $\hat{\beta}_X$ is asymptotically normally distributed whose variance estimation is

$$\begin{aligned} (s.e.\{\hat{\beta}_Y\})^2 &= \left(\frac{\partial}{\partial \beta_X} \beta_Y\right)^2 (s.e.\{\hat{\beta}_X\})^2 + \left(\frac{\partial}{\partial (\sigma_{\epsilon_X}^2)} \beta_Y\right)^2 (s.e.\{\hat{\sigma}_{\epsilon_X}^2\})^2 + \left(\frac{\partial}{\partial (\sigma_Y^2)} \beta_Y\right)^2 (s.e.\{\hat{\sigma}_Y^2\})^2 \\ &= \frac{\left(\frac{\sigma_X^2}{\sigma_Y^2} - \beta_X^2\right)^2 (s.e.\{\hat{\beta}_X\})^2 + \left(\frac{\sigma_X^2}{\sigma_Y^2}\right) (s.e.\{\hat{\sigma}_{\epsilon_X}^2\})^2 + \left(\frac{\beta_X^2 \sigma_{\epsilon_X}^4}{\sigma_Y^2}\right) (s.e.\{\hat{\sigma}_Y^2\})^2}{\left(\frac{\sigma_X^2}{\sigma_Y^2} + \beta_X^2\right)^4}. \end{aligned} \tag{22}$$

For the quantities in the formula (22), $s.e.\{\hat{\beta}_X\}$ can be gotten from outputs of standard linear regression fit packages, and

$$(s.e.\{\hat{\sigma}_Y^2\})^2 = \frac{2\sigma_Y^4}{n_F - 1}, \quad (s.e.\{\hat{\sigma}_{\epsilon_X}^2\})^2 = \frac{2\sigma_{\epsilon_X}^4}{n_s - 2},$$

from the variance formula of the Chi-square distribution.

Plug-in the point estimators for each quantity into equation (22), we estimate the standard error for $\hat{\beta}_Y$ as

$$s.e.\{\hat{\beta}_Y\} = \sqrt{\frac{\left(\frac{\hat{\sigma}_{\epsilon_X}^2}{\hat{\sigma}_Y^2} - \hat{\beta}_X^2\right)^2 (s.e.\{\hat{\beta}_X\})^2 + \left(\frac{2\hat{\beta}_X^2 \hat{\sigma}_{\epsilon_X}^4}{\hat{\sigma}_Y^2}\right) \left(\frac{1}{n_s - 2} + \frac{1}{n_F - 1}\right)}{\left(\frac{\hat{\sigma}_{\epsilon_X}^2}{\hat{\sigma}_Y^2} + \hat{\beta}_X^2\right)^4}}. \tag{23}$$

Appendix B Mathematical derivation of power formulas

For the standard simple linear regression on a data set of size n , the power of an α level t-test (testing the zero slope null hypothesis) is given

$$power = P(NF_{ncp=nf^2, df_1=1, df_2=n-2} > FQ_{\alpha, df_1=1, df_2=n-2}), \tag{24}$$

where $NF_{ncp=nf^2, df_1=1, df_2=n-2}$ denotes a random variable following a non-central F-distribution with noncentral parameter nf^2 , degrees of freedoms of 1 and $n - 2$, and $FQ_{\alpha, df_1=1, df_2=n-2}$ denotes the α upper quantile of a central F-distribution with degrees of freedoms of 1 and $n - 2$. Here f is the Cohen's effect size defined as

$$f^2 = \frac{R^2}{1 - R^2} = \frac{\rho^2}{1 - \rho^2}, \tag{25}$$

where R^2 is the proportion of variation in the data explained by the regression equation.

We are doing the reverse-regression hypothesis test for

$$H_0 : \beta_X = 0 \quad \text{versus} \quad H_A : \beta_X \neq 0.$$

If we have observations of both X and Y on the full data set, then the power is given by formula (26) with $n = n_F$. Now we derive the power formula when the test is conducted on the selectively sampled subset. Besides the changes in sample size from n_F to n_S , the effect size also changes from the full data set to the selectively sampled subset. The R^2 in (25) would be larger in the subset due to selectively sampling the extreme values. To derive the change in effect size, we reexpress (25) as

$$f^2 = \frac{R^2}{1 - R^2} = \frac{1}{\frac{1}{R^2} - 1} = \frac{1}{\frac{1}{\rho^2} - 1} = \frac{\beta_X^2 \sigma_Y^2}{\sigma_{\epsilon_X}^2},$$

where the last equality comes from equation (17). Notice that both quantities β_X^2 and $\sigma_{\epsilon_X}^2$ remain invariant on the full data set and the subset since the reverse-regression model (11) holds on both data sets. σ_Y^2 differs in the two data sets. Since only the extreme γ proportion of Y is selected, the variance of selected Y_S is bigger on the selectively sampled subset than the variance of Y on the full data set. To calculate $\sigma_{Y_S}^2$, we note that the selected Y_S do not follow the $N(\mu_Y, \sigma_Y^2)$ distribution anymore. Rather it follows the normal distribution truncated at the upper and lower $(\gamma/2)$ -tails. Let $z_{\gamma/2}$ denotes the upper $(\gamma/2)$ -quantile of the standard normal distribution $N(0, 1)$. Let $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ denotes the density function of the standard normal distribution $N(0, 1)$. Then the variance of selected Y_S is

$$\sigma_{Y_S}^2 = \sigma_{Y_F}^2 2 \int_{z_{\gamma/2}}^{\infty} x^2 \frac{1}{\gamma} \phi(x) dx. \tag{26}$$

Therefore, compared to the effect size on the full data set, the effect size f^2 on the subset increases by a factor of $2 \int_{z_{\gamma/2}}^{\infty} x^2 \frac{1}{\gamma} \phi(x) dx$. Since the sample size n_S of the subset is γ proportion of the full data set size n_F so that $n_S/\gamma = n_F$, the power of t-test on the subset becomes

$$\begin{aligned} \text{power} &= P(NF_{ncp=n_S f^2 2 \int_{z_{\gamma/2}}^{\infty} x^2 \frac{1}{\gamma} \phi(x) dx, df_1=1, df_2=n_S-2} > FQ_{\alpha, df_1=1, df_2=n_S-2}) \\ &= P(NF_{ncp=n_F f^2 2 \int_{z_{\gamma/2}}^{\infty} x^2 \phi(x) dx, df_1=1, df_2=\gamma n_F-2} > FQ_{\alpha, df_1=1, df_2=\gamma n_F-2}). \end{aligned} \tag{27}$$

Acknowledgements We would like to thank Steve George and Mark Bishop for their support and helpful comments of this work, and Brian Bouverat and the Metabolism and Translational Science Core of the Claude D. Pepper Older Americans Independence Center for biomarker analyses.

Author Contributions All authors commented on the draft and the interpretation of the findings, read and approved the final manuscript. AAD and NJD jointly contributed to drafting the article. AAD: conception of proposed estimator, drafting the methodological sections of article, critical revision of the article, final approval of the version to be published. NJD: drafting the non-methodological sections of article, writing the code for and performing the simulations, data analysis, preparing figures and tables, critical revision of the article. SSW: conception the work, supervision of the data collection, data analysis and interpretation, critical revision of the article, final approval of the version to be published.

Funding Open access funding provided by Northeastern University Library. Dr. Wu was partially supported by the funding from the National Institutes of Health and National Institute of Arthritis and Musculoskeletal and Skin Diseases (AR055899). All authors were independent from this funding source.

Data Availability The BISP2 example dataset is available from the corresponding author upon request.

Code Availability The simulation code that supports the findings of this study is available upon request.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical approval The University of Florida Institutional Review Board approved the Biopsychosocial Influence of Shoulder Pain trial (IRB201500556) protocol.

Informed Consent All participants provided informed consent for eligibility screening, trial enrollment, study intervention, and the current study of association analysis between biomarkers and pain outcomes.

Consent to Participate All study procedures were followed in accordance with the relevant guidelines (e.g. Declaration of Helsinki) under the Ethics approval and consent to participate heading.

Consent for Publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Wagner JA (2002) Overview of biomarkers and surrogate endpoints in drug development. *Dis Mark* 18(2):41–46. <https://doi.org/10.1155/2002/929274>
2. Strimbu K, Tavel JA (2010) What are biomarkers? *Curr Opin HIV and AIDS* 5(6):463–466. <https://doi.org/10.1097/COH.0b013e32833ed177>
3. Schisterman EF, Albert PS (2012) The biomarker revolution. *Stat Med* 31(22):2513–2515. <https://doi.org/10.1002/sim.5499>.The
4. Atkinson AJ, Colburn WA, DeGruttola VG, DeMets DL, Downing GJ, Hoth DF, Oates JA, Peck CC, Schooley RT, Spilker BA, Woodcock J, Zeger SL (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 69(3):89–95. <https://doi.org/10.1067/mcp.2001.113989>
5. The Luminex FLEXMAP 3D® System - UF ICBR (2020). <https://biotech.ufl.edu/the-luminex-flexmap-3d-system-a-multiplexed-analytical-platform-for-novel-biomarker-discovery/>
6. Albert PS, Schisterman EF (2012) Novel statistical methodology for analyzing longitudinal biomarker data. *Stat Med* 31(22):2457–2460. <https://doi.org/10.1002/sim.5500>.Novel
7. Darvasi A, Soller M (1992) Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theor Appl Genet* 85(2–3):353–359. <https://doi.org/10.1007/BF00222881>
8. Lander ES, Botstein D (2012) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Proc Am Control Conf*. <https://doi.org/10.1109/acc.2012.6315381>

9. Muranty H, Goffinet B (1997) Selective genotyping for location and estimation of the effect of a quantitative trait locus. *Biometrics* 53(2):629–643
10. Sen S, Satagopan JM, Churchill GA (2005) Quantitative trait locus study design from an information perspective. *Genetics* 170:447–464. <https://doi.org/10.1534/genetics.104.038612>
11. Holt D, Smith TMF, Winter PD (1980) Regression analysis of data from complex surveys. *J R Stat Soc* 143(4):474–487
12. Rabier C-E (2014) On statistical inference for selective genotyping. *J Stat Plan Inference* 147:24–52
13. Carey G, Williamson J (1991) Linkage analysis of quantitative traits: increased power by using selected samples. *Am J Human Genet* 49:786–796
14. Van Gestel S, Houwing-Duistermaat JJ, Adolfsen R, Van Duijn CM, Van Broeckhoven C (2000) Power of selective genotyping in genetic association analyses of quantitative traits. *Behav Genet* 30(2):141–146. <https://doi.org/10.1023/A:1001907321955>
15. Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB (2002) Two-stage designs for gene-disease association studies. *Biometrics* 58(1):163–170. <https://doi.org/10.1111/j.0006-341X.2002.00163.x>
16. Satagopan JM, Venkatraman ES, Begg CB (2004) Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* 60(3):589–597. <https://doi.org/10.1111/j.0006-341X.2004.00207.x>
17. Lawless JF, Kalbfleisch JD, Wild CJ (1999) Semiparametric methods for response-selective and missing data problems in regression. *J R Stat Soc* 61(2):413–438
18. Zhou H, Weaver MA, Qin J, Longnecker MP, Wang MC (2002) A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics* 58(June):413–421
19. Weaver MA, Zhou H (2005) An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *J Am Stat Assoc* 100(470):459–469. <https://doi.org/10.1198/016214504000001853>
20. Tan Z, Qin G, Zhou H (2016) Estimation of a partially linear additive model for data from an outcome-dependent sampling design with a continuous outcome. *Biostatistics* 17(4):663–676. <https://doi.org/10.1093/biostatistics/kxw015>
21. Wang X, Zhou H (2006) A semiparametric empirical likelihood method for biased sampling schemes with auxiliary covariates. *Biometrics* 62(4):1149–1160. <https://doi.org/10.1111/j.1541-0420.2006.00612.x>
22. Wang X, Wu Y, Zhou H (2009) Outcome- and auxiliary-dependent subsampling and its statistical inference. *J Biopharm Stat* 19(6):1132–1150. <https://doi.org/10.1080/10543400903243025>
23. Zhou H, Song R, Wu Y, Qin J (2011) Statistical inference for a two-stage outcome-dependent sampling design with a continuous outcome. *Biometrics* 67(1):194–202. <https://doi.org/10.1111/j.1541-0420.2010.01446.x>
24. Song R, Zhou H, Kosorok MR (2009) A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika* 96(1):221–228. <https://doi.org/10.1093/biomet/asn073>
25. Xu W, Zhou H (2012) Mixed effect regression analysis for a cluster-based two-stage outcome-auxiliary-dependent sampling design with a continuous outcome. *Biostatistics* 13(4):650–664. <https://doi.org/10.1093/biostatistics/kxs013>
26. Yu J, Zhou H, Cai J (2021) Accelerated failure time model for data from outcome-dependent sampling. *Lifetime Data Anal* 27(1):15–37. <https://doi.org/10.1007/s10985-020-09508-y>
27. Schildcrout JS, Mumford SL, Chen Z, Heagerty PJ, Rathouz PJ (2012) Outcome-dependent sampling for longitudinal binary response data based on a time-varying auxiliary variable. *Stat Med* 31(22):2441–2456. <https://doi.org/10.1002/sim.4359>
28. Zelnick LR, Schildcrout JS, Heagerty PJ (2018) Likelihood-based analysis of outcome-dependent sampling designs with longitudinal data. *Stat Med* 37(13):2120–2133. <https://doi.org/10.1002/sim.7633>
29. Borsa PA, Parr JJ, Wallace MR, Wu SS, Dai Y, Fillingim RF, George SZ (2018) Genetic and psychological factors interact to predict physical impairment phenotypes following exercise-induced shoulder injury. *J Pain Res* 11:2497–2508. <https://doi.org/10.2147/JPR.S171498>
30. George SZ, Bishop MD, Wu SS, Staud R, Borsa PA, Wallace MR, Greenfield WH, Dai Y, Fillingim RF (2022) Biopsychosocial influence on shoulder pain: results from a randomized pre-clinical trial of exercise-induced muscle injury. *Pain* 1:1–15. <https://doi.org/10.1097/j.pain.0000000000002700>

31. Keller S, Bann CM, Dodd SL, Schein J, Mendoza RR, Cleeland CS (2004) Validity of the brief pain inventory for use in documenting the outcomes of patients with noncancer pain. *Clin J Pain* 20:309–318
32. R Core Team (2021) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing. <https://www.R-project.org/>
33. Pan Y, Zhou H, Weaver M, Qin G, Cai J (2018) ODS: statistical methods for outcome-dependent sampling designs. R package version 0.2.0. <https://CRAN.R-project.org/package=ODS>
34. Lehmann EL, Casella G (1998) Theory of point estimation, 2nd edn. Springer, New York