Biometrics 64, 1–25 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

Causal Meta-Analysis by Integrating Multiple Observational Studies with Multivariate Outcomes

Subharup Guha

Department of Biostatistics, University of Florida, Gainesville, Florida, U.S.A.

email: s.guha@ufl.edu

and

Yi Li

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A.

email: yili@umich.edu

Summary: Integrating multiple observational studies to make unconfounded causal or descriptive comparisons of group potential outcomes in a large natural population is challenging. Moreover, retrospective cohorts, being convenience samples, are usually unrepresentative of the natural population of interest and have groups with unbalanced covariates. We propose a general covariate-balancing framework based on pseudo-populations that extends established weighting methods to the meta-analysis of multiple retrospective cohorts with multiple groups. Additionally, by maximizing the effective sample sizes of the cohorts, we propose a FLEXible, Optimized, and Realistic (FLEXOR) weighting method appropriate for integrative analyses. We develop new weighted estimators for unconfounded inferences on wide-ranging population-level features and estimands relevant to group comparisons of quantitative, categorical, or multivariate outcomes. Asymptotic properties of these estimators are examined. Through simulation studies and meta-analyses of TCGA datasets, we demonstrate the versatility and reliability of the proposed weighting strategy, especially for the FLEXOR pseudo-population.

Key words: FLEXOR; Pseudo-population; Retrospective cohort; Unconfounded comparison; Weighting.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

The study of differential patterns of oncogene expression levels across cancer subtypes has aroused great interest because it unveils new tumorigenesis mechanisms and can improve cancer screening and treatment (Kumar et al., 2020). In a multi-site breast cancer study conducted at seven medical centers, including, for example, Memorial Sloan Kettering, Mayo Clinic, and University of Pittsburgh, the goal was to compare the mRNA expression levels of eight targeted breast cancer genes, namely, COL9A3, CXCL12, IGF1, ITGA11, IVL, LEF1, PRB2, and SMR3B (e.g., Christopoulos et al., 2015) in the disease subtypes infiltrating ductal carcinoma (IDC) and infiltrating lobular carcinoma (ILC), which account for nearly 80% and 10% of breast cancer cases in the United States (Wright, 2022; Tran, 2022). The data reposited at The Cancer Genome Atlas (TCGA) portal (NCI, 2022) include demographic, clinicopathological, and biomarker measurements; some study-specific attributes are summarized in Web Table 2 of Supplementary Materials. Each breast cancer patient's outcome is a vector of mRNA expression measurements for these eight targeted genes.

Inference focuses on interpreting biomarker comparisons between the disease subtypes IDC and ILC in the context of a larger disease population in the U.S., e.g., SEER breast cancer patients (Surveillance Research Program, NCI, 2023). The estimands of interest include contrasts and gene-gene pairwise correlations, alongside disease subtype-specific summaries (e.g., means, standard deviations, and medians). Understanding gene expression and co-expression patterns in different subtypes of breast cancer among national-level patients is crucial for developing feasible guidelines for regulating targeted therapies and precision medicine (Schmidt et al., 2016). As revealed by Web Table 2 of Supplementary Materials, naive group comparisons based on the TCGA patient cohorts are severely confounded by the high degree of covariate imbalance between the IDC and ILC subtypes.

More broadly, covariate balance is vitally important in observational studies where inter-

est focuses on unconfounded causal comparisons of group potential outcomes (Robins and Rotnitzky, 1995; Rubin, 2007) in a large *natural population* such as the U.S. population. The *observed populations* of convenience samples such as observational studies are usually unrepresentative of this natural population. Theoretical and simulation studies have demonstrated the conceptual and practical advantages of weighting over other covariate-balancing techniques like matching and regression adjustment (Austin, 2010). As a result, weighting methods have widespread applicability in diverse research areas such as political science, sociology, and healthcare (Lunceford and Davidian, 2004). For analyzing cohorts consisting of two groups, the propensity score (PS) (Rosenbaum and Rubin, 1983) plays a central role. In these studies, the average treatment effect (ATE) and average treatment effect on the treated group (ATT) are overwhelmingly popular estimands (Robins et al., 2000). However, the inverse probability weights (IPW) on which these estimators rely may be unstable when some PSs are near 0 or 1 (Li and Li, 2019).

Several researchers have proposed variations of ATE based on truncated subpopulations of scientific or statistical interest (Crump et al., 2006; Li and Greene, 2013). Most weighting methods, implicitly or explicitly, provide unbiased inferences for a specific *pseudo-population*, a covariate-balanced construct that often differs substantially from the real but mostly unknown natural population of interest. For example, Li et al. (2018) showed that IPWs correspond to a *combined* pseudo-population and introduced the *overlap* pseudo-population, wherein the weights minimize the asymptotic variance of the weighted average treatment effect for the overlap pseudo-population (ATO). For single observational studies comprising two or more groups, Li and Li (2019) proposed the *generalized overlap* pseudo-population that minimizes the sum of asymptotic variances of weighted estimators of pairwise group differences. For multiple observational studies with two groups, Wang and Rosner (2019) developed an integrative approach for Bayesian inferences on ATE. For single observational

studies with two groups, Mao et al. (2019) obtained analytical variance expressions of modified IPW estimators adjusted for the estimated PS and augmented the estimators with outcome models for improved efficiency. Zeng et al. (2023) explored weighting procedures in single study, multiple group settings with censored survival outcomes.

However, these methods have several limitations. *First*, they are theoretically guaranteed to be effective for a specific set of outcome types and estimands under certain theoretical conditions (e.g., equal variances of univariate group-specific outcomes). As study endpoints may be continuous, categorical, or multivariate, inference procedures for disparate outcome types have been inadequately explored. Further, scientific interests may necessitate alternative estimands than ATE, ATT or ATO, such as distribution percentiles, standard deviations, pairwise correlations of multivariate outcomes, and unplanned estimands suggested during post hoc analyses. *Second*, these methods may imply group assignment changes for some subjects that are sometimes difficult to justify for a meaningful, generalizable pseudo-population (Li et al., 2018; Li and Li, 2019). *Lastly*, very few methods can accommodate the integration of multiple observational studies with multiple unbalanced groups as encountered in the TCGA datasets. One potential use of the existing weighing methods to achieve covariate balance is by creating a new categorical variable that combines study and group information. However, it is unclear how to conduct unconfounded group comparisons independent of the "nuisance" study factor. Furthermore, the pseudo-populations generated by this approach are often impractical, and inferential accuracies for common estimands are frequently suboptimal. There is a critical need for developing efficient approaches that enable the integration of multiple observational studies and multiple unbalanced groups and the construction of pseudo-populations that resemble the natural population of interest.

To fill this gap, we extend the propensity score to the multiple propensity score and propose a new class of pseudo-populations and multi-study balancing weights to effectuate

data integration and causal meta-analyses. Compared to the existing weighting methods, our work presents two main advances. First, our framework enables unconfounded inferences on a wide variety of population-level group features as well as planned or unplanned estimands relevant to group comparisons. Second, the framework allows us to derive efficient estimators within this proposed family of pseudo-populations. Specifically, by maximizing the effective sample size, we further obtain a FLEXible, Optimized, and Realistic (FLEXOR) weighting method and derive new weighted estimators which are efficient for a variety of quantitative, categorical, and multivariate outcomes, are applicable to different weighting strategies, and effectively utilize multivariate outcome information. For example, the estimators yield efficient estimates of various functionals of group-specific potential outcomes, e.g., contrasts of means and medians, correlations, and percentiles.

The rest of the paper is organized as follows. Section 2 introduces some basic notation, theoretical assumptions, and a general covariate-balancing framework for meta-analysis. We further introduce FLEXOR, an optimized pseudo-population, as its special case. Section 3 develops unconfounded integrative estimators applicable to different weighting methods, estimands, and response types, and establishes asymptotic properties. Section 4 presents the finite sample performance of the proposed methodology, especially when used in conjunction with the FLEXOR weights. Section 5 meta-analyzes the aforementioned TCGA studies and detects differential targeted gene expression and co-expression patterns across the two major breast cancer subtypes in the Unites States. Section 6 concludes with some final remarks.

2. Integration of Observational Studies with Multiple Unbalanced Groups

2.1 *Notation and basic assumptions*

We aim to compare K subpopulations or groups (e.g., disease subtypes) of participants belonging to a large natural population such as the U.S. patient population. Beyond basic

summaries (e.g., group prevalences) from preexisting registries, no additional information is available about the natural population. The investigation comprises J observational studies. We assume J and K are not large. For $i = 1 \ldots, N$, let $Z_i \in \{1, \ldots, K\}$ denote the group and $S_i \in \{1, \ldots, J\}$ denote the observational study. We assume that each participant belongs to exactly one observational study and each study includes at least one participant in each group. Additionally, there are p covariates shared by all the studies and denoted by $\mathbf{X}_i \in \mathcal{X} \subset \mathcal{R}^p$ for the ith participant. The motivating TCGA database comprises $K = 2$ groups corresponding to breast cancer subtype IDC and ILC, and $p = 30$ covariates of $N = 450$ breast cancer patients in $J = 7$ observational studies. The ith participant's potential outcome is $\mathbf{Y}_i^{(z)} = (Y_{i1}^{(z)}, \ldots, Y_{iL}^{(z)})' \in \mathcal{R}^L$, i.e., the outcome had the patient belonged to group $z = 1, \ldots, K$. The observed outcome is $\mathbf{Y}_i = \mathbf{Y}_i^{(Z_i)}$. In the TCGA example, vectors $\mathbf{Y}_i^{(1)}$ and $\mathbf{Y}_i^{(2)}$ represent counterfactual mRNA measurements of disease subtypes IDC and ILC on $L = 8$ targeted genes, and the observed $\mathbf{Y}_i \in \mathcal{R}^8$ contains mRNA measurements of breast cancer subtype Z_i with which participant i is actually diagnosed.

The participant-specific measurements are a random sample from an *observed distribution*, $p[S, Z, \mathbf{X}, \mathbf{Y}]_+$, where $p[\cdot]_+$ generically represents distributions or densities with respect to the observed population. Extending Rubin (2007) and Imbens (2000), we assume (A) **Stable unit treatment value assumption (SUTVA)**: Given subjects' covariates, the study and group memberships do not influence the potential outcomes, and no two versions of grouping lead to different potential outcomes; (B) **Study-specific weak unconfoundedness**: Given study S and covariate vector $\mathbf{X}$, group membership Z is independent of the potential outcomes $\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(K)}$; and (C) **Positivity**: Joint density $p[S = z, Z = z, \mathbf{X} = \mathbf{x}]_+$ is strictly positive for all $(s, z, \mathbf{x})$. Assumption (B) states that $p[\mathbf{Y}^{(z)} \mid S, Z, \mathbf{X}]_+ = p[\mathbf{Y}^{(z)} \mid S, \mathbf{X}]_+$. Assumption (C) guarantees that the study and group memberships and covariates do not have deterministic relationships and often holds when J and K are not large.

2.2 *A new family of pseudo-populations*

We first extend variations of the propensity score (e.g., Rosenbaum and Rubin, 1983) to the *multiple propensity score* (MPS) of the vector (S, Z). For $\mathbf{x} \in \mathcal{X} \subset \mathcal{R}^p$, the MPS

$$\delta_{sz}(\mathbf{x}) = p\big[S = s, Z = z \mid \mathbf{X} = \mathbf{x}\big]_+ \quad \text{for } (s, z) \in \Sigma \equiv \{1, \ldots, J\} \times \{1, \ldots, K\}. \tag{1}$$

It then follows that the joint density $p\big[S = s, Z = z, \mathbf{X} = \mathbf{x}\big]_+ = \delta_{sz}(\mathbf{x}) f_+(\mathbf{x})$, where $f_+(\mathbf{x}) = p[\mathbf{X} = \mathbf{x}]_+$ represents the marginal covariate density in the observed population. As the MPS is unknown in observational studies, we can estimate it by regressing the combinations of (S_i, Z_i) on covariate $\mathbf{x}_i$ $(i = 1, \ldots, N)$. In single studies with two groups, the PS is usually estimated using logistic regression (Mao et al., 2019). For estimating MPS, we recommend multinomial logistic regression: $\log\big(\delta_{sz}(\mathbf{x})/\delta_{11}(\mathbf{x})\big) = \boldsymbol{\omega}'_{sz}\mathbf{x}$ for $(s, z) \neq (1, 1)$, so that $\boldsymbol{\omega} = \big\{\boldsymbol{\omega}_{sz} : (s, z) \neq (1, 1)\big\}$ is a $(JK - 1)p$-dimensional parameter. If we define $\boldsymbol{\omega}_{11}$ to be the vector of p zeros, then $\delta_{sz}(\mathbf{x}) = \exp(\boldsymbol{\omega}'_{sz}\mathbf{x}) / \sum_{s^*=1}^{J} \sum_{z^*=1}^{K} \exp(\boldsymbol{\omega}'_{s^*z^*}\mathbf{x})$ for all $(s, z) \in \Sigma$.

Consider a pseudo-population with attributes fully or partially prescribed by the investigator via two probability vectors: (i) relative amounts of information extracted from the studies, quantified by probability tuple $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)$; and (ii) relative group prevalence, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$. For instance, in the TCGA breast cancer studies, setting $\gamma_j = 1/7$ extracts equal information from each study, whereas $\boldsymbol{\theta} = (8/9, 1/9)$ constrains the pseudo-population to the known U.S. proportions of breast cancer subtypes IDC and ILC (Wright, 2022; Tran, 2022). If some or all components of $\boldsymbol{\gamma}$ or $\boldsymbol{\theta}$ are unknown, subsequent inferences can optimize the pseudo-population over the multiple possibilities for these quantities.

For multiple observational studies, the participant study memberships $S_1, \ldots, S_N$ are primarily influenced by the J study designs and unknown factors driving participation; moreover, study participant characteristics can differ substantially across studies, especially in cancer investigations. To address these issues, we aim to design a pseudo-population for achieving theoretical covariate balance between the K groups. In other words, we construct

a pseudo-population wherein the study memberships, the group memberships and patient characteristics are mutually independent, i.e., $S \perp Z \perp \mathbf{X}$, so that

$$p\big[S = s, Z = z, \mathbf{X} = \mathbf{x}\big] = \gamma_s \, \theta_z \, f_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x}), \quad \text{for } (s, z, \mathbf{x}) \in \Sigma \times \mathcal{X}. \tag{2}$$

Here and hereafter, $p[\cdot]$ denotes a distribution or density with respect to the designed pseudo-population, whereas $p[\cdot]_+$ corresponds to the observed population, as mentioned earlier. Equation (2) further emphasizes that although S, Z, and $\mathbf{X}$ are independent in the pseudo-population, they may share some distributional parameters. More explicitly, the subscripts of $f_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x})$ emphasize that the pseudo-population density of $\mathbf{X}$ may depend on $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$.

Next, consider the relationship between the pseudo-population covariate density, $f_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x})$, and the marginal observed covariate density, $f_+(\mathbf{x})$. Assuming a common dominating measure for the densities and a common support, $\mathcal{X}$, there exists without loss of generality a positive *tilting function* (e.g., Li et al., 2018) denoted by $\eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}$ such that $f_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x}) \propto \eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x}) f_+(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. Therefore, $f_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x}) = \eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x}) f_+(\mathbf{x}) / \mathbb{E}_+[\eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{X})]$ where $\mathbf{X} \sim f_+$ and $\mathbb{E}_+(\cdot)$ denotes expectations under the observed distribution. Intuitively, high tilting function values correspond to covariate space regions with high pseudo-population weights. Let $\mathcal{S}_J$ denote the unit simplex in $\mathcal{R}^J$. Different choices of $\boldsymbol{\gamma} \in \mathcal{S}_J$, $\boldsymbol{\theta} \in \mathcal{S}_K$, and tilting function $\eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}$ identify different pseudo-populations with structure (2).

Balancing weights for integration of multiple studies. To efficiently meta-analyze multiple studies (with $J > 1$), we propose the *multi-study balancing weight*, defined as the ratio of the joint densities with respect to the pseudo-population and observed population. More specifically, for any $(s, z, \mathbf{x}) \in \Sigma \times \mathcal{X}$, the multi-study balancing weight

$$\rho_{\boldsymbol{\gamma},\boldsymbol{\theta}}(s, z, \mathbf{x}) = \frac{p\big[S = s, Z = z, \mathbf{X} = \mathbf{x}\big]}{p\big[S = s, Z = z, \mathbf{X} = \mathbf{x}\big]_+} = \frac{\gamma_s \, \theta_z \, f_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x})}{\delta_{sz}(\mathbf{x}) \, f_+(\mathbf{x})} = \frac{\gamma_s \, \theta_z \, \eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x})}{\delta_{sz}(\mathbf{x}) \, \mathbb{E}_+[\eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{X})]}. \tag{3}$$

As $\rho_{\boldsymbol{\gamma},\boldsymbol{\theta}}(s, z, \mathbf{x}) \times p\big[S = s, Z = z, \mathbf{X} = \mathbf{x}\big]_+ = p\big[S = s, Z = z, \mathbf{X} = \mathbf{x}\big]$, the balancing weight serves to redistribute the observed distribution's relative mass to match that of the pseudo-

population. Defining the *unnormalized weight function* as $\tilde{\rho}_{\boldsymbol{\gamma},\boldsymbol{\theta}}(s,z,\mathbf{x}) = \gamma_s\theta_z\eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x})/\delta_{sz}(\mathbf{x})$, the *unnormalized weight* of the ith participant is $\tilde{\rho}_i = \tilde{\rho}_{\boldsymbol{\gamma},\boldsymbol{\theta}}(s_i,z_i,\mathbf{x}_i)$. For a general pseudo-population (e.g., FLEXOR pseudo-population introduced in the sequel), the unnormalized weights, even within a study-group combination, may depend on $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ through the tilting function. As discussed later, the unnormalized weights can be utilized to provide unconfounded inferences on various potential outcome features for a general pseudo-population.

The proposed pseudo-populations and balancing weights are general, encompassing many well-known weighting methods in single-study settings. For example, in single studies, assume equally prevalent pseudo-population groups ($\theta_z = 1/K$) in expression (2). A constant tilting function yields IPWs when $K = 2$ and generalized IPWs (Imbens, 2000) when $K > 2$. On the other hand, $\eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x}) = 1/\sum_z \delta_z^{-1}(\mathbf{x})$ produces overlap weights (Li et al., 2018) when $K = 2$, and generalized overlap weights (Li and Li, 2019) when $K > 2$. Again, if $\eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x}) = \delta_{z'}(\mathbf{x})$ for a group z', then the pseudo-population's covariate density, $f_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x})$, matches the observed covariate density of the group z' participants.

The choice of different tilting functions in (2) naturally extends several weighting methods designed for single studies to meta-analytical settings. For example, assuming equally weighted studies and equally prevalent groups, i.e., $\gamma_s = 1/J$ and $\theta_z = 1/K$, a constant tilting function $\eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x}) \propto 1$ and $\eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x}) = 1/\sum_s\sum_z \delta_{sz}^{-1}(\mathbf{x})$, respectively, produces extensions of the combined (Li et al., 2018) and generalized overlap (Li and Li, 2019) pseudo-populations appropriate for meta-analyzing multiple studies with multiple groups. We refer to these proposed pseudo-populations as the *integrative combined* (IC) and *integrative generalized overlap* (IGO) pseudo-populations, respectively. Similarly, for a fixed group z', the tilting function $\eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x}) = \sum_s \delta_{sz'}(\mathbf{x})$ gives a pseudo-population whose marginal covariate density equals the observed covariate density of group z' participants irrespective of their study

memberships. Given the availability of different tilting functions, an important question arises: which choice is optimal and in what sense? We address this below.

Effective sample size. A widely used measure of a pseudo-population's inferential accuracy is the *effective sample size* (ESS), $\mathcal{Q}(\boldsymbol{\gamma}, \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}) = N/\big[1 + \mathrm{Var}_+\big\{\rho_{\boldsymbol{\gamma},\boldsymbol{\theta}}(S, Z, \mathbf{X})\big\}\big] = N/\mathbb{E}_+\big\{\rho^2_{\boldsymbol{\gamma},\boldsymbol{\theta}}(S, Z, \mathbf{X})\big\}$, which relies on the second moment (provided it exists) of the balancing weights in the observed population (e.g., McCaffrey et al., 2013). The ESS is asymptotically equivalent to the sample ESS, $\tilde{\mathcal{Q}}(\boldsymbol{\gamma}, \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}) = N^2(\sum_{i=1}^N \tilde{\rho}_i)^2/\sum_{i=1}^N \tilde{\rho}_i^2$. Informally, the ESS is the hypothetical sample size from the pseudo-population containing the same information as N samples from the observed population, and it is always less than N unless the pseudo-population and observed population are identical.

An optimized case: FLEXOR pseudo-population. We propose *FLEXOR* as a member of pseudo-population family (2) that maximizes the ESS or minimizes the variation of the balancing weights, subject to any problem-dictated constraints on the vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$. That is, if the triplet $(\breve{\boldsymbol{\gamma}}, \breve{\boldsymbol{\theta}}, \breve{\eta}_{\breve{\boldsymbol{\gamma}},\breve{\boldsymbol{\theta}}})$ identifies the FLEXOR pseudo-population and $(\boldsymbol{\gamma}, \boldsymbol{\theta})$ is known to belong to a subset, Υ, of $\mathcal{S}_J \times \mathcal{S}_K$, then $\mathcal{Q}(\breve{\boldsymbol{\gamma}}, \breve{\boldsymbol{\theta}}, \breve{\eta}_{\breve{\boldsymbol{\gamma}},\breve{\boldsymbol{\theta}}}) = \sup_{(\boldsymbol{\gamma},\boldsymbol{\theta})\in\Upsilon} \sup_{\eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}} \mathcal{Q}(\boldsymbol{\gamma}, \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma},\boldsymbol{\theta}})$.

A two-step procedure for constructing the FLEXOR pseudo-population. Starting with an initial $(\boldsymbol{\gamma}, \boldsymbol{\theta}) \in \Upsilon$, we iteratively perform the following steps until convergence:

- **Step I** For a fixed $(\boldsymbol{\gamma}, \boldsymbol{\theta})$, maximize sample ESS $\tilde{\mathcal{Q}}(\boldsymbol{\gamma}, \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma},\boldsymbol{\theta}})$ over all tilting functions, $\eta_{\boldsymbol{\gamma},\boldsymbol{\theta}}$. This gives the *best fixed-*$(\boldsymbol{\gamma}, \boldsymbol{\theta})$ *pseudo-population* identified by $(\boldsymbol{\gamma}, \boldsymbol{\theta}, \breve{\eta}_{\boldsymbol{\gamma},\boldsymbol{\theta}})$. The analytical form of $\breve{\eta}_{\boldsymbol{\gamma},\boldsymbol{\theta}}$ for the theoretical ESS is given in Theorem 1 below. Set function $\eta = \breve{\eta}_{\boldsymbol{\gamma},\boldsymbol{\theta}}$.
- **Step II** For a fixed tilting function η, maximize $\tilde{\mathcal{Q}}(\boldsymbol{\gamma}, \boldsymbol{\theta}, \eta)$ over all $(\boldsymbol{\gamma}, \boldsymbol{\theta}) \in \Upsilon$ to obtain the *best fixed-*η *pseudo-population*, identified by the triplet $(\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}}, \eta)$. This parametric maximization over $\Upsilon \subset \mathcal{S}_J \times \mathcal{S}_K$ can be quickly performed in R using the `optim` function or by Gauss-Seidel or Jacobi algorithms. Set $(\boldsymbol{\gamma}, \boldsymbol{\theta}) = (\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}})$.

In our experience, convergence is attained within only a few iterations. The converged

pseudo-population with the largest ESS yields the FLEXOR pseudo-population. The following theorem gives the analytical expression for the global maximum of $\mathcal{Q}(\boldsymbol{\gamma}, \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma},\boldsymbol{\theta}})$ mentioned in Step I. See Web Appendix A.1 of Supplementary Materials for the proof.

THEOREM 1: *Suppose probability vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ have strictly positive elements and are held fixed. Let $\boldsymbol{\Xi}$ be the set of tilting functions for which the ESS, $\mathcal{Q}(\boldsymbol{\gamma}, \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma},\boldsymbol{\theta}})$, of pseudo-population (2) is finite. Maximizing $\mathcal{Q}(\boldsymbol{\gamma}, \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma},\boldsymbol{\theta}})$ over all tilting functions $\eta_{\boldsymbol{\gamma},\boldsymbol{\theta}} \in \boldsymbol{\Xi}$, the optimal fixed-$(\boldsymbol{\gamma}, \boldsymbol{\theta})$ pseudo-population's tilting function, denoted by $\breve{\eta}_{\boldsymbol{\gamma},\boldsymbol{\theta}}$, has the expression:*

$$\breve{\eta}_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x}) = \left(\sum_{s=1}^{J} \sum_{z=1}^{K} \frac{\gamma_s^2 \theta_z^2}{\delta_{sz}(\mathbf{x})} \right)^{-1}, \qquad \mathbf{x} \in \mathcal{X}. \tag{4}$$

The unnormalized weight function for the optimal fixed-$(\boldsymbol{\gamma}, \boldsymbol{\theta})$ pseudo-population is then

$$\tilde{\rho}_{\boldsymbol{\gamma},\boldsymbol{\theta}}(s, z, \mathbf{x}) = \frac{1}{\gamma_s \theta_z} \left(\frac{\gamma_s^2 \theta_z^2 / \delta_{sz}(\mathbf{x})}{\sum_{t=1}^{J} \sum_{u=1}^{K} \gamma_t^2 \theta_u^2 / \delta_{tu}(\mathbf{x})} \right), \qquad \text{for } (s, z) \in \Sigma \text{ and } \mathbf{x} \in \mathcal{X}. \tag{5}$$

The optimal fixed-$(\boldsymbol{\gamma}, \boldsymbol{\theta})$ pseudo-population's balancing weights, evaluated as in (3), are uniformly bounded. The ESS of the optimal fixed-$(\boldsymbol{\gamma}, \boldsymbol{\theta})$ pseudo-population is $N\mathbb{E}_+\left[\breve{\eta}_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{X})\right]$ with the expectation taken over $\mathbf{X} \sim f_+(\mathbf{x})$, the observed population's covariate density.

It can be shown that the optimal tilting function $\breve{\eta}_{\boldsymbol{\gamma},\boldsymbol{\theta}}(\mathbf{x})$ apportions low importance to outlying regions of covariate space $\mathcal{X}$ where $\delta_{sz}(\mathbf{x})$ is approximately 0 for some $(s, z) \in \Sigma$. Furthermore, the optimal tilting function emphasizes covariate regions where the group propensities $\delta_1(\mathbf{x}), \ldots, \delta_K(\mathbf{x})$ match the group proportions $\theta_1, \ldots, \theta_K$ of the larger natural population. In particular, in pseudo-populations with equally prevalent groups, the tilting function promotes covariate regions where the group propensities are approximately equal.

3. Meta-analyses of Group Potential Outcomes

Causal meta-analyses generally follow a two-stage inferential procedure (e.g., Rubin, 2007). In *Stage 1*, the "outcome free" analysis only utilizes covariate information to estimate the propensity scores, as done in Section 2. In *Stage 2*, for the pseudo-population of interest, the

procedure makes unconfounded comparisons of group potential outcomes via estimands such as pairwise difference of group means. For any known pseudo-population belonging to family (2), the procedure accommodates wide-ranging group-level features of the endpoints using the available multivariate outcome information. Additionally, we derive analytical expressions for the asymptotic variances of the proposed multivariate estimators.

Suppose potential outcome vectors $\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(K)}$ have a common support, $\mathcal{Y} \subset \mathcal{R}^L$. To ensure that SUTVA, weak unconfoundedness, and positivity of the observed population also hold for the pseudo-population, we assume identical conditional distributions:

$$p[\mathbf{Y}^{(z)} \mid S, Z, \mathbf{X}] = p[\mathbf{Y}^{(z)} \mid S, Z, \mathbf{X}]_+ \quad \text{for group } z = 1, \ldots, K, \tag{6}$$

where $p[\cdot|\cdot]_+$ and $p[\cdot|\cdot]$ denote the observed and pseudo-population conditional densities, respectively. Unlike the observed population, the covariate-balanced pseudo-population entails $p[\mathbf{Y} \mid Z = z] = p[\mathbf{Y}^{(z)}]$, enabling us to construct weighted estimators of various features of the pseudo-population potential outcomes.

Let $\mathbb{E}[\cdot]$ denote expectations with respect to the pseudo-population. Let $\Phi_1, \ldots, \Phi_M$ be real-valued functions having domain $\mathcal{Y}$. We wish to infer pseudo-population means of transformed potential outcomes, $\mathbb{E}[\Phi_1(\mathbf{Y}^{(z)})], \ldots, \mathbb{E}[\Phi_M(\mathbf{Y}^{(z)})]$ for $z = 1, \ldots, K$. Appropriate choices of Φ_m correspond to pseudo-population inferences about group-specific marginal means, medians, variances, and CDFs of potential outcome components. Equivalently, writing $\mathbf{\Phi}(\mathbf{Y}^{(z)}) = \big(\Phi_1(\mathbf{Y}^{(z)}), \ldots, \Phi_M(\mathbf{Y}^{(z)})\big)' \in \mathcal{R}^M$, the inferential focus is the vector, $\boldsymbol{\lambda}^{(z)} = \mathbb{E}[\mathbf{\Phi}(\mathbf{Y}^{(z)})]$.

For real-valued functions ψ with domain $\mathcal{R}^M$, we estimate $\psi(\boldsymbol{\lambda}^{(z)})$. For example, if the first two components of $\mathbf{Y}^{(z)}$ are quantitative, then defining $\Phi_1(\mathbf{Y}^{(z)}) = Y_1^{(z)}$, $\Phi_2(\mathbf{Y}^{(z)}) = Y_2^{(z)}$, $\Phi_3(\mathbf{Y}^{(z)}) = Y_1^{(z)}Y_2^{(z)}$, and $\psi(t_1, t_2, t_3) = t_3 - t_1 t_2$, we obtain $\psi(\boldsymbol{\lambda}^{(z)}) = \text{cov}(Y_1^{(z)}, Y_2^{(z)})$ as the pseudo-population covariance of $Y_1^{(z)}$ and $Y_2^{(z)}$ in the zth group. The pseudo-population correlation of pairwise components of $\mathbf{Y}^{(z)}$ can be estimated from estimates of the covariance and standard deviations, as in the motivating breast cancer studies, where the goal is to

estimate the pairwise correlations of the eight targeted genes in groups $z \in \{1, 2\}$ (i.e., IDC and ILC subtypes). For a second example, let $y_{11}, \ldots, y_{1M}$ be a fine grid of prespecified points in the support of the first component $Y_1^{(z)}$ and $\Phi_m(\mathbf{Y}^{(z)}) = \mathcal{I}(Y_1^{(z)} \leqslant y_{1m})$. For $\psi(t_1, \ldots, t_M) = t_m$, the pseudo-population CDF of $Y_1^{(z)}$ evaluated at y_{1m} equals $\psi(\boldsymbol{\lambda}^{(z)})$. Similarly, for $\psi(t_1, \ldots, t_M) = t_{m^*}$ where $m^* = \arg\min_m |t_m - 0.5|$, the approximate pseudo-population median of $Y_1^{(z)}$ is given by $\psi(\boldsymbol{\lambda}^{(z)})$.

Using the unnormalized weights $\tilde{\rho}_1, \ldots, \tilde{\rho}_N$ [defined underneath equation (3)] of a pseudo-population, we estimate $\mathbb{E}[\boldsymbol{\Phi}(\mathbf{Y}^{(z)})]$ as random vector

$$\bar{\boldsymbol{\Phi}}_z = \frac{\sum_{i=1}^N \tilde{\rho}_i \, \boldsymbol{\Phi}(\mathbf{Y}_i) \, \mathcal{I}(Z_i = z)}{\sum_{i=1}^N \tilde{\rho}_i \, \mathcal{I}(Z_i = z)}. \tag{7}$$

The following theorem and corollaries study asymptotic properties of random vector $\bar{\boldsymbol{\Phi}}_z$ as an estimator of multivariate feature $\mathbb{E}[\boldsymbol{\Phi}(\mathbf{Y}^{(z)})]$. Part 2(a) of the theorem considers a simpler situation where the MPS is known. As discussed in Mao et al. (2019) and Zeng et al. (2023), Part 2(b) considers a more realistic situation in which the MPS is estimated. The proofs are available in Web Appendix A.2 of Supplementary Materials.

THEOREM 2: *Let $\mathbb{E}_+[\cdot]$ and $\mathbb{E}[\cdot]$ respectively denote expectations with respect to the observed population and a pseudo-population of the form (2). Let observed probability $P_+[S = s]$ be strictly positive for study $s = 1, \ldots, J$. Suppose the conditional distributions of the potential outcomes are weakly unconfounded, as described in Section 2, and satisfy assumption (6). Suppose the multi-study balancing weight (3) is such that $\mathbb{E}_+[\rho^2_{\boldsymbol{\gamma},\boldsymbol{\theta}}(S, Z, \mathbf{X})]$ is finite. For $m = 1, \ldots, M$, let Φ_m be a real-valued function with domain $\mathcal{Y}$ such that $\mathbb{E}_+[\rho^2_{\boldsymbol{\gamma},\boldsymbol{\theta}}(S, Z, \mathbf{X})\Phi_m(\mathbf{Y})]$ is finite. For group $z = 1, \ldots, K$, interest focuses on the pseudo-population moment, $\mathbb{E}[\boldsymbol{\Phi}(\mathbf{Y}^{(z)})]$, also denoted by vector $\boldsymbol{\lambda}^{(z)} = (\lambda_1^{(z)}, \ldots, \lambda_M^{(z)})'$. For estimator $\bar{\boldsymbol{\Phi}}_z$ defined in (7), as $N \to \infty$:*

(1) ***Consistency****: $\bar{\boldsymbol{\Phi}}_z \xrightarrow{p} \boldsymbol{\lambda}^{(z)}$.*

(2) ***Asymptotic normality****: Consider the following situations:*

(a) ***Known MPS***: *Suppose multiple propensity score (1) is known. Let variance matrix*

$$\boldsymbol{\Sigma}_1^{(z)} = \frac{1}{\theta_z^2}\mathbb{E}_+\Big(\rho^2_{\boldsymbol{\gamma},\boldsymbol{\theta}}(S,Z,\mathbf{X})\,\mathcal{I}(Z=z)\,\big(\boldsymbol{\Phi}(\mathbf{Y}_i)-\boldsymbol{\lambda}^{(z)}\big)\big(\boldsymbol{\Phi}(\mathbf{Y}_i)-\boldsymbol{\lambda}^{(z)}\big)'\Big).$$

Then $\sqrt{N}\big(\bar{\boldsymbol{\Phi}}_z-\boldsymbol{\lambda}^{(z)}\big)\xrightarrow{d} N_M\big(\mathbf{0},\boldsymbol{\Sigma}_1^{(z)}\big)$.

(b) ***Estimated MPS***: *Suppose the MPS is estimated using multinomial logistic regression as outlined after definition (1). Let* $\hat{\boldsymbol{\omega}}$ *be the MLE of parameter* $\boldsymbol{\omega}$ *that determines the unnormalized weights* $\tilde{\rho}_1,\ldots,\tilde{\rho}_N$ *in estimator (7). We denote the variance matrix of Part 2(a) by* $\boldsymbol{\Sigma}_1^{(z)}(\boldsymbol{\omega})$ *to make explicit its dependence on* $\boldsymbol{\omega}$. *Define variance matrix* $\boldsymbol{\Sigma}_2^{(z)}(\boldsymbol{\omega})=\boldsymbol{\Sigma}_1^{(z)}(\boldsymbol{\omega})+\mathbf{D}^{(z)}(\boldsymbol{\omega})$, *where* $\mathbf{D}^{(z)}(\boldsymbol{\omega})$ *is given in Web Appendix A.2 of Supplementary Materials. Then* $\sqrt{N}\big(\bar{\boldsymbol{\Phi}}_z-\boldsymbol{\lambda}^{(z)}\big)\xrightarrow{d} N_M\big(\mathbf{0},\boldsymbol{\Sigma}_2^{(z)}(\boldsymbol{\omega})\big)$.

COROLLARY 1: *Suppose the MPS is estimated using multinomial logistic regression. Let* ψ *be a real-valued differentiable function with domain* $\mathcal{R}^M$. *Let* $\nabla\psi(\boldsymbol{\lambda})=\partial\psi(\boldsymbol{\lambda})/\partial\boldsymbol{\lambda}$ *denote the gradient vector of length* M *at* $\boldsymbol{\lambda}$. *With* $\boldsymbol{\lambda}^{(z)}=\mathbb{E}[\boldsymbol{\Phi}(\mathbf{Y}^{(z)})]$, *suppose gradient vector* $\nabla\psi(\boldsymbol{\lambda}^{(z)})$ *is non-zero at* $\boldsymbol{\lambda}^{(z)}$. *With matrix estimator* $\mathbf{G}_2^{(z)}(\boldsymbol{\omega})$ *defined as in Part 2(b) of Theorem 2, set* $\tau_z(\boldsymbol{\omega})=\nabla'\psi(\boldsymbol{\lambda}^{(z)})\,\mathbf{G}_2^{(z)}(\boldsymbol{\omega})\,\nabla\psi(\boldsymbol{\lambda}^{(z)})$. *Then* $\psi(\bar{\boldsymbol{\Phi}}_z)$ *is an asymptotically normal estimator of* $\psi(\boldsymbol{\lambda}^{(z)})$: $\sqrt{N}\big(\psi(\bar{\boldsymbol{\Phi}}_z)-\psi(\boldsymbol{\lambda}^{(z)})\big)\xrightarrow{d} N\big(0,\tau_z^2(\boldsymbol{\omega})\big)$.

Remark. Theorem 2 and its corollaries summarize several noteworthy features of estimator (7), in that it: (i) is applicable to the balancing weights of any pseudo-populations, including IC, IGO, and FLEXOR weights; (ii) generalizes plug-in sample moment estimators (Li and Li, 2019) to multiple groups and studies, while accommodating mixed-type multivariate outcomes, (iii) exploits known or researcher-supplied information about the group proportions of the pseudo-population; as mentioned, the FLEXOR weights are typically set θ_z equal to the known group prevalences of the larger population. By contrast, $\theta_z=1/K$ for most other weighting methods, and (iv) extends Mao et al. (2019) by quantifying the sampling errors in multiple group settings; matrix $\mathbf{D}^{(z)}(\boldsymbol{\omega})$ in Part 2(b) represents the

adjustment due to MPS estimation, and in the event that parameter $\boldsymbol{\omega}$ is known, this adjustment term vanishes and matrix $\boldsymbol{\Sigma}_2^{(z)}(\boldsymbol{\omega})$ of Part 2(b) equals $\boldsymbol{\Sigma}_1^{(z)}$ of Part 2(a).

Group comparisons. Consider estimation of the pseudo-population moment $\mathbb{E}[\Phi_m(\mathbf{Y}^{(z)})]$ using $\bar{\Phi}_{zm}$. Applying standard results (e.g., Johnson et al., 2002, Chapter 5), we can construct approximate $100(1-\alpha)\%$ confidence intervals *simultaneously* for all possible linear combinations of $\mathbb{E}[\Phi_m(\mathbf{Y}^{(1)})], \ldots, \mathbb{E}[\Phi_m(\mathbf{Y}^{(K)})]$. In particular, for large N, using the mth diagonal element of $\hat{\boldsymbol{\Sigma}}^{(z)}$ defined in Theorem 2, the interval $\sum_{z=1}^K a_z\,\bar{\Phi}_{zm} \pm \sqrt{\chi^2_K(\alpha)\sum_{z=1}^K a_z^2\left(s^{(z)}_{mm}\right)^2/N}$ contains $\sum_{z=1}^K a_z\,\mathbb{E}[\Phi_m(\mathbf{Y}^{(z)})]$ with approximate probability $(1-\alpha)$ simultaneously for all scalars $a_1, \ldots, a_K$. Various pseudo-population features can then be compared between the K groups. Writing $\lambda^{(zm)} = \mathbb{E}[Y_m^{(z)}]$, we could estimate $\lambda^{(1m)} - \lambda^{(2m)}$ (e.g., average difference between the mth gene's mRNA expression levels for IDC and ILC breast cancer patients) and, when $K > 2$, $\lambda^{(1m)} - \frac{1}{K-1}\sum_{z=2}^K \lambda^{(zm)}$ (e.g., for the mth gene, average difference between the mRNA expression levels for a reference group and the average of the other groups). We could also estimate ratios of means such as $\lambda^{(zm)}/\lambda^{(1m)}$, ratios of mean differences such as $\left(\lambda^{(3m)} - \lambda^{(1m)}\right)/\left(\lambda^{(2m)} - \lambda^{(1m)}\right)$, group-specific standard deviations, percentiles, ratios of medians, and ratios of coefficients of variation. Under mild conditions, these estimators are consistent and asymptotically normal, and their asymptotic variances are available by applying Corollary 1 and the delta method. If N_z is small for some groups, such as rare or undersampled treatments, the asymptotic confidence intervals may not have proper coverage and we could employ bootstrap methods to construct confidence intervals.

In single studies ($J = 1$), Hirano et al. (2003) and Zeng et al. (2023) have shown that treating IPWs as known counter-intuitively overestimates the variance of pairwise group mean comparisons. However, with multiple studies and arbitrary functions of group-specific features $\psi(\boldsymbol{\lambda}^{(1)}), \ldots, \psi(\boldsymbol{\lambda}^{(K)})$, this is not generally guaranteed because matrix $\mathbf{D}^{(z)}(\boldsymbol{\omega})$ of Theorem 2 may be neither positive nor negative definitive for a general pseudo-population (2).

4. Simulation Study

We used simulated datasets to evaluate different weighting strategies for inferring the population-level features of two subject groups and assessed the accuracy of the Section 3 asymptotic variance expression for the mean group differences. Mimicking the motivating TCGA breast cancer studies, we simulated $R = 500$ independent datasets, each consisting of $J = 7$ observational studies, $K = 2$ groups, and $L = 1$ (i.e., univariate) outcomes for $\tilde{N}$ subjects whose covariate vectors were sampled with replacement from the $N = 450$ TCGA breast cancer patients. We first took $\tilde{N} = 500$ subjects in two simulation scenarios, labeled "high" and "low," to represent the relative degrees of covariate similarity or balance among the $JK = 14$ study-group combinations; in other words, the low similarity scenario represented higher confounding levels. We then applied the Section 3 procedure to meta-analyze the four studies in each artificial dataset. Additionally, by increasing $\tilde{N}$ from 125, to 250, and then to 500 subjects, we compared the asymptotic and bootstrap-based variances of the group mean difference, $(\lambda^{(1)} - \lambda^{(2)})$, where $\lambda^{(z)} = \mathbb{E}[Y^{(z)}]$.

As a common initial step to all 500 artificial datasets, we performed k-means clustering of the covariates, $\mathbf{X}_1, \ldots, \mathbf{X}_N$, of the TCGA datasets and detected lower-dimensional structure by aggregating them into $Q = 10$ clusters with centers $\mathbf{q}_1, \ldots, \mathbf{q}_Q \in \mathcal{R}^p$ and $m_1, \ldots, m_Q$ allocated number of covariates. Independently for the artificial datasets $r = 1, \ldots, 500$ comprising $\tilde{N}$ patients each, we generated the data as follows:

(1) **Natural population** Generate cluster relative weights, $\boldsymbol{\pi}^{(r)} = (\pi_1^{(r)}, \ldots, \pi_Q^{(r)}) \sim \mathcal{D}_Q(\mathbf{1}_Q)$, denoting the Dirichlet distribution on the unit simplex $\mathcal{S}_Q$ and $\mathbf{1}_Q$ representing the vector of Q ones. Let the number of patients in the large natural population be $N_0 = 10^6$. For the natural population patients, sample their cluster memberships from the mixture distribution of integers: $c_{ir}^{(0)} \overset{\text{i.i.d.}}{\sim} \sum_{u=1}^{Q} \pi_u^{(r)} \zeta_u$ where ζ_u represents a point mass at u. Thence, pick covariate $\mathbf{x}_{ir}^{(0)}$ uniformly from the $m_{c_{ir}^{(0)}}$ TCGA covariates allocated previously to the $c_{ir}^{(0)}$th k-means

cluster. Generate the "known" relative group proportions in the natural population: $\boldsymbol{\theta}^{(r)} \sim \mathcal{D}_K(\mathbf{1}_K)$, for $K = 2$ groups. Fix the association between group memberships and covariates: $\delta_z^{(r)}(\mathbf{x}) \propto 1$ if $z = 1$ and $\delta_z^{(r)}(\mathbf{x}) \propto \exp\big(\omega_0^{(r)} + \omega_1^{(r)} \sum_{t=1}^{p} x_t / \frac{1}{N_0} \sum_{i=1}^{N_0} \sum_{t=1}^{p} x_{irt}^{(0)}\big)$ if $z = 2$. Here, $\omega_1^{(r)}$ equals 1 and 0.1 in the high and low similarity scenarios respectively, with $\omega_0^{(r)}$ chosen so that $\delta_z^{(r)}(\mathbf{x}_{ir}^{(0)})$, averaged over the natural population, equals $\theta_z^{(r)}$.

(2) **Covariates** For subject $i = 1, \ldots, \tilde{N}$, sample covariate vector $\tilde{\mathbf{x}}_i^{(r)} = (\tilde{x}_{i1}^{(r)}, \ldots, \tilde{x}_{ip}^{(r)})'$ with replacement from the $N = 450$ TCGA covariate vectors.

(3) **Study and group memberships** Study $s_i^{(r)}$ and group $z_i^{(r)}$ were generated as follows:

(a) *Multiple propensity score* Define the group-specific study propensities as follows: $\log\big(\delta_{S=s|Z=z}(\mathbf{x}) / \delta_{S=1|Z=z}(\mathbf{x})\big) = sz\omega^{(r)} \sum_{t=1}^{p} \tilde{x}_{it}^{(r)} / \frac{1}{\tilde{N}} \sum_{i'=1}^{\tilde{N}} \sum_{t=1}^{p} \tilde{x}_{i't}^{(r)}$ for $s = 2, \ldots, J$ and $z = 1, 2$. We set similarity parameter $\omega^{(r)}$ equal to 0.5 (0.05) in the high and low similarity scenario, respectively. Assuming the same group PS as the natural population, the MPS is available as $\delta_{sz}(\mathbf{x}) = \delta_{s|z}(\mathbf{x})\delta_z(\mathbf{x})$. For patient $i = 1, \ldots, \tilde{N}$, evaluate their probability vector $\boldsymbol{\delta}^{(r)}(\mathbf{x}_i) = \big(\delta_{11}^{(r)}(\mathbf{x}_i), \ldots, \delta_{JK}^{(r)}(\mathbf{x}_i)\big)$

(b) *Study-group memberships* For patient $i = 1, \ldots, \tilde{N}$, generate $(s_i^{(r)}, z_i^{(r)})$ from the categorical distribution with parameter $\boldsymbol{\delta}^{(r)}(\mathbf{x}_i)$.

(4) **Subject-specific observed outcomes** Generate $Y_i^{(r)} \mid Z_i = z_i^{(r)} \overset{\text{indep}}{\sim} N\big(z_i^{(r)} \sum_{t=1}^{p} \tilde{x}_{it}^{(r)}, \tau_r^2\big)$, with τ_r^2 chosen to achieve an approximate R-squared of 0.9.

Subsequently, we disregarded knowledge of all simulation parameters and analyzed each artificial dataset using the proposed methods. As discussed in Section 3, during Stage 1 of the inferential procedure, we estimated the MPS of each dataset. We then evaluated the unnormalized balancing weights, $\tilde{\rho}_1, \ldots, \tilde{\rho}_{\tilde{N}}$, for the IC, IGO, and FLEXOR pseudo-populations. The computational costs of evaluating the FLEXOR weights were negligible.

[Table 1 about here.]

Define *percent ESS* as the ESS for 100 participants. For $\tilde{N} = 500$ subjects, Table 1 presents summaries of the percent ESS of the FLEXOR, IGO, and IC pseudo-populations in the low and high similarity scenarios. Unsurprisingly, all three pseudo-populations had substantially higher ESS in the less challenging high similarity scenario in which the covariates were almost balanced even before applying the weighting methods. In both scenarios, the IC and IGO pseudo-populations had similar ESS and a median ESS of approximately 32% (74%) in the low (high) simulation scenarios. The FLEXOR pseudo-population had substantially higher ESS in every dataset and scenario, and median ESS of 87.26% (95.19%%) in the low (high) scenarios corresponding to 436.3 and 475.95 subjects, respectively.

We applied the Section 3 strategy to make weighted inferences about functionals of the group-specific means $\lambda^{(z)}$ and standard deviations $\sigma^{(z)}$ of the zth group's potential outcomes. The sufficient conditions of Theorem 2 and its corollaries are satisfied by the potential outcome features and pseudo-populations. Since the estimands depend on the pseudo-population, we evaluated each estimator's accuracy relative to the true value of its corresponding estimand computed using Monte Carlo methods.

[Table 2 about here.]

For various estimands and both similarity scenarios, and averaging over the 500 artificial datasets comprising $\tilde{N} = 500$ subjects each, Table 2 displays the absolute biases, variances, and coverages of nominally 95% confidence intervals for the FLEXOR, IGO, and IC pseudo-populations. For each artificial dataset and weighting method, the three performance measures were estimated using 500 independent bootstrap samples. For each potential outcome feature (row), and separately for the absolute bias and variance performance measures (three-column block), a pseudo-population (column) is marked in bold if it significantly outperforms the other competing pseudo-populations. In general, the IGO and IC weights had comparable performances for these data. The three methods had somewhat similar

accuracies and reasonable coverages in the high similarity scenario where the covariates were almost balanced across the study-group combinations. However, in the more realistic and challenging low similarity simulation scenario, the best results typically corresponded to the FLEXOR pseudo-population, which often substantially outperformed the other methods. Somewhat unexpectedly, this included the mean group difference ($\lambda^{(1)} - \lambda^{(2)}$), for which IGO weights are theoretically optimal under additional assumptions such as homoscedasticity (see Li and Li, 2019, for single studies); the simulation mechanism did not comply with these sufficient conditions. The results demonstrate the advantages of the FLEXOR strategy which focuses on stabilizing the balancing weights rather than inferences about specific estimands.

Finally, we compared the bootstrap-based and asymptotic variances of estimator (7) for unconfounded inferences about the mean group difference, $\lambda^{(1)} - \lambda^{(2)}$. For an increasing number of subjects, i.e., $\tilde{N} = 125$, 250, and 500, we generated 500 artificial datasets in the high and low similarity scenarios. For any dataset, the asymptotic variance of weighted estimator $\hat{\lambda}^{(1)} - \hat{\lambda}^{(2)}$ is available by applying Theorem 2 and the subsequently discussed group comparison strategies. This theoretical limiting value can be compared to the variance estimate based on $B = 500$ bootstrap samples. Web Table 1 of Supplementary Materials compares these numbers for the simulation scenarios and sample sizes. We find that when the sample size is relatively small (i.e., $\tilde{N} \leqslant 250$), there is a substantial difference between the asymptotic and bootstrap-based variances. This difference indicates that a sufficiently large number of samples may be required for the asymptotic variance to be reliable. However, for $\tilde{N} = 500$ subjects, the two variances match very well, giving us the confidence to use asymptotic variances in the TCGA data analysis with a comparable number of patients.

5. Data Analysis

To understand breast cancer oncogenesis, we analyzed the $J = 7$ motivating TCGA studies using mRNA expression measurements on $L = 8$ targeted genes and $p = 30$ demographic

and clinicopathological covariates for $N = 450$ patients. The participants are partitioned into $K = 2$ groups determined by cancer subtypes IDC and ILC, constituting approximately 80% and 10% of U.S. breast cancer cases (Wright, 2022; Tran, 2022); the study-specific percentages in Web Table 2 of Supplementary Materials are significantly different.

The ESS of the IC weights was 25.7% or 115.7 patients. The IGO weights had a similar ESS of 26.4% or 118.7 patients. The FLEXOR population had a higher ESS of 40.9% or 183.9 patients, while also guaranteeing that the weight-adjusted composition of IDC and ILC patients in each TCGA study matched the composition of U.S. breast cancer patients. Applying the Section 3 procedure, we estimated population-level functionals of the group potential outcomes for the FLEXOR, IC, and IGO pseudo-populations. For example, for the lth biomarker, the group-specific mean $\lambda_l^{(z)}$ and standard deviation $\sigma^{(z)}$ were estimated by setting $\mathbf{\Phi}(\mathbf{y}) = (y_l, y_l^2)'$ in Theorem 2 and $\psi(t_1, t_2) = \sqrt{t_2 - t_1^2}$ in Corollary 1. Median $M_l^{(z)}$ was estimated by first estimating the CDF of potential outcome $Y_l^{(z)}$ for a fine grid of points. Group comparison estimands like $\lambda_l^{(1)} - \lambda_l^{(2)}$ and $\sigma^{(1)}/\sigma^{(2)}$ were estimated by applying appropriately defined functionals to the estimates of $\lambda_l^{(1)}$, $\lambda_l^{(2)}$, $\sigma_l^{(1)}$, and $\sigma_l^{(2)}$. The estimate and 95% confidence interval based on $B = 100$ bootstrap samples are displayed in Table 3 for each feature (row), pseudo-population (column), and genes COL9A3, CXCL12, IGF1, and ITGA11 (block). The results for the genes IVL, LEF1, IC, and SMR3B are displayed in Web Table 3 of Supplementary Materials. For each gene-estimand combination, a confidence interval for the IC or IGO pseudo-population is marked in bold whenever the FLEXOR pseudo-population's confidence interval was *narrower*; we find that the FLEXOR pseudo-population often provided the most precise (narrowest) confidence intervals.

For FLEXOR, the confidence intervals for $\lambda_l^{(1)} - \lambda_l^{(2)}$ reveal that the mean potential outcomes were significantly different between the disease subtypes for genes CXCL12, IGF1, LEF1, PRB2, and SMR3B. Additionally, the standard deviation of the IDC and IDL poten-

tial outcomes for FLEXOR were substantially different for the genes COL9A3, PRB2, and IVL; the respective confidence intervals for $\sigma_l^{(1)}/\sigma_l^{(2)}$ excluded 1. If required, the group-specific medians could be compared by inferences on $M_l^{(1)}/M_l^{(2)}$ or $M_l^{(1)} - M_l^{(2)}$.

[Table 3 about here.]

Next, we estimated the correlation between the potential outcomes of the l_1th and l_2th biomarker in the zth group: for $M = 3$ and $\mathbf{y} \in \mathcal{R}^8$, we assumed an M-variate function, $\mathbf{\Phi}(\mathbf{y}) = (\Phi_1(\mathbf{y}), \Phi_2(\mathbf{y}), \Phi_3(\mathbf{y}))'$, with component functions, $\Phi_1(\mathbf{y}) = y_{l_1}$, $\Phi_2(\mathbf{y}) = y_{l_3}$, and $\Phi_3(\mathbf{y}) = y_{l_1} y_{l_2}$. For the zth group, we estimated $\boldsymbol{\lambda}^{(z)} = \left(\mathbb{E}[Y_{l_1}^{(z)}], \mathbb{E}[Y_{l_2}^{(z)}], \mathbb{E}[Y_{l_1}^{(z)} Y_{l_2}^{(z)}]\right)'$ for a pseudo-population by applying Theorem 2. Setting $\psi(t_1, t_2, t_3) = t_3 - t_1 t_2$, we then applied Corollary 1 to estimate pseudo-population covariance, $\psi(\boldsymbol{\lambda}^{(z)}) = \text{cov}(Y_{l_1}^{(z)}, Y_{l_2}^{(z)})$. Using the estimated standard deviations $\sigma_{l_1}^{(z)}$ and $\sigma_{l_2}^{(z)}$ for the pseudo-population, as described above, we estimated the correlation. Independent estimates from $B = 100$ bootstrap samples were used to compute 95% confidence intervals of the true correlation between the l_1th and l_2th gene pair in the zth group. Web Tables 4-6 of Supplementary Materials present 95% confidence intervals of the group-specific correlations for each gene pair and weighting method.

[Table 4 about here.]

Table 4 lists the significantly correlated gene pairs for each disease subtype. For the FLEXOR pseudo-population and IDC disease subtype, gene CXCL12 was significantly co-expressed with the IGF1, ITGA11, and LEF1; gene IGF1 was co-expressed with ITGA11 and LEF1; gene COL9A3 was co-expressed with LEF1 and PRB2; and gene LEF1 was co-expressed with IVL and ITGA11. For disease subtype ILC, only the CXCL12 - IGF1 gene pair was significantly correlated according to FLEXOR. The differential correlation pattern for the FLEXOR pseudo-population was, therefore, the gene pairs (CXCL12, ITGA11), (IGF1, ITGA11), (COL9A3, LEF1), (CXCL12, LEF1), (IGF1, LEF1), (ITGA11, LEF1), (IVL, LEF1), and (COL9A3, PRB2). Detecting these variations in gene co-expression patterns

between the IDC and ILC subtypes of breast cancer patients in the United States is crucial for informing precision medicine and targeted therapies (Schmidt et al., 2016).

[Figure 1 about here.]

By contrast, Table 4 shows that the differential correlation pattern of the IC pseudo-population comprised just five gene pairs, and was identical to the IGO pseudo-population's pattern. Although these gene pairs were also detected by the FLEXOR pseudo-population, the latter detected additional co-expressed gene pairs. Figure 1 graphically summarizes the number of differentially correlated gene pairs discovered by the weighting methods and (biased) unadjusted analyses. Recent literature on breast cancer gene ontology substantiates the distinctive findings of FLEXOR. The genes IVL and LEF1 are highly expressed in basal and metaplastic human breast cancers, and the cell adhesion and ECM receptor pathways, containing the genes ITGA11 and LEF1, are deregulated (Williams et al., 2022). The focal adhesion and cell cycle pathways, containing the genes COL9A3 and LEF1, are affected by WNT signaling gene set mutations caused by breast cancer metastases (Paul, 2020).

6. Conclusion

In multiple retrospective cohorts, the integrative analysis of mixed-type multivariate outcomes to accurately compare multiple groups is a challenging problem. We formulate new frameworks for covariate-balanced pseudo-populations that extend existing weighting methods to meta-analytical investigations and design a novel, estimand-agnostic FLEXOR pseudo-population that maximizes the effective sample size by a cost-effective iterative procedure. We propose generally applicable weighted estimators for a wide variety of population-level univariate or multivariate features relevant to multigroup comparisons, e.g., contrasts and ratios of means, medians, and standard deviations, and correlation coefficients.

The methodology has a wide range of meta-analytical applications, including multi-arm

RCTs. A component of the multi-study balancing weights is considerably simplified if the s^*th study is an RCT, in which case the study-specific group MPS $\delta_{z|s^*}(\mathbf{x})$ equals $1/K$. In general, the theoretical results hold for a mix of observational studies and RCTs, although the study MPS must still be estimated because the subject-study allocations are usually non-random for multiple studies. The methodology may be generalized in several other directions, such as increased efficiency by adding an outcome modeling component (Mao et al., 2019; Zeng et al., 2023); transportability (Westreich et al., 2017) and data-fusion (Bareinboim and Pearl, 2016; Dahabreh et al., 2023) problems, which incorporate additional information in the form of random samples from the natural population; flexible machine learning methods for MPS estimation that achieve achieving $\sqrt{N}$ inference (Chernozhukov et al., 2018). Furthermore, in recent years, weighting approaches are frequently challenged and rendered ineffectual by high-dimensional genetic or genomic measurements. Our future research will explore these avenues.

Acknowledgements

This work was supported by NSF and NIH. We thank the Editor, AE, and two referees for many insightful comments that improved the content and presentation of the paper.

Supplementary Materials

Web Appendices, Tables, and Figures, and code referenced in Sections 1-5 are available with this paper at the Biometrics website on Oxford Academic.

Data Availability

The data that support the findings in this paper are openly available in The Cancer Genome Atlas (TCGA) portal at https://www.cancer.gov/tcga.

References

Austin, P. C. (2010). The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational

studies. *Statistics in Medicine* **29,** 2137–2148.

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* **113,** 7345–7352.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Christopoulos, P. F., Msaouel, P., and Koutsilieris, M. (2015). The role of the insulin-like growth factor-1 system in breast cancer. *Molecular Cancer* **14,** 1–14.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2006). Moving the goalposts: addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical report, National Bureau of Economic Research.

Dahabreh, I. J., Robertson, S. E., Petito, L. C., Hernán, M. A., and Steingrimsson, J. A. (2023). Efficient and robust methods for causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a target population. *Biometrics* **79,** 1057–1072.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71,** 1161–1189.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87,** 706–710.

Johnson, R. A., Wichern, D. W., et al. (2002). *Applied Multivariate Statistical Analysis*. Prentice hall Upper Saddle River, NJ.

Kumar, B., Chand, V., Ram, A., Usmani, D., and Muhammad, N. (2020). Oncogenic mutations in tumorigenesis and targeted therapy in breast cancer. *Current Molecular Biology Reports* **6,** 116–125.

Li, F. and Li, F. (2019). Propensity score weighting for causal inference with multiple

treatments. *The Annals of Applied Statistics* **13,** 2389–2415.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* **113,** 390–400.

Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *The International Journal of Biostatistics* **9,** 215–234.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* **23,** 2937–2960.

Mao, H., Li, L., and Greene, T. (2019). Propensity score weighting analysis and treatment effect discovery. *Statistical methods in medical research* **28,** 2439–2454.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine* **32,** 3388–3414.

NCI (2022). Genomic data commons data portal. https://portal.gdc.cancer.gov/.

Paul, M. R. (2020). *The Genomic Evolution of Breast Cancer Metastasis.* PhD thesis, University of Pennsylvania.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11,** 550–560.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90,** 122–129.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine* **26,** 20–36.

Schmidt, K. T., Chau, C. H., Price, D. K., and Figg, W. D. (2016). Precision oncology

medicine: the clinical relevance of patient-specific biomarkers used to optimize cancer treatment. *The Journal of Clinical Pharmacology* **56,** 1484–1499.

Surveillance Research Program, NCI (2023). SEER*Explorer: An interactive website for SEER cancer statistics [Internet]. Available from https://seer.cancer.gov/statistics-network/explorer/.

Tran, H.-T. (2022). Invasive lobular carcinoma. https://www.hopkinsmedicine.org/health/conditions-and-diseases/breast-cancer/invasive-lobular-carcinoma.

Wang, C. and Rosner, G. L. (2019). A Bayesian nonparametric causal inference model for synthesizing randomized clinical trial and real-world evidence. *Statistics in Medicine* **38,** 2573–2588.

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., and Cole, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology* **186,** 1010–1014.

Williams, R., Jobling, S., Sims, A. H., Mou, C., Wilkinson, L., Collu, G. M., Streuli, C. H., Gilmore, A. P., Headon, D. J., and Brennan, K. (2022). Elevated edar signalling promotes mammary gland tumourigenesis with squamous metaplasia. *Oncogene* **41,** 1040–1049.

Wright, P. (2022). Invasive ductal carcinoma. https://www.hopkinsmedicine.org/health/conditions-and-diseases/breast-cancer/invasive-ductal-carcinoma-idc.

Zeng, S., Li, F., Hu, L., and Li, F. (2023). Propensity score weighting analysis of survival outcomes using pseudo-observations. *Statistica Sinica* **33,** 2161–2184.

Received October 2007. *Revised February* 2008. *Accepted March* 2008.

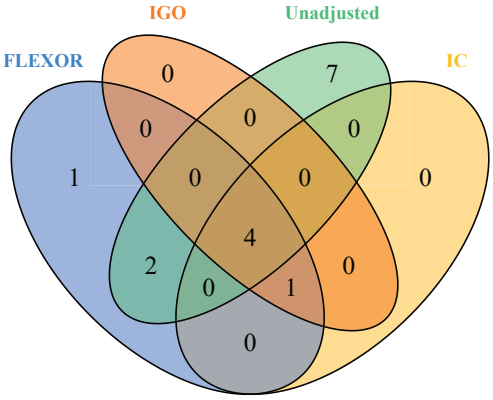

Figure 1: Venn diagram of the differential correlation pattern of the targeted gene pairs for the three weighting methods and unweighted analysis.

	Low similarity			High similarity		
	FLEXOR	**IGO**	**IC**	**FLEXOR**	**IGO**	**IC**
Minimum	78.37	20.52	20.61	85.81	30.80	30.62
First quartile	85.51	29.91	29.83	94.08	55.67	55.61
Median	87.26	32.07	31.92	95.19	73.73	73.70
Mean	87.20	31.97	31.87	95.03	70.21	69.90
Third quartile	89.02	34.46	34.50	96.17	86.09	85.62
Maximum	93.92	42.56	42.87	98.59	94.52	94.61

Table 1: For the 500 simulated datasets, percentage ESS summaries for the three pseudo-populations in the low and high simulation scenarios with $\tilde{N} = 500$ subjects.

Low similarity scenario									
	Absolute bias $\times 10^2$			***Standard deviation*** $\times 10$			***Coverage*** (%)		
Estimand	**FLEXOR**	**IGO**	**IC**	**FLEXOR**	**IGO**	**IC**	**FLEXOR**	**IGO**	**IC**
$\lambda^{(1)}$	**2.9** (0.1)	4.1 (0.1)	3.8 (0.1)	**2.9** (0.0)	3.6 (0.0)	3.4 (0.0)	97	93	95
$\lambda^{(2)}$	**4.5** (0.1)	8.4 (0.2)	6.6 (0.1)	**4.6** (0.1)	6.5 (0.1)	5.7 (0.1)	98	88	89
$\sigma^{(1)}$	**2.6** (0.1)	3.0 (0.1)	3.1 (0.1)	2.6 (0.1)	2.6 (0.0)	2.8 (0.1)	95	90	90
$\sigma^{(2)}$	**4.4** (0.1)	6.0 (0.1)	6.2 (0.1)	**3.8** (0.1)	4.5 (0.1)	4.8 (0.1)	93	89	89
$\lambda^{(1)} - \lambda^{(2)}$	**4.6** (0.1)	7.9 (0.2)	7.4 (0.2)	**4.4** (0.1)	6.1 (0.0)	6.1 (0.0)	96	89	90
High similarity scenario									
	Absolute bias $\times 10^2$			***Standard deviation*** $\times 10$			***Coverage*** (%)		
$\lambda^{(1)}$	2.8 (0.1)	3.3 (0.1)	2.7 (0.1)	2.8 (0.0)	3.2 (0.0)	2.8 (0.0)	97	96	96
$\lambda^{(2)}$	4.4 (0.1)	5.7 (0.1)	4.2 (0.1)	4.5 (0.0)	5.5 (0.0)	4.6 (0.0)	97	95	97
$\sigma^{(1)}$	3.0 (0.1)	2.9 (0.1)	2.7 (0.1)	2.2 (0.0)	2.3 (0.0)	2.5 (0.0)	94	94	94
$\sigma^{(2)}$	6.1 (0.2)	5.9 (0.1)	**5.2** (0.1)	3.9 (0.0)	4.1 (0.1)	4.5 (0.1)	93	94	94
$\lambda^{(1)} - \lambda^{(2)}$	4.3 (0.1)	4.7 (0.1)	4.5 (0.1)	**4.1** (0.0)	4.4 (0.0)	4.4 (0.0)	97	95	96

Table 2: In the two simulation scenarios with $\tilde{N} = 500$ subjects each, averaging over 500 artificial datasets, the absolute biases, variances, and 95% confidence interval coverages of some potential outcome features for the FLEXOR, IGO, and IC pseudo-populations. For each artificial dataset and weighting method, the three performance measures were estimated using 500 independent bootstrap samples. The estimated standard errors are displayed in parentheses. For each feature (row), and separately for the absolute bias and variance measures (three-column block), a weighting strategy (column) is marked in bold if it significantly outperforms the other two strategies.

Estimand	**FLEXOR**	**IC**	**IGO**
COL9A3 ($l = 1$)			
$\lambda_l^{(1)}$	-0.05 $(-0.27, 0.23)$	-0.11 $(\mathbf{-0.33, 0.21})$	-0.09 $(-0.25, 0.22)$
$\lambda_l^{(2)}$	-0.11 $(-0.36, 0.20)$	-0.16 $(\mathbf{-0.44, 0.21})$	-0.19 $(\mathbf{-0.49, 0.27})$
${\sigma_l}^{(1)}$	1.03 $(0.87, 1.26)$	0.97 $(\mathbf{0.84, 1.31})$	0.93 $(\mathbf{0.83, 1.36})$
${\sigma_l}^{(2)}$	0.68 $(0.54, 0.86)$	0.69 $(\mathbf{0.47, 0.90})$	0.69 $(\mathbf{0.49, 0.91})$
$\mathrm{M}_l^{(1)}$	-0.19 $(-0.47, 0.07)$	-0.23 $(\mathbf{-0.46, 0.13})$	-0.23 $(-0.40, 0.11)$
$\mathrm{M}_l^{(2)}$	-0.05 $(-0.52, 0.37)$	0.07 $(\mathbf{-0.57, 0.49})$	-0.05 $(\mathbf{-0.54, 0.48})$
$\lambda_l^{(1)} - \lambda_l^{(2)}$	0.06 $(-0.33, 0.43)$	0.05 $(\mathbf{-0.48, 0.54})$	0.09 $(\mathbf{-0.36, 0.59})$
$\sigma_l^{(1)}/\sigma_l^{(2)}$	1.52 $(1.15, 2.09)$	1.41 $(\mathbf{1.11, 2.24})$	1.35 $(\mathbf{1.09, 2.24})$
CXCL12 ($l = 2$)			
Estimand	**FLEXOR**	**IC**	**IGO**
$\lambda_l^{(1)}$	-0.03 $(-0.22, 0.21)$	-0.03 $(\mathbf{-0.23, 0.29})$	0.02 $(\mathbf{-0.31, 0.22})$
$\lambda_l^{(2)}$	0.59 $(0.23, 0.88)$	0.55 $(\mathbf{0.11, 1.01})$	0.58 $(\mathbf{0.26, 1.01})$
${\sigma_l}^{(1)}$	0.91 $(0.84, 1.16)$	0.97 $(\mathbf{0.84, 1.21})$	0.94 $(\mathbf{0.83, 1.21})$
${\sigma_l}^{(2)}$	0.80 $(0.52, 1.10)$	0.83 $(\mathbf{0.49, 1.27})$	0.82 $(\mathbf{0.54, 1.20})$
$\mathrm{M}_l^{(1)}$	-0.15 $(-0.20, 0.36)$	-0.16 $(\mathbf{-0.32, 0.36})$	-0.09 $(\mathbf{-0.35, 0.38})$
$\mathrm{M}_l^{(2)}$	0.68 $(0.44, 1.01)$	0.69 $(\mathbf{0.10, 1.16})$	0.58 $(\mathbf{0.21, 1.08})$
$\lambda_l^{(1)} - \lambda_l^{(2)}$	-0.62 $(-1.00, -0.12)$	-0.58 $(\mathbf{-1.18, -0.08})$	-0.56 $(\mathbf{-1.08, -0.17})$
$\sigma_l^{(1)}/\sigma_l^{(2)}$	1.14 $(0.87, 1.99)$	1.17 $(\mathbf{0.71, 2.32})$	1.14 $(\mathbf{0.75, 1.94})$
IGF1 ($l = 3$)			
Estimand	**FLEXOR**	**IC**	**IGO**
$\lambda_l^{(1)}$	0.04 $(-0.21, 0.23)$	0.10 $(\mathbf{-0.28, 0.31})$	0.13 $(\mathbf{-0.30, 0.30})$
$\lambda_l^{(2)}$	0.82 $(0.54, 1.09)$	0.84 $(\mathbf{0.52, 1.16})$	0.82 $(\mathbf{0.52, 1.17})$
${\sigma_l}^{(1)}$	0.81 $(0.80, 1.12)$	0.86 $(\mathbf{0.78, 1.18})$	0.87 $(\mathbf{0.75, 1.16})$
${\sigma_l}^{(2)}$	0.76 $(0.47, 0.95)$	0.82 $(\mathbf{0.45, 1.13})$	0.76 $(\mathbf{0.45, 1.02})$
$\mathrm{M}_l^{(1)}$	-0.01 $(-0.18, 0.34)$	0.06 $(\mathbf{-0.23, 0.47})$	0.10 $(\mathbf{-0.26, 0.45})$
$\mathrm{M}_l^{(2)}$	0.95 $(0.60, 1.22)$	0.95 $(\mathbf{0.43, 1.28})$	0.88 $(\mathbf{0.52, 1.32})$
$\lambda_l^{(1)} - \lambda_l^{(2)}$	-0.77 $(-1.22, -0.43)$	-0.74 $(\mathbf{-1.19, -0.33})$	-0.69 $(\mathbf{-1.18, -0.31})$
$\sigma_l^{(1)}/\sigma_l^{(2)}$	1.06 $(0.94, 2.03)$	1.05 $(\mathbf{0.83, 2.33})$	1.14 $(\mathbf{0.82, 2.26})$
ITGA11 ($l = 4$)			
Estimand	**FLEXOR**	**IC**	**IGO**
$\lambda_l^{(1)}$	0.01 $(-0.28, 0.22)$	0.03 $(\mathbf{-0.37, 0.24})$	0.07 $(-0.29, 0.17)$
$\lambda_l^{(2)}$	0.01 $(-0.48, 0.26)$	-0.02 $(\mathbf{-0.53, 0.27})$	0.07 $(\mathbf{-0.63, 0.28})$
${\sigma_l}^{(1)}$	0.92 $(0.83, 1.10)$	0.96 $(\mathbf{0.80, 1.16})$	0.94 $(\mathbf{0.83, 1.19})$
${\sigma_l}^{(2)}$	0.81 $(0.60, 1.03)$	0.93 $(\mathbf{0.54, 1.07})$	0.98 $(\mathbf{0.56, 1.15})$
$\mathrm{M}_l^{(1)}$	0.14 $(-0.28, 0.41)$	0.19 $(\mathbf{-0.49, 0.48})$	0.20 $(\mathbf{-0.36, 0.39})$
$\mathrm{M}_l^{(2)}$	-0.02 $(-0.54, 0.32)$	-0.22 $(\mathbf{-0.72, 0.26})$	-0.09 $(\mathbf{-0.55, 0.35})$
$\lambda_l^{(1)} - \lambda_l^{(2)}$	0.01 $(-0.28, 0.49)$	0.05 $(\mathbf{-0.41, 0.56})$	0.00 $(\mathbf{-0.44, 0.60})$
$\sigma_l^{(1)}/\sigma_l^{(2)}$	1.14 $(0.89, 1.62)$	1.03 $(\mathbf{0.86, 2.01})$	0.96 $(\mathbf{0.82, 1.84})$

Table 3: For four targeted genes, estimates and 95% bootstrap confidence levels (shown in parenthesis) of different population-level estimands of the potential outcomes of group 1 (IDC cancer subtype, denoted by superscript 1) and group 2 (ILC cancer subtype, denoted by superscript 2) with FLEXOR, IC, and IGO weights. An IC or IGO confidence interval is highlighted in bold if it is *wider* than the FLEXOR confidence interval. All numbers are rounded to 2 decimal places. See Section 5 for further explanation.

Infiltrating Ductal Carcinoma	
Pseudo-population	**Significantly correlated gene pairs**
FLEXOR	CXCL12-IGF1, CXCL12-ITGA11, IGF1-ITGA11, COL9A3-LEF1, CXCL12-LEF1, IGF1-LEF1, ITGA11-LEF1, IVL-LEF1, COL9A3-PRB2
IC	CXCL12-IGF1, CXCL12-ITGA11, IGF1-ITGA11, CXCL12-LEF1, IGF1-LEF1, COL9A3-PRB2
IGO	CXCL12-IGF1, CXCL12-ITGA11, IGF1-ITGA11, CXCL12-LEF1, IGF1-LEF1, COL9A3-PRB2
Infiltrating Lobular Carcinoma	
Pseudo-population	**Significantly correlated gene pairs**
FLEXOR	CXCL12-IGF1
IC	CXCL12-IGF1
IGO	CXCL12-IGF1

Table 4: Co-expressed gene pairs for each pseudo-population and breast cancer subtype.