Data Integration with Oracle Use of External Information from Heterogeneous Populations

Yuqi Zhai and Peisong Han

Department of Biostatistics, University of Michigan, Ann Arbor, MI

ABSTRACT

It is common to have access to summary information from external studies. Such information can be useful for an internal study of interest to improve parameter estimation efficiency when incorporated. However, external studies may target populations different from the internal study, in which case an incorporation of the corresponding information may introduce estimation bias. We develop a penalized constrained maximum likelihood (PCML) method that simultaneously (a) selects the external studies whose information is useful for internal model fitting and (b) incorporates the corresponding information into internal estimation. The PCML estimator has the same efficiency as an oracle estimator that fully incorporates the useful external information alone. We establish estimation consistency, parametric rate of convergence, external information selection consistency, asymptotic normality, and oracle efficiency. An algorithm for implementation is provided, together with a data-adaptive tuning parameter selection. Supplemental materials are available online containing some details referred to throughout the article.

1. Introduction

Data integration has become an active research area due to increasing availability of data from many sources in the era of data science. Data from different sources oftentimes contain information that can help make a better decision or a more accurate conclusion compared to using any single data source alone, if the information is properly incorporated. For instance, in survey sampling, although a probability-based sample is desired to ensure the representativeness of the population, the financial cost in practice may limit the sample size of the probability-based sample. On the other hand, nonprobability samples may be easy to obtain with large sizes and can then be used to improve the precision and/or accuracy for estimation. Another example is that, in genetics research, evidence from multiple genome-wide association studies can be integrated to help better identify genetic determinants for diseases, whereas a single study alone may not have the desired power.

Statistical methods for data integration vary depending on many factors, including the types of information to be combined. Methods that can deal with summary or aggregate information are particularly attractive because of their less demand on data sharing and data storage, as well as ethical considerations such as maintaining confidentiality and privacy of study participants. The need for statistical methods that can incorporate summary information from external studies into an internal analysis arises in many areas. Summary information has become widely available. For instance, in survey sampling aggregate information such as stratified population means is oftentimes available from published census reports, and in biomedical and public health research aggregate information such as demographic distributions and model fitting results is oftentimes available from published articles. Our development in this article is under such a setting to incorporate useful summary information from external studies to improve estimation efficiency for an internal study that has individual-level data.

Some authors have considered similar settings, but most of them made the assumption that the internal and external study populations are the same (e.g., Imbens and Lancaster 1994; Wu and Sitter 2001; Chen, Sitter, and Wu 2002; Chaudhuri, Handcock, and Rendall 2008; Qin et al. 2015; Chatterjee et al. 2016; Huang, Qin, and Tsai 2016; Cheng et al. 2019; Gu et al. 2019; Han and Lawless 2019; Huang and Qin 2020; Zhang et al. 2020; Han, Taylor, and Mukherjee 2022). Such an assumption can be easily violated in practice since, for example, the demographic distribution and outcome prevalence often vary between study populations, in which case methods based on this assumption are no longer valid. In the presence of population heterogeneity, some authors proposed to shrink the internal study results toward the external information (Estes, Mukherjee, and Taylor 2018; Gu, Taylor, and Mukherjee 2021).

In this article we consider the setting where the focus is on the internal study population, and the goal of integrating external information is to assist the internal analysis to improve estimation efficiency. This may be the case, for example, when the internal study has a clear target population and is based on a careful design with a well-controlled sampling. In this case, only the information from the external studies that target the
same population as the internal study should be incorporated, as otherwise the external information may lead to estimation bias for internal analysis. Based on this consideration we make it explicit that the internal study population is the target for inference, and the external information that may introduce estimation bias should be discarded. It is worth to point out that there are settings where it is desired to incorporate information from all available external studies, not only from some particular ones, but these settings are not the focus of this article.

In the possible presence of study population heterogeneity, we develop a method that can simultaneously select and incorporate the information from those external studies that target the same population as the internal study and discard the information from the rest. The external information is formulated as moment constraints on the internal study model. The constraints corresponding to external studies that target the same population as the internal study are valid and should be incorporated to help efficiency improvement, and those corresponding to the other external studies are invalid and should be discarded. This formulation makes the data integration problem into a selection of valid moment constraints. We then further formulate it as a variable selection problem by introducing nuisance parameters that represent the biases of the moment constraints under the internal data distribution and select the ones with zero biases. Variable selection can be achieved by shrinkage techniques that estimate some parameters exactly as zeros through a penalization on the nuisance parameters.

The method we develop is penalized constrained maximum likelihood (PCML). The constrained maximum likelihood (CML) type methods have been considered by many authors for data integration when the internal and external study populations are the same (e.g., Qin 2000; Qin et al. 2015; Chatterjee et al. 2016; Huang, Qin, and Tsai 2016; Han and Lawless 2019; Zhang et al. 2020; Han, Taylor, and Mukherjee 2022). In the presence of population heterogeneity, we make use of adaptive group Lasso penalties (Tibshirani 1996; Zou 2006; Yuan and Lin 2006; Wang and Leng 2008) on the CML as a way to simultaneously select and incorporate useful external information into internal analysis. There are situations where, for an external study, the vector of introduced nuisance bias parameters is not a zero vector but does have some zero components. Such an external study can provide some information that is still useful for internal efficiency improvement without introducing estimation bias, even if the study population is not exactly the same as the internal one. Section 2.4 contains a detailed relevant discussion. To account for such situations, we consider both group-wise and component-wise shrinkage for selecting the moment constraints to ensure a maximal incorporation of useful information. The PCML method makes an oracle use of the external information in the sense that the PCML estimator has the same efficiency as the oracle CML estimator that knows which external information is useful and fully incorporates that information alone. Compared to a recently proposed two-step procedure (Sheng et al. 2021) that first conducts a hypothesis test for population heterogeneity and then assumes a nuisance model to link the external information to the internal study, our method simultaneously selects and incorporates the valid external information without specifying any additional models beyond the internal study model.

Our proposed method has implicit connections to some literature on penalized empirical likelihood (Tang and Leng 2010; Leng and Tang 2012; Chang, Tang, and Wu 2018), due to the connections between the CML-type methods and the empirical likelihood (Han and Lawless 2019). But the settings are different. In our data integration setting some external studies provide invalid moment constraints due to population heterogeneity, whereas the penalized empirical likelihood assumes all moment constraints are valid.

We provide a detailed theoretical investigation of the PCML method. Under a set of regularity conditions, including assumptions on the convergence rate of the tuning parameter, we establish the asymptotic properties of the PCML estimator as follows. First, estimation consistency is established by explicitly exploiting the saddle-point representation of the PCML method. Second, the convergence rate of the PCML estimator is shown to be the parametric $\sqrt{n}$-rate. Third, external study selection consistency is established by showing that the nuisance parameters representing the biases of moment constraints are estimated exactly as zero with probability approaching one when the true biases are zero. Fourth, the asymptotic normal distribution is derived jointly for both the internal model parameters and the nuisance parameters representing the nonzero biases of the moment constraints. And last, the asymptotic variance of the PCML estimator for the internal model parameters is shown to be equal to the asymptotic variance of the oracle CML estimator.

2. The Proposed Method

2.1. Setting and Notation

We consider the setting where (a) an internal study collects individual-level data to fit a parametric regression model, (b) some external studies have fitted similar regression models using less detailed covariates with large sample sizes and their model fitting results are available, and (c) these external studies are conducted for possibly different populations. The aim is to incorporate external information that is useful to improve the internal model fitting, since the external information uncertainty is low due to their large sample sizes. One major challenge is how to identify and incorporate only the useful external information, because the information from external studies that do not target the internal study population may introduce estimation bias when incorporated.

This setting is motivated by research in many areas, particularly biomedicine and public health. For example, the internal study collects new covariates such as newly discovered biomarkers, as well as certain conventional covariates such as demographic variables, to investigate their associations with a disease outcome. The internal study sample size may not be large due to budget or technique restrictions. On the other hand, the associations between the outcome and some of the conventional covariates have been established by external studies with large sample sizes, with results available in published articles. Such external information, if incorporated into internal analysis, may substantially improve internal model fitting.

To fix notation, let $(Y_i, X_i^T, Z_i^T)^T, i = 1, \ldots, n$, denote the individual-level data from a random sample collected by the
internal study, where $Y$ is the outcome variable, $X$ is the vector of conventional covariates that are typically collected by studies on the same outcome, and $Z$ is the vector of covariates that are only collected by the internal study. We allow $Z$ to be the null set if the internal study only collects $X$. The main interest is to fit a parametric regression model $f(Y|X,Z; \beta)$ for the distribution $f(Y|X,Z)$, where $\beta$ is a $q$-dimensional vector of parameters with true value $\beta_0$ such that $f(Y|X,Z; \beta_0) = f(X|Z)$. With no additional information, $\beta_0$ can be estimated by the maximum likelihood estimator (MLE) $\hat{\beta}_{\text{MLE}}$ that maximizes the likelihood $\prod_{i=1}^n f(Y_i|X_i,Z_i; \beta)$.

Suppose there are $K$ external studies on the same outcome $Y$ that can potentially provide useful information to improve the internal study population parameter estimation. In this article we consider $K$ to be a fixed finite number. The $k$th external study, $k = 1, \ldots, K$, used covariates $X_{ik}$ and fitted a model $f_{ik}(Y|X_{ik}; \theta_{ik})$ for $f_{ik}(Y|X_{ik})$. Here, for generality, we allow $X_{ik}$ to be a possibly coarsened version of $X$, such as a subset or a categorization of some components of $X$, the subscript of $f_{ik}$ is to explicitly indicate that the $k$th external study population may be different from the internal study population, and $\theta_{ik}$ is the parameters for this model, which is possibly misspecified by the external study. Let $h_{ik}(Y,X_{ik}; \theta_{ik})$ denote the $d_k$-dimensional score function for the model $f_{ik}(Y|X_{ik}; \theta_{ik})$. The $k$th external study then provides an estimate $\hat{\theta}_{ik}$ that is the solution to the corresponding score equation. When the external study sample size is large, the uncertainty in $\hat{\theta}_{ik}$ is negligible compared to the internal study and we will use notation $\theta^{*}_{ik}$ instead of $\hat{\theta}_{ik}$, where $\theta^{*}_{ik}$ is the probability limit of $\hat{\theta}_{ik}$ under the external study. The assumption that the external study uncertainty is negligible compared to the internal study has been made by many authors (e.g., Chaudhuri, Handcock, and Rendall 2008; Qin et al. 2015; Chatterjee et al. 2016; Huang, Qin, and Tsai 2016; Han and Lawless 2019; Zhang et al. 2020; Sheng et al. 2021), and they are closely connected to the empirical likelihood literature (Qin and Lawless 1994; Owen 2001). When all study populations are the same, the CML estimator defined through (3) is more efficient than the MLE and the efficiency gain comes from the integration of the external summary information. With heterogeneous populations, however, in general the CML method no longer works in the sense that the CML estimator can be severely biased after incorporation of the external information.

### 2.2. The CML Method Assuming Population Homogeneity

Hereafter we will use $E(\cdot)$ to denote expectations under the internal study data distribution. When all study populations are the same, (1) has become $0 = E[h_{ik}(Y,X_{ik}; \theta^{*}_{ik})] = E[E[h_{ik}(Y,X_{ik}; \theta^{*}_{ik})|X_i]]$. Thus, defining $U_{ik}(X,Z; \beta, \theta^{*}_{ik}) = \int h_{ik}(Y,X_{ik}; \theta^{*}_{ik})f(Y|X,Z; \beta) dY$, we then have

$$E \left[ U_{ik}(X,Z; \beta_0, \theta^{*}_{ik}) \right] = 0, \quad (2)$$

which summarizes the information from the $k$th external study in the form of moment constraints under the internal study covariate distribution.

To incorporate the external summary information in (2) into estimating $\beta_0$, the CML method introduces a discrete distribution $p_i \geq 0$ on the internal study covariate data $(X_i^T, Z_i^T)^T$, $i = 1, \ldots, n$, and the CML estimator $\hat{\beta}_{\text{CML}}$ for $\beta_0$ is defined through

$$\max_{\beta} \prod_{i=1}^n p_i \prod_{k=1}^K f(Y_i|X_i,Z_i; \beta) \quad \text{subject to} \quad p_i \geq 0, \quad \sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i g(X_i,Z_i; \beta) = 0, \quad (3)$$

where $g(X,Z; \beta) = [U_{1}(X,Z; \beta, \theta^{*}_{1})^T, \ldots, U_{K}(X,Z; \beta, \theta^{*}_{K})^T]^T$. The CML-type estimators have been proposed and studied by many authors under different settings (e.g., Qin 2000; Qin et al. 2015; Chatterjee et al. 2016; Huang, Qin, and Tsai 2016; Han and Lawless 2019; Zhang et al. 2020; Sheng et al. 2021), and they are closely connected to the empirical likelihood literature (Qin and Lawless 1994; Owen 2001). When all study populations are the same, the CML estimator defined through (3) is more efficient than the MLE and the efficiency gain comes from the integration of the external summary information. With heterogeneous populations, however, in general the CML method no longer works in the sense that the CML estimator can be severely biased after incorporation of the external information.

### 2.3. The PCML Method for Heterogeneous Populations

In the presence of heterogeneous populations, the moment constraints in (2) may no longer be valid. To account for this, we introduce some unknown nuisance parameters $\gamma_{0(k)}$, where $\mathcal{Y}_{0(k)} = E \left[ U_{0(k)}(X,Z; \beta_0, \theta^{*}_{0(k)}) \right]$, to represent the bias of the moment constraints resulted from the population difference. Thus, the moment constraints from all external studies can be reparameterized as $E[g(X,Z; \beta_0) = 0$, where $\gamma_{0} = (\gamma_{0(1)}, \ldots, \gamma_{0(K)})^T$. The zero components of $\gamma_{0}$ identify the external studies that are based on the same population as the internal study and whose summary information should be incorporated. It is desirable to estimate the zero components of $\gamma_{0}$ to be exact zeros, which will simultaneously select the external studies that provide useful information and incorporate the information into internal model fitting. The shrinkage estimation techniques can help achieve this goal.
Among the many shrinkage techniques available in the literature that are capable of shrinking the parameter estimates to exactly zero, the Lasso (Tibshirani 1996) is one of the most widely used due to its simplicity and effectiveness. Zou (2006) developed the adaptive Lasso (aLasso) so that both the selection of the zero parameters and the estimation of the nonzero parameters are consistent and the final estimator is as efficient as if the zero parameters are removed from the model before estimation, the so-called oracle property (see also Fan and Li 2001). Therefore, we adopt the aLasso to deal with multiple external studies. Since we are considering multiple external studies, intuition suggests that the shrinkage needs to be carried out at the study level so that an external study should no longer be considered if it is for a different population. Such a group-wise shrinkage can be achieved based on the group Lasso (gLasso) developed by Yuan and Lin (2006). The adaptive version of group Lasso (agLasso) by Wang and Leng (2008) ensures the consistency of both group selection and parameter estimation, as well as the oracle property of the final estimator. Thus, we adopt the agLasso to deal with multiple external studies.

Based on all the considerations so far, we propose the PCML estimator \( \hat{\beta} \) for \( \beta_0 \) that is the \( \beta \)-component of \( \langle \hat{\beta}, \hat{\gamma} \rangle \) defined through

\[
\min_{\beta, \gamma} \left[ \sum_{i=1}^{n} \log f_i(\beta) + \max_{\rho} \left( \sum_{i=1}^{n} \log \{1 - \rho^T [g_i(\beta) - \gamma]\} \right) \right] + n \sum_{k=1}^{K} \hat{P}_{\lambda_k}(\gamma(k)) ,
\]

subject to

\[
p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i g(X_i, Z_i; \beta) - \gamma = 0,
\]

where

\[
\hat{P}_{\lambda_k}(\gamma(k)) = \lambda_n \| \hat{\gamma}(k) \|^w \| \gamma(k) \|
\]

is the agLasso penalty with tuning parameter \( \lambda_n > 0 \), \( \| \cdot \| \) is the Euclidean norm, \( \hat{\gamma}(k) \) is some first-step consistent estimator of \( \gamma(0(k)) \), and \( w > 0 \) is some user-specified positive number. The most natural choice for \( \hat{\gamma}(k) \) in the setting we consider is to take the corresponding components from \( \gamma = n^{-1} \sum_{i=1}^{n} g(X_i, Z_i; \beta_0) \). A common choice for \( w \) is \( w = 1 \) or 2 (e.g., Zou 2006; Wang and Leng 2008).

Compared to the optimization in (3) for the CML estimator, the optimization in (4) for the proposed PCML estimator has an agLasso penalty that shrinks the estimate of \( \gamma_0 \) toward zero. When the degree of shrinkage is properly chosen through the tuning parameter \( \lambda_n \), some \( \gamma(0(k)) \) will be estimated exactly as zeros and the corresponding information summarized in the moment constraints (2) will be automatically incorporated into the estimation of \( \beta_0 \). Furthermore, when only the \( \gamma(0(k)) \) corresponding to the external studies that are for the same population as the internal study are estimated as zeros, the resulting PCML estimator for \( \beta_0 \) will be consistent and have improved efficiency compared to the MLE. The penalization in (4) allows simultaneous selection of useful external information and estimation of \( \beta_0 \) incorporating that information.

Using the Lagrange multiplier method, it is easy to show that the PCML constrained optimization in (4) can be equivalently written as

\[
\min_{\beta, \gamma} \left[ \sum_{i=1}^{n} \log f_i(\beta) + \max_{\rho} \left( \sum_{i=1}^{n} \log \{1 - \rho^T [g_i(\beta) - \gamma]\} \right) \right] + n \sum_{k=1}^{K} \tilde{P}_{\gamma_k}(\beta(k)) ,
\]

where \( f_i(\beta) = f(Y_i|X_i, Z_i; \beta), g_i(\beta) = g(X_i, Z_i; \beta), \) and \( \rho \) is the Lagrange multiplier. The expression in (6) is the so-called saddle-point representation in the empirical likelihood literature (e.g., Owen 2001; Newey and Smith 2004) and is the expression used both for derivation of the asymptotic properties and for the numerical implementation in later sections.

### 2.4. Group-wise Shrinkage versus Component-wise Shrinkage

The agLasso penalty (5) is based on the intuition that an external study should no longer be considered for information integration if its population is different from the internal study. The penalty (5) ensures that data integration is carried out in a group-wise manner at the study level. However, a further investigation reveals that not all components of (2) are necessarily invalid when the external study has a different population. In below we present two such examples, both of which have an appreciable degree of generality despite the concrete numbers in Example 2. Both examples are based on a factorization of the joint distribution \( f(Y, X, Z) = f(Y|X)f(Z|X)f(X) \). Example 1 shows that (2) may still hold if the difference between the internal and external study populations is only in \( X \). Example 2 shows that, in the presence of a difference in any of \( f(Y|X, Z), f(Z|X) \) and \( f(X) \), some components in (2) may still hold even though the rest do not.

**Example 1.** Suppose that the internal and external studies have different distributions for \( X \) but share the same distribution for both \( Y|X, Z \) and \( Z|X \), and thus, they also share the same distribution for \( Y|X \). Suppose that the external study used a correctly specified model \( f(Y|X; \theta) \), which implies that \( \mathbb{E}[h(Y, X; \theta^*)|X] = 0 \). Note that in this case, due to the correct specification of \( f(Y|X; \theta) \), the moment equality is conditional on \( X \) and thus, holds regardless of the difference in the \( X \) distribution between the internal and external studies. Thus, the same calculation leading to (2) shows that \( \mathbb{E}[U(X, Z, \beta_0, \theta^*)|X] = 0 \), which then implies (2).

**Example 2.** Suppose that, for the internal study, \( X \sim N(-0.5, 0.5^2) \), \( Z|X \sim N(X + X^2, 1.25^2) \), and \( Y|X, Z \sim N(\beta_1 + \beta_2 X, 1.25^2) \) with \( \beta_0 = (0.5, 1, 0.5)^T \). For the external study with data generated as in Cases (a)–(c) below, the model \( Y|X \sim N(\theta_1 + \theta_2 X, \sqrt{1.25^2}) \) is fitted, which leads to \( h(Y, X; \theta) = (1, X)^T(Y - \theta_1 - \theta_2 X) \) and \( U(X, Z; \beta, \theta) = (1, X)^T(\beta + \beta_2 X + \beta_2 Z - \theta_1 - \theta_2 X) \). Some calculation shows that under the internal study \( \mathbb{E}(X^2) = 0.5, \mathbb{E}(Z) = 0 \) and \( \mathbb{E}(XZ) = 0 \), and thus, \( \mathbb{E}(U(X, Z; \beta, \theta)) = (\beta_1 - 0.5 \beta_2 - \theta_1 + 0.5 \theta_2, -0.5 \beta_1 + 0.5 \beta_2 + 0.5 \theta_1 - 0.5 \theta_2)^T \). Now consider three cases for the external study data distribution. (a) The distributions of \( Z|X \) and \( Y|X, Z \) are the same as the internal study while \( X \sim N(0, \sqrt{0.5^2}) \). Some calculation shows that \( \theta^* = (0.75, 1.5)^T \), which then leads to \( \mathbb{E}(U(X, Z; \beta_0, \theta^*)) = (0, -0.125)^T \). (b) The distributions of \( X \)
3. Asymptotic Properties of the PCML Method

3.1. Estimation Consistency and √n-Convergence

We first establish the consistency of the proposed estimator \((\hat{\beta}, \hat{\gamma})\). The assumptions needed on the model \(f(Y|X,Z;\beta)\) and the moment function \(g(X,Z;\beta)\) are similar to those for the consistency of the MLE and the empirical likelihood estimator (e.g., Newey and McFadden 1994; Qin and Lawless 1994; Newey and Smith 2004). In addition, the penalty function needs to be small enough compared to the likelihood function and this is achieved through an assumption on the turning parameter \(\lambda_n\).

**Assumption 1.** (i) \(B \times T\), the parameter space where \((\beta_0, \gamma_0)\) lies, is compact;

(ii) \(\mathbb{E}[\log f(Y|X,Z;\beta)]\) is uniquely maximized at \(\beta_0 \in B\);

(iii) \(\log f(Y|X,Z;\beta)\) is continuous at each \(\beta \in B\) with probability one;

(iv) \(\mathbb{E}\left[\sup_{\beta,\gamma \in B \times T} \|g(X,Z;\beta) - \gamma\|^2\right] < \infty\) for some \(\alpha > 2;\)

(v) \(\mathbb{E}\left[\|g(X,Z;\beta_0) - \gamma_0\|^2\right]\) is nonsingular;

(vi) \(\sup_{\beta \in B} n^{-1/2} \sum_{k=1}^n \left[I(\beta) - \mathbb{E}[I(\beta)]\right] = O_p(1)\) for \(I(\beta) = \log f(Y|X,Z;\beta) + g(X,Z;\beta)\);

(vii) \(\lambda_n = O_p(n^{-\xi})\) for some \(\xi\) with \(1/\alpha < \xi < 1/2\).

Here Assumption 1(vi) is a functional Central Limit Theorem and is a standard result in the empirical processes theory (Donsker’s Theorem, e.g., Andrews 1994; van der Vaart and Wellner 1996; van der Vaart 2000; Kosorok 2008). It is a uniform version of the standard Central Limit Theorem that holds under the typical regularity conditions (e.g., Newey and McFadden 1994). Assumption 1(vii) makes sure that the shrinkage effect when estimating the nonzero components of \(\gamma_0\) disappears as \(n \to \infty\). Under Assumption 1, the consistency of \((\hat{\beta}, \hat{\gamma})\) is given by Theorem 1. The proof makes use of the saddle-point representation in (6). This proof, together with the proofs of all other theorems, is given in the supplementary materials.

**Theorem 1 (Consistency).** Under Assumption 1, the PCML estimator \((\hat{\beta}, \hat{\gamma})\) converges to \((\beta_0, \gamma_0)\) in probability as \(n \to \infty\).

To establish the \(\sqrt{n}\)-convergence of \((\hat{\beta}, \hat{\gamma})\) we need some additional assumptions.

**Assumption 2.** (i) \((\beta_0, \gamma_0)\) is in the interior of \(B \times T\);

(ii) \(g(X,Z;\beta)\) is continuously differentiable in some neighborhood \(B_K'\) of \(\beta_0\) and \(\mathbb{E}[\sup_{\beta \in B_K'} \|\partial g(\beta)/\partial \beta\|] < \infty;\)

(iii) \(\log f(Y|X,Z;\beta)\) is twice continuously differentiable in some neighborhood \(B_K\) of \(\beta_0\) and \(\mathbb{E}[\sup_{\beta \in B_K} \|\partial^2 g(\beta)/\partial \beta \partial \beta\|] < \infty,\) where \(\partial g(\beta)/\partial \beta\);

(iv) \(\mathbb{E}[(\partial^2 g(\beta)/\partial \beta \partial \beta)^{-1}]\) is nonsingular;

(v) \(\lambda_n = o_p(n^{-1/2}).\)

The assumptions needed on \(f(Y|X,Z;\beta)\) and \(g(X,Z;\beta)\) are similar to those in Newey and McFadden (1994), Newey and Smith (2004), and Liao (2013). The \(\sqrt{n}\)-convergence requires that the tuning parameter converges to zero as fast as possible so that the penalty term in (4) is asymptotically small compared to the likelihood term in (4). Assumption 2(v) ensures that \(\lambda_n\) converges to zero fast enough to obtain \(\sqrt{n}\)-convergence of \((\hat{\beta}, \hat{\gamma})\). Under Assumptions 1 and 2, the \(\sqrt{n}\)-convergence of \((\hat{\beta}, \hat{\gamma})\) is given by Theorem 2. Theorem 2 also gives the \(\sqrt{n}\)-convergence of the Lagrange multiplier \(\hat{\rho}\) corresponding to \((\hat{\beta}, \hat{\gamma})\), and this result is oftentimes of independent interest. For example, the tuning parameter selection in Section 4.2 makes use of this result.

**Theorem 2 (\(\sqrt{n}\)-Consistency).** Under Assumptions 1 and 2, we have (i) \(\|\hat{\beta} - \beta_0\| = O_p(n^{-1/2});\) (ii) \(\|\hat{\gamma} - \gamma_0\| = o_p(n^{-1/2});\) and (iii) \(\hat{\rho} = \arg \max \sum_{k=1}^n \log(1 - \rho^T[g(\hat{\beta} - \hat{\gamma})],\) the Lagrange multiplier as in (6), exists with probability approaching one and \(\|\hat{\rho}\| = O_p(n^{-1/2}).\)

3.2. External Study Selection Consistency

Let \(K_0 = \{k : \gamma_0(k) \neq 0, k = 1, \ldots, K\}\) and \(K_{\neq 0} = \{k : \gamma_0(k) \neq 0, k = 1, \ldots, K\}\) denote the index sets for the zero and nonzero groups in \(\gamma_0\), respectively, corresponding to external studies that are for the same population as the internal study and those for different populations. Let \(\hat{K}_0 = \{k : \hat{\gamma}(k) = 0, k = 1, \ldots, K\}\).
1, \ldots, K) and \hat{\mathcal{K}}_{\neq 0} = \{k : \hat{\gamma}(k) \neq 0, k = 1, \ldots, K\} denote the index sets for the zero and nonzero groups in \hat{\gamma}, respectively, corresponding to external studies that are selected by the PCML method for information integration and those are not selected.

The consistency of \hat{\gamma} from Theorem 1 implies that \hat{\gamma} falls into a shrinking neighborhood of \gamma_0 with probability approaching one, and thus, for those \gamma_0(k) \neq 0 we must have \gamma(k) \neq 0 with probability approaching one. However, consistency of \hat{\gamma} alone does not imply \gamma(k) = 0 with probability approaching one for those \gamma_0(k) = 0, and thus, does not imply external study selection consistency. To ensure the selection consistency, we impose a further condition on the convergence rate of the tuning parameter \lambda_n. This condition is based on taking \gamma = n^{1/2+w/2}\lambda_n \rightarrow \infty as n \rightarrow \infty.

Theorem 3. Under Assumptions 1, 2 and 3, we have \lim_{n \rightarrow \infty} P(\hat{\mathcal{K}}_{\neq 0} = \mathcal{K}_{\neq 0}) = 1.

3.3. Asymptotic Distribution and Oracle Information Integration

To derive the asymptotic distribution of the proposed PCML estimator, rewrite \gamma_0 as \gamma_0^T = (\gamma_0^T, \gamma_{0,\neq 0}^T) without loss of generality, where \gamma_0 = 0 contains those \gamma_0(k) that \gamma_0(k) = 0 and \gamma_{0,\neq 0} contains those \gamma_0(k) that \gamma_0(k) \neq 0. Correspondingly, write \gamma = (\gamma^T, \gamma_{\neq 0}^T) = (\gamma_0^T, \gamma_{0,\neq 0}^T), \gamma as \gamma^T = (\gamma_0^T, \gamma_{\neq 0}^T), and \hat{\gamma} as \hat{\gamma}^T = (\hat{\gamma}_0^T, \hat{\gamma}_{\neq 0}^T). Define \eta = (\beta^T, \gamma_{\neq 0}^T), \tilde{\eta} = (\beta^T, \gamma_{\neq 0}^T), and \hat{\eta} = (\hat{\beta}^T, \hat{\gamma}_{\neq 0}^T). Because \hat{\gamma}_{\neq 0} = 0 with probability approaching one based on Theorem 3, we just need to derive the asymptotic distribution of \hat{\gamma}. The result is given by the following theorem.

Theorem 4 (Asymptotic Normality). Under Assumptions 1, 2 and 3, we have \sqrt{n}(\hat{\gamma} - \eta_0) \xrightarrow{d} \mathcal{N}(0, S + G_0^T \Omega_0^{-1} G_0^{-1})^{-1}, where \text{S} = \text{diag}(S_0, 0), S_0 = E[s(\beta_0)(s(\beta_0)^T)]^{1/2}, s(\beta) = \partial \log(\ell)(X, Z; \beta)/\partial \beta, G_0 = E[\partial(g(X, Z; \beta_0) - \gamma_0)/\partial \eta], and \Omega_0 = E[(g(X, Z; \beta_0) - \gamma_0)(g(X, Z; \beta_0) - \gamma_0)^T].

From Theorem 4, some calculations lead to the asymptotic distribution for the PCML estimator \hat{\beta}.

Theorem 5 (Oracle Efficiency). Under Assumptions 1, 2, and 3, we have

\[ \sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, (S_0 + G_0^T \Omega_0^{-1} G_0)^{-1}) \]

(7)

where \text{G}_0 = E[\partial(g(X, Z; \beta_0) - \beta)/\partial \beta] and \Omega_0 = E[(g(X, Z; \beta_0) - \gamma_0)^2].

Compared to the MLE based on the internal study data alone, whose asymptotic variance is \text{S}_0^{-1}, the proposed PCML estimator \hat{\beta} has a smaller asymptotic variance because \text{G}_0^T \Omega_0^{-1} \text{G}_0 is positive-definite. On the other hand, the asymptotic variance in (7) is the same as that of the oracle CML estimator defined in (3) with only \text{g}(X, Z; \beta) used. In other words, the proposed estimator has the same efficiency as that of the oracle CML estimator incorporating only useful external information.

This optimal estimation efficiency for the parameter of interest, together with the external study selection consistency from Theorem 3, implies the oracle use of information from external studies in the presence of population heterogeneity.

4. Implementation

4.1. Implementation Based on Saddle-Point Representation

The numerical implementation of the proposed PCML method is based on the saddle-point representation (6) and consists of two loops, following the recommendation from the empirical likelihood literature (e.g., Owen 2001; Kitamura 2007; Han and Lawless 2019). The inner loop computes the Lagrange multiplier \rho(\hat{\beta}, \gamma) at a given value of (\beta, \gamma), and the outer loop updates (\beta, \gamma).

Specifically, the inner loop is max_{\eta} \sum_{i=1}^n \log (1 - \rho^T [g_i(\beta, \gamma)]) as in (6). When the given value (\beta, \gamma) is close to the true value (\beta_0, \gamma_0), which is indeed the case during the implementation if the initial value of (\beta, \gamma) is taken to be the consistent estimator (\hat{\beta}_{\text{MLE}}, \hat{\gamma}), the inner loop is a concave maximization with a unique maximizer (e.g., Han 2014). Thus, the inner loop can be easily implemented based on the Newton-Raphson algorithm, for which the initial value can be simply set as \rho = 0 because of Theorem 2.

To present the outer loop, let \tilde{\rho}(\hat{\beta}, \gamma) denote the computed Lagrange multiplier from the inner loop at a given (\beta, \gamma) and \tilde{\rho}(\hat{\beta}, \gamma) the components of \tilde{\rho}(\hat{\beta}, \gamma) corresponding to \gamma_0. The outer loop computes the PCML estimator (\hat{\beta}, \hat{\gamma}) in the following steps.

Step 0. Take the initial value (\hat{\beta}^{(0)}, \hat{\gamma}^{(0)}) = (\hat{\beta}_{\text{MLE}}, \hat{\gamma}).

With \hat{\beta}^{(i)} \hat{\gamma}^{(i)} available from the \text{i}th iteration (i = 0, 1, 2, \ldots), in the (l+1)th iteration the outer loop obtains \hat{\beta}^{(l+1)} and \hat{\gamma}^{(l+1)} based on a block coordinate descent procedure.

Step 1. For k = 1, \ldots, K sequentially, set \hat{\gamma}^{(l+1)} equal to 0 if

\[ \frac{1}{n} \sum_{i=1}^n \left[ \hat{\rho}(\hat{\beta}^{(l)}, \hat{\gamma}^{(l+1)}(\hat{\gamma}(k)) - \gamma(\hat{\gamma}(k))) \right]^T [g_i(\hat{\beta}^{(l)}) - \hat{\gamma}^{(l+1)}(\hat{\gamma}(k))] \]

\[ \leq \frac{\lambda_n}{\|\hat{\gamma}(k)\|^W} \] (8)

and equal to the root of the equation

\[ \frac{\lambda_n}{\|\hat{\gamma}(k)\|^W} \|\hat{\gamma}(k)\| + \frac{1}{n} \sum_{i=1}^n \rho(k) \hat{\rho}(\hat{\beta}^{(l)}, \hat{\gamma}^{(l+1)}(\gamma(k))) \]

\[ \left[ g_i(\beta^{(l)}) - \hat{\gamma}^{(l+1)}(\gamma(k)) \right] \] (9)
as an equation for \( y_{(k)} \) if (8) does not hold, where

\[
\hat{y}^{(l+\frac{1}{2})}(y_{(k)}) \equiv \left[ (\hat{y}^{(l+1)}_{(1)})^T, \ldots, (\hat{y}^{(l+1)}_{(k-1)})^T, y_{(k)}^T, (\hat{y}^{(l)}_{(k)})^T, \ldots, (\hat{y}^{(l)}_{(K)})^T \right]^T.
\]

Step 2. Set \( \hat{\beta}^{(l+1)} \) equal to the root of the equation

\[
\frac{1}{n} \sum_{i=1}^{n} g_i(\hat{\beta}) + \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{\partial g_i(\beta)}{\partial \beta} \right]^T \hat{\rho}(\beta, \hat{y}^{(l+1)}) = 0.
\]

(10)
as an equation for \( \beta \).

Step 3. Repeat Step 1 and Step 2 until convergence such that \( \| \hat{y}^{(l+1)} - \hat{\beta}^{(l)} \| \) and \( \| y^{(l+1)} - \hat{y}^{(l)} \| \) are smaller than some prespecified small number and \( \hat{\Sigma}^{(l+1)} = \hat{\Sigma}^{(l)} \) where \( \hat{\Sigma}^{(l)} = \{ k : \hat{\rho}_{ik} = 0, k = 1, \ldots, K \} \).

Equations (9) and (10) are the first-order condition of the saddle-point representation (6) with respect to \( y_{(k)} \) when \( y_{(k)} \neq 0 \) and \( \beta \), respectively, treating \( \hat{\rho}(\beta, y) \) as an implicit function of \( \beta \) and \( y \). These equations can be solved based on the Newton–Raphson algorithm, for which the calculation of the Jacobian matrices of the left-hand sides of (9) and (10) needs to again treat \( \hat{\rho}(\beta, y) \) as an implicit function of \( \beta \) and \( y \). The expression of the Jacobian matrix for (10) is the same as that in Han and Lawless (2019) and the expression for (9) can be similarly derived. Details are omitted here due to their lengthy expressions.

4.2. Tuning Parameter Selection

The rate of convergence of the tuning parameter \( \lambda_n \) is crucial when deriving the asymptotic properties of the PCML estimator in Section 3, and Assumptions 2(v) and 3 specify some sufficient conditions on the convergence rate that guarantee the \( \sqrt{n} \)-convergence and the oracle property of the PCML estimator. For practical implementation, however, we need an effective way of selecting a concrete value for the tuning parameter.

Note from (8) that \( y_{(0)}(k) \) is estimated exactly as zero if

\[
\left| \frac{\hat{\rho}_{(k)}(\hat{\beta}, \hat{y}_{(-k)})}{\sqrt{n}} \right| \leq \frac{1}{\| g_i(\hat{\beta}) - \hat{y}_{(-k)} \|} \leq \frac{1}{\sqrt{n}\lambda_n \| y_{(-k)} \|^{w}}.
\]

(11)

where \( \hat{y}_{(-k)} = (\hat{y}^{T}_{(1)}, \ldots, \hat{y}^{T}_{(k-1)}, 0, \hat{y}^{T}_{(k+1)}, \ldots, \hat{y}^{T}_{(K)})^T \). For any \( y_{(0)}(k) \neq 0 \), \( \hat{\rho}(\beta, \hat{y}_{(-k)}) \) is of order \( O_p(1) \) and thus, the left-hand side of (11) is asymptotically bounded away from zero, in which case to avoid estimating \( y_{(0)}(k) \) to be zero \( \sqrt{n}\lambda_n \) needs to converge to zero as fast as possible, since \( \| y_{(-k)} \|^{w} \) converges to a nonzero constant. With all \( y_{(0)}(k) \neq 0 \) estimated as nonzeros, for any \( y_{(0)}(k) = 0 \), \( \hat{\rho}(\beta, \hat{y}_{(-k)}) \) is of order \( O_p(n^{-1/2}) \) and the left-hand side of (11) is of order \( O_p(1) \), and in addition \( \| y_{(k)} \| = O_p(n^{-1/2}) \). Therefore, to estimate \( y_{(0)}(k) = 0 \) exactly as zero

\[ n^{1/2+w/2}\lambda_n \text{ needs to diverge to infinity as fast as possible. These considerations agree with Assumptions 2(v) and 3. To balance these rate requirements on } \lambda_n, \text{ we choose } \lambda_n = Cn^{-1/2-w/4}, \text{ where } C \text{ is a positive constant.} \]

We now discuss how to select \( C > 0 \) by following the idea in Liao (2013). From the proof of Theorem 4 it is seen that \( \sqrt{n}\hat{\rho} \rightarrow \mathbf{Y}v \), where \( \mathbf{Y} = \Omega^{-1}A(\text{diag}(\Omega, S_0))^{1/2} \), \( \mathbf{v} \) is a \((d + q)\) dimensional standard Gaussian random vector, and \( d = \dim(\rho_0) = \sum_{k=1}^{d_k} \). On the other hand, under \( \sqrt{n}\)-consistent estimation, the left-hand side of (11) has the same asymptotic distribution as \( \sqrt{n}\hat{\rho}_{(k)} = \sqrt{n}\hat{\rho}_s, \) where \( \hat{\rho}_s \) is the \( d_k \times d \) matrix that selects the components \( \rho_{ik} \) from \( \rho \). Therefore, to account for the study heterogeneity of the left-hand side of (11) and to normalize the linear combination of \( v \), we allow the \( C \) in the tuning parameter \( \lambda_n = Cn^{-1/2-w/4} \) to be study-specific and choose \( C_{(k)} = \| B_k \mathbf{Y} \| _F \), where \( \| \| _F \) is the Frobenius norm and \( \mathbf{Y} \) is an estimate of \( \mathbf{Y} \) with a preliminary PCML estimator plugged in. For the preliminarily estimated tuning parameter can be taken as \( \lambda_n = n^{-1/2-w/4} \) with \( C = 1 \).

5. Simulation Studies

5.1. Simulation Setup

For the internal study there are four covariates, \( X_1, X_2, X_3 \) and \( Z \), which are generated as follows; \( (X_1, \tilde{X}_2) \sim \mathcal{N}(0, \Sigma) \) with unit variances and correlation coefficient 0.3, \( X_2 = I(\tilde{X}_2 > 0) \), \( X_3 \sim \text{Exponential}(1) \), and \( Z|X \sim \mathcal{N}(X_1 + X_3, 1^2) \). Given \( X \) and \( Z \), \( Y \) is generated from a Bernoulli distribution with \( \log (P(Y = 1|X, Z)) = (X_1, X_2, X_3, X_1Z, X_1X_2)^\top \beta_0 \), with \( \beta_0 = (-0.5, 0.5, -1.5, 1, 0.5, -0.5) \). The internal study model is the corresponding logistic regression with \( \beta^T = (\beta_1, \beta_{X_1}, \beta_{X_2}, \beta_{X_3}, \beta_Z, \beta_XZ) \).

We consider two external studies. External study 1 has the same data distribution as the internal study and measures only \( Y, X_2 \) and \( X_3 \) to fit the logistic regression model logit\( (P(Y = 1|X_2, X_3)) = \{ \xi_{(1)} + \theta_{(1)1}X_2 + \theta_{(1)2}X_3 \} \). External study 2 has a different covariate distribution. Specifically, \( (X_1, \tilde{X}_2) \sim \mathcal{N}(−0.5, 0.25), \Sigma) \) with the same \( \Sigma \) as the internal study, \( X_2 = I(\tilde{X}_2 > 0) \), \( X_3 \sim \text{Exponential}(0.5) \), and \( Z|X \sim \mathcal{N}(0.25 + 0.5X_1, 0.5X_1, 1^2) \). Conditional on the covariates, the outcome distribution in external study 2 is the same as that in the internal study. External study 2 measures only \( Y, X_1 \) and \( X_3 \) to fit the logistic regression model logit\( (P(Y = 1|X_1, X_3)) = \{ \xi_{(2)} + \theta_{(2)1}X_1 + \theta_{(2)2}X_3 \} \). The \( \mathbf{h}_{(k)}(Y, X_{(k)}; \theta_{(k)}), k = 1, 2, \) are the corresponding score functions for these two external logistic regression models. In the supplementary materials we provide a detailed expression for (4) that defines the PCML estimator under this particular simulation setting.

Since external study 1 has the same data distribution as the internal study, we have \( y_{(0)}(1) = 0 \) and thus, incorporating the information from this external study should improve the efficiency for internal parameter estimation. For external study 2, some numerical calculation based on a sample size \( 10^6 \) for both the internal and external studies shows that \( y_{(0)}(2) = (-0.1651, -0.0036, -0.0957) \). The second component of \( y_{(0)}(2) \)
is very close to zero, and thus, part of the information available from external study 2 may be helpful for efficiency improvement as well. To evaluate the numerical performance of the PCML method, we consider three scenarios where the external summary information is available from (i) external study 1 only, (ii) external study 2 only, and (iii) both external studies. The MLE and the CML estimators are included for comparisons. In each scenario, both the group-wise shrinkage and the component-wise shrinkage are applied. We take \( w = 2 \) in the penalty function (5).

For the external studies we consider two sample sizes, 50,000 and 3000, corresponding to large and moderate sizes, respectively. In both cases two internal sample sizes, \( n = 300 \) and 800, are considered. When the external sample size is 50,000, all replications use the same external study data. When the external sample size is 3000, in each replication the external data are regenerated together with the internal data. Table 1 contains the results for external sample size 50,000, and Table S1 in the supplementary materials contains the results for external sample size 3000, both based on 1000 replications. The observations from these two tables are very similar.

### 5.2. Simulation Observations

When only using External Study 1, the CML estimator CML-1 is the oracle CML estimator and has a substantial reduction of empirical standard errors, compared to the MLE, for the estimates of \( \beta_3, \beta_{X_2} \), and \( \beta_{X_1} \) corresponding to the regressors used in External Study 1. This observation is in full agreement with the existing CML literature that the efficiency improvement occurs mainly for the estimates corresponding to external study covariates. The PCML estimator with group-wise shrinkage (PCMLg-1) has a performance very close to CML-1, especially when \( n = 800 \). Even with \( n = 300 \), compared to the MLE, PCMLg-1 has substantially smaller empirical standard errors for the estimates of \( \beta_3, \beta_{X_2}, \), and \( \beta_{X_1} \). The PCML estimator with component-wise shrinkage (PCMLc-1) has a performance almost identical to CML-1.

The closeness in performance in this case between PCMLg-1, PCMLc-1 and CML-1 is because the PCML method is able to automatically incorporate all the information available from External Study 1. For PCMLg-1, the selection rate of External Study 1 is 97.1\% for \( n = 300 \) and 98.7\% for \( n = 800 \), and for PCMLc-1 the selection rate of the entire External Study 1 is 99.6\% for \( n = 300 \) and 99.8\% for \( n = 800 \).

When only using External Study 2, the CML estimator CML-2 clearly has a large bias because this external study data distribution is different from the internal study. On the other hand, the CML estimator CML-2 that only uses the second moment constraint out of the three that External Study 2 provides has little bias. In addition, CML-2o has a considerably smaller empirical standard error for the estimate of \( \beta_{X_1} \) compared to the MLE. This improved performance of CML-2o over the MLE is because that the second component of \( \gamma_{0(2)} \) is very close to zero. Thus, CML-2o may be treated as the oracle CML estimator in this case for comparison purposes. In this case the PCML estimator with group-wise shrinkage (PCMLg-2) has identical results to the MLE, since the group-wise shrinkage detected the distribution difference and made no use of the external information for both \( n = 300 \) and \( n = 800 \). The PCML estimator with component-wise shrinkage (PCMLc-2) has a performance almost identical to the oracle CML-2o when \( n = 800 \), showing the effectiveness of the component-wise shrinkage in integrating useful external information in the presence of population heterogeneity. The rate of estimating the second component of \( \gamma_{0(2)} \) and only this component exactly as zero is 99.8\% in this case. When \( n = 300 \), due to randomness in the internal data, PCMLc-2 sometimes estimates the third component of \( \gamma_{0(2)} \) also as zero, whose true value is \( -0.0957 \). Specifically, when \( n = 300 \) the rate of estimating the second component of \( \gamma_{0(2)} \) alone as zero is 75.4\% and estimating the second and third components but not the first as zero is 23.5\%. Incorporating information from the third moment constraint leads to slight bias and larger empirical standard errors for PCMLc-2 compared to CML-2o, but all these disappear when \( n = 800 \).

When using both external studies, the CML estimator CML-12 has a large bias. Compared to both CML-1 and CML-2o, the oracle CML estimator CML-12o that uses all moment constraints from External Study 1 and the second moment constraint from External Study 2 has a further reduction in empirical standard errors for certain estimates. The PCML estimator with group-wise shrinkage (PCMLg-12) has a performance almost identical to PCMLg-1, especially when \( n = 800 \), since the group-wise shrinkage correctly selected External Study 1 with rate 96.6\% when \( n = 300 \) and rate 98.7\% when \( n = 800 \), and never selected External Study 2.

The PCML estimator with component-wise shrinkage (PCMLc-12) has a performance almost identical to the oracle CML-12o when \( n = 800 \). When \( n = 300 \) PCMLc-12 has a slightly larger empirical standard error compared to CML-12o due to occasionally estimating the third component of \( \gamma_{0(2)} \) as zero. Specifically, when \( n = 300 \) the rate of correctly selecting External Study 1 together with only the second moment constraint from External Study 2 is 97.2\%, and the rate becomes 99.4\% when \( n = 800 \). Compared to PCMLc-1, PCMLc-12 shows a better overall efficiency, especially when \( n = 800 \), due to the integration of additional useful information from External Study 2. Compared to PCMLc-2, the efficiency improvement of PCMLc-12 is substantial. Compared to PCMLg-12, PCMLc-12 has a clear reduction in the empirical standard error for the estimate of \( \beta_{X_1} \), corresponding to the covariate \( X_1 \) that is used only by External Study 2.

Based on all these observations, the PCML method is very effective in incorporating useful external information in the presence of study population heterogeneity. Especially, the PCML estimator based on component-wise shrinkage can make a partial use of the information from an external study that is not selected by the group-wise shrinkage. The numerical performance is overall excellent even with a small internal sample size.

### 5.3. Bootstrap for Inference

In finite samples, the standard error of the PCML estimator \( \hat{\beta} \) calculated based on the asymptotic distribution (7) does not properly account for the finite-sample study selection error, and thus, may lead to poor inferences about \( \beta_0 \). A theoretical development and investigation of a method that takes the study
The risk of developing high-grade prostate cancer (Gleason score ≤ 7) and certain risk factors. The effects of some commonly considered risk factors, including age, race, the prostate specific antigen (PSA) level, the digital rectal examination (DRE) finding and prior biopsy result, have been studied extensively in the literature. Among the studies, Thompson et al. (2006) built an online risk calculator for calculating the risk of developing high-grade prostate cancer, using data collected in the 1990s from 5519 men in the placebo group of the Prostate Cancer Prevention Trial (PCPT). This PCPT risk calculator is the first online prostate cancer risk assessment tool and is among the most widely used ones. Detailed information about the study, including the model behind this risk calculator, is provided in Thompson et al. (2006).

Recent research on the biological mechanisms related to the progression of prostate cancer shows that two specific biomarkers, TMPRSS2:ERG (T2:ERG) and prostate cancer antigen 3 (PCA3), may lead to a better early detection of the disease (e.g., Tomlins et al. 2016). Therefore, it is of great interest to study the effects of both the aforementioned conventional risk factors and the new biomarkers on the risk of prostate cancer after adjusting for selection error into account is challenging and is beyond the scope of this article. Instead, we evaluate the performance of the bootstrap method for a numerical calculation of the standard error. The results are summarized in Table 2. It is seen that, when \( n = 300 \) the bootstrap standard errors overall overestimate the empirical standard errors, but the overestimation becomes much milder when \( n = 800 \), in which case the difference is less for component-wise shrinkage compared to group-wise shrinkage. In the presence of overestimation, bootstrap will lead to more conservative inference. Overall, when the internal sample size is not very small, the bootstrap method seems to have an acceptable performance and provide a feasible way for standard error calculation in the absence of other formal methods.

### 6. Data Application

We apply the proposed method to study the association between the risk of developing high-grade prostate cancer (Gleason score ≥ 7) and certain risk factors. The effects of some commonly considered risk factors, including age, race, the prostate specific antigen (PSA) level, the digital rectal examination (DRE) finding and prior biopsy result, have been studied extensively in the literature. Among the studies, Thompson et al. (2006) built an online risk calculator for calculating the risk of developing high-grade prostate cancer, using data collected in the 1990s from 5519 men in the placebo group of the Prostate Cancer Prevention Trial (PCPT). This PCPT risk calculator is the first online prostate cancer risk assessment tool and is among the most widely used ones. Detailed information about the study, including the model behind this risk calculator, is provided in Thompson et al. (2006).

Recent research on the biological mechanisms related to the progression of prostate cancer shows that two specific biomarkers, TMPRSS2:ERG (T2:ERG) and prostate cancer antigen 3 (PCA3), may lead to a better early detection of the disease (e.g., Tomlins et al. 2016). Therefore, it is of great interest to study the effects of both the aforementioned conventional risk factors and the new biomarkers on the risk of prostate cancer after adjusting for...

### Table 1. Simulation results summarized based on 1000 replications with external study sample size 50,000.

| Internal sample size \( n = 300 \) | \( \beta_c \) | \( \beta_{x_1} \) | \( \beta_{x_2} \) | \( \beta_{x_3} \) | \( \beta_Z \) | \( \beta_{x_1}Z \) | \( \beta_{x_2}Z \) | \( \beta_{x_3}Z \) |
|----------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| MLE                              | Bias         | ESE          | RMSE         | Bias         | ESE          | RMSE         | Bias         | ESE          |
|                                 | -0.024       | 0.268        | 0.269        | -0.008       | 0.116        | 0.131        | -0.006       | 0.131        |
|                                 | 0.022        | 0.259        | 0.260        | 0.023        | 0.157        | 0.129        | 0.023        | 0.159        |
|                                 | -0.337       | 0.333        | 0.335        | -0.028       | 0.160        | 0.160        | -0.026       | 0.159        |
|                                 | 0.036        | 0.254        | 0.256        | 0.017        | 0.166        | 0.166        | 0.015        | 0.166        |
|                                 | 0.021        | 0.166        | 0.167        | 0.021        | 0.166        | 0.167        | 0.021        | 0.166        |
|                                 | -0.022       | 0.117        | 0.119        | -0.022       | 0.118        | 0.120        | -0.022       | 0.118        |

| Internal sample size \( n = 800 \) | \( \beta_c \) | \( \beta_{x_1} \) | \( \beta_{x_2} \) | \( \beta_{x_3} \) | \( \beta_Z \) | \( \beta_{x_1}Z \) | \( \beta_{x_2}Z \) | \( \beta_{x_3}Z \) |
|----------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| MLE                              | Bias         | ESE          | RMSE         | Bias         | ESE          | RMSE         | Bias         | ESE          |
|                                 | -0.008       | 0.160        | 0.161        | -0.008       | 0.169        | 0.169        | -0.008       | 0.169        |
|                                 | 0.023        | 0.148        | 0.148        | 0.023        | 0.148        | 0.148        | 0.023        | 0.148        |
|                                 | -0.028       | 0.090        | 0.090        | -0.028       | 0.095        | 0.095        | -0.028       | 0.095        |
|                                 | 0.017        | 0.010        | 0.010        | 0.017        | 0.013        | 0.013        | 0.017        | 0.013        |
|                                 | -0.022       | 0.007        | 0.007        | -0.022       | 0.009        | 0.009        | -0.022       | 0.009        |

**NOTE:** ESE: empirical standard error; RMSE: root mean squared error; MLE: maximum likelihood estimator using internal study data alone; CML: constrained maximum likelihood; PCMLg: penalized constrained maximum likelihood estimators with group-wise and component-wise shrinkage, respectively. −1, −2, −2o, −12, −12o: With external study 1 only, with external study 2 only, with the second moment constraint from external study 2 only, with both external studies, with external study 1 and the second moment constraint from external study 2, respectively.
We fit the logistic regression model logit\((P(Y = 1)) = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_5 + \beta_6 \log_2(\text{Z}_1 + 1) + \beta_7 \text{Z}_2\). Here \(Y\) is the high-grade prostate cancer status, \(X_1\) is the PSA level (ng/ml), \(X_2\) is age, \(X_3\) is a binary indicator of an abnormal DRE result, \(X_5\) is a binary indicator of negative previous biopsies, \(X_5\) is a binary indicator of being African American, \(Z_1\) is the PCA3 score, and \(Z_2\) is a binary indicator dichotomized at the sample median of the T2:ERG score (Cheng et al. 2019). When fitting this model, we will incorporate the information available from Thompson et al. (2006) that led to the PCPT risk calculator, a logistic regression model given by 

\[
\hat{\beta}_1 = 0.3634 + 0.001 X_1 + 0.010 X_2 + 0.094 X_3 + 0.002 X_4 + 0.005 X_5 + 0.06 N_2.909.
\]

for the internal study cohort the median PSA level is 4.6 ng/ml and 36.5% have a PSA level ≤ 4.0 ng/ml. The heterogeneity in the study cohorts can also be clearly seen from 

\[
\tilde{\gamma} = n^{-1} \sum_{i=1}^{n} g(X_i, Z_i; \hat{\beta}_{ML}) \]

In this application we have 

\[
\tilde{\gamma} = (0.042, 0.056, 2.909, 0.006, -0.004, -0.004). \]

The large component 2.909 clearly indicates a cohort heterogeneity. On the other hand, however, the last three components of \(\gamma\) are very close to zero, showing that part of the external information may be useful to improve the internal estimation. In our analysis the group-wise shrinkage did not lead to information integration. The component-wise shrinkage did estimate the last three components of \(\gamma\) as well as the second component, exactly to be 0.

Table 3 contains the analysis results. Due to population heterogeneity, the CML estimates for the effects of DRE, prior biopsy and race are quite different from the MLE. In contrast, the PCML estimates are considerably closer to the MLE, with some effect change observed for prior biopsy and race. For the MLE, the effects of all covariates but race (indicator of being African American) are highly significant. In the internal study cohort there are only 81 African Americans, and this small number leads to the nonsignificance of the corresponding effect \((p\text{-value} = 0.788)\). The PCML method incorporates part of the information from Thompson et al’s (2006) cohort, which includes 175 African Americans. The information integration leads to a better estimate of the race effect together with a reduced standard error, resulting in a significance \((p\text{-value} = 0.002)\) that is in agreement with the general findings in existing literature. Based on the PCML method, while

### Table 2. Results of the bootstrap method for standard error calculation, with external study sample size 50,000, 1000 replications, and 200 bootstrap samples for each replication.

| Race | PCA3 | Age | Prior biopsy | T2:ERG |
|------|------|-----|--------------|--------|
| 0.125 | 0.175 | 0.175 | 0.175 | 0.175 |

**NOTE:** ESE, empirical standard error; BSE, average of the bootstrap standard errors over 1000 replications.

### Table 3. Analysis results for the prostate cancer data with \(n = 1218\).

| Estimate | MLE | Std. Err | p-value | CML | Std. Err | p-value | PCML | Std. Err | p-value |
|----------|-----|----------|---------|------|----------|---------|------|----------|---------|
| Intercept | -7.236 | 0.698 | < 0.001 | -7.050 | 0.261 | < 0.001 | -7.131 | 0.683 | < 0.001 |
| PSA | 0.638 | 0.089 | < 0.001 | 0.894 | 0.027 | < 0.001 | 0.464 | 0.143 | < 0.001 |
| Age | 0.035 | 0.010 | < 0.001 | 0.010 | 0.085 | < 0.001 | 0.033 | 0.010 | < 0.001 |
| DRE | 0.586 | 0.193 | 0.002 | 0.963 | 0.057 | < 0.001 | 0.512 | 0.181 | < 0.005 |
| Biopsy | -0.974 | 0.234 | < 0.001 | -0.293 | 0.061 | < 0.001 | -0.414 | 0.178 | < 0.020 |
| Race | -0.087 | 0.324 | 0.788 | 0.749 | 0.097 | < 0.001 | 0.594 | 0.189 | < 0.002 |
| PCA3 | 0.364 | 0.058 | < 0.001 | 0.365 | 0.064 | < 0.001 | 0.363 | 0.053 | < 0.001 |
| T2:ERG | 0.545 | 0.172 | 0.002 | 0.548 | 0.194 | 0.005 | 0.556 | 0.167 | 0.001 |

**NOTE:** MLE, maximum likelihood estimate; CML, constrained maximum likelihood; PCML, penalized constrained maximum likelihood; Std. Err, standard error. The standard errors for the PCML estimates are calculated based on 200 bootstrap samples.
having had previous negative biopsies is significantly associated with a decreased risk of high-grade prostate cancer, having a higher PSA level, older age, abnormal DRE results, being African American, and higher PCA3 and T2:REG scores are all associated with significantly increased risk.

7. Discussion

We considered two penalties in this article, the adaptive group Lasso (agLasso) penalty for group-wise selection of external studies and the adaptive Lasso (aLasso) penalty for component-wise selection of external study moment constraints. It is hard to have a general rule on when to use which penalty, as the choice may depend on many factors, including the respective covariates used by the internal and external studies, the forms they are included and the dimensions. In our experience the group-wise selection is more conservative in selecting external information. Therefore, a possible approach would be to first carry out a group-wise selection using the agLasso penalty. If none of the external studies are selected or if the selected studies only cover a small subset of the internal study covariates, a component-wise selection using the aLasso penalty can be employed. It is worth to point out that, alternative penalties may be considered to achieve the same theoretical properties. One example would be the SCAD penalty (Fan and Li 2001), which is a widely used alternative to the Lasso-type penalties. Another would be the penalties proposed by Breheny and Huang (2009) that lead to bi-level variable selection and, when applied to our data integration setting, could achieve oracle external information selection while maintaining the group structure of external studies. We will investigate this in a future project.

We made the assumption that the external study uncertainty is negligible compared to the internal study following the literature. Under some settings different from the one considered in this article, there have been recent developments on accounting for the uncertainty associated with external studies when their sample sizes are not much larger than the internal study (e.g., Han and Lawless 2019; Zhang et al. 2020). A very interesting finding is that, under certain scenarios, the external study uncertainty may reduce the internal estimation variance (Han and Lawless 2019), an observation similar to that using estimated weights helps to reduce the asymptotic variance compared to the true weights for the inverse probability weighting method in missing data literature (e.g., Robins, Rotnitzky, and Zhao 1994; Liang et al. 2004). This deserves an investigation under the setting considered in this article.

In our simulation studies, results in Tables 1 and S1 (in supplementary materials) showed the excellent performance of the PCML method for point estimation. For standard error calculation we evaluated the bootstrap method, the overall numerical performance of which seems to be acceptable. In some unreported simulation studies of the bootstrap method, as a way to account for the external information uncertainty when the external sample size was 3000, for each bootstrap sample from the internal study data we generated a value $\hat{\theta}_k$ from the normal distribution with mean $\theta_k$ and variance the corresponding covariance matrix of $\hat{\theta}_k$. We then used the bootstrap samples paired with the generated $\hat{\theta}_k$’s to compute the bootstrap PCML estimates, which lead to the bootstrap standard error. But such a way of accounting for external information uncertainty considerably overestimated the empirical standard error. As a future research topic, we will investigate standard error calculation that can properly account for the external information uncertainty.

In this article we considered that the number of external studies $K$ is finite and fixed. There are many practical settings where a large number of external studies may be available, for which a more appropriate theoretical framework would be to let $K$ increase with the internal sample size. In this case, other than the studies that provide invalid information, there may also be studies that provide redundant information when they have similar model structures and variables. Such redundant information not only affects the computation but also brings theoretical complications. We are currently developing methods that deal with this more general and more challenging setting. Some other extensions of the proposed method are also of interest. When the new covariates collected by the internal study are high-dimensional, a variable selection may be needed to build a sparse internal model. Such a setting is similar to the one in Sheng et al. (2021) and can be achieved by adding an additional penalty for variable selection. Another extension is to take into account the design of studies. In this article the internal study data are a random sample, but in practice a biased sampling is often used for data collection, such as case–control sampling, and it is of vital importance to take these study designs into consideration.

Supplementary Materials

The online supplementary materials contain an Appendix and R code.

Appendix: It contains (a) detailed proofs of all theorems, (b) an expression of Equation (4) under our simulation setting in Section 5, and (c) some additional simulation results. (Appendix.pdf)

R code: R code is provided for computing the estimator PCMLc-12 in Table 1. (Rcode-PCMLc-12.R)

Acknowledgments

We thank the Editor, Associate Editor, and two referees for their constructive comments that have helped improve the quality of the article.

References

Andrews, D. W. (1994), “Empirical Process Methods in Econometrics,” in Handbook of Econometrics (Vol. 4), pp. 2247–2294, Amsterdam: Elsevier. [1005]

Breheny, P., and Huang, J. (2009), “Penalized Methods for Bi-level Variable Selection,” Statistics and Its Interface, 2, 369–380. [1011]

Chang, J., Tang, C. Y., and Wu, T. T. (2018), “A New Scope of Penalized Empirical Likelihood with High-Dimensional Estimating Equations,” The Annals of Statistics, 46, 3185–3216. [1002]

Chatterjee, N., Chen, Y.-H., Maas, P., and Carroll, R. J. (2016), “Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-Level Information From External Big Data Sources,” Journal of the American Statistical Association, 111, 107–117. [1001,1002,1003]

Chaudhuri, S., Handcock, M. S., and Rendall, M. S. (2008), “Generalised Linear Models Incorporating Population Level Information: An Empirical Likelihood Based Approach,” Journal of the Royal Statistical Society, Series B, 70, 311–328. [1001,1003]

Chen, J., Sitter, R. R., and Wu, C. (2002), “Using Empirical Likelihood Methods to Obtain Range Restricted Weights in Regression Estimators for Surveys,” Biometrika, 89, 230–237. [1001]
