Electronic Journal of Statistics
Vol. 18 (2024) 5304–5329
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2327

Integrating external summary information under population heterogeneity and information uncertainty

Yuqi Zhai

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109,
e-mail: yqzhai@umich.edu

and

Peisong Han

Biostatistics Innovation Group, Gilead Sciences, Foster City, CA 94404,
e-mail: hanpeisong@gmail.com

Abstract: We develop a doubly penalized constrained maximum likelihood (dPCML) method for using summary information from external studies to improve estimation efficiency for an internal study that has individual-level data, in the presence of study population heterogeneity and external information uncertainty. The dPCML method can simultaneously select and incorporate the external information that agrees with the internal study while properly accounting for the uncertainty of the external information. It allows partial information where only some but not all parameter estimates from external models are reported and/or certain parameters are known to be unequal between the internal and external studies. It can still effectively account for the external information uncertainty with only external sample sizes available instead of standard errors of parameter estimates. It covers some existing data integration methods as special cases. A detailed theoretical investigation is carried out to establish asymptotic properties of the dPCML estimator, including estimation consistency, external information selection consistency, and asymptotic normality. We also provide an algorithm for implementation and conduct comprehensive simulation studies. As an application, we build an updated model to study the risk of having high-grade prostate cancer by integrating information from two widely used risk calculators.

Keywords and phrases: Constrained maximum likelihood, data integration, empirical likelihood, shrinkage, summary information, uncertainty.

Received April 2024.

Contents

1. Introduction

Data integration has become an increasingly attractive research area due to the growing availability of data from multiple sources. Integrating data from different sources can lead to a better decision process and/or more insightful conclusions compared to using a single data source. Statistical methods that leverage summary information are of particular interest because of their minimal demand on data sharing, data storage and computational power, as well as ethical considerations such as maintaining confidentiality and privacy of study participants.

Summary information from external studies can be very useful to improve parameter estimation efficiency for model fitting for an internal study of interest, especially when the internal study population is the target for inference and the internal sample size is not large. There has been a large literature on integrating external summary information, and many existing methods make the assumption that the external study populations for which the summary information is generated are the same as the internal study population of interest (e.g., Imbens and Lancaster 1994; Qin 2000; Wu and Sitter 2001; Chen et al. 2002; Chaudhuri et al. 2008; Qin et al. 2015; Chatterjee et al. 2016; Huang et al. 2016; Cheng et al. 2019; Gu et al. 2019; Huang and Qin 2020; Han et al. 2023) or the distribution of the outcome given the covariates does not differ across studies (e.g., Han and Lawless 2019; Kundu et al. 2019; Zhang et al. 2020; Sheng et al. 2021). In practice, however, such an assumption oftentimes does not hold, in which case these methods may yield substantial estimation biases for internal model parameters.

In the presence of study population heterogeneity, some authors proposed to shrink the internal study results towards the external information as a way to integrate the summary information (e.g., Estes et al. 2018; Gu et al. 2021). Such methods become less effective when the internal study is designed to target a specific population and the goal of integrating external information is to improve estimation efficiency of the internal analysis rather than shifting the analysis to align with external studies. In such a case, only the external information that agrees with the internal study population should be incorporated, as otherwise

the external information can introduce estimation bias. Taylor et al. (2023) and Choi et al. (2023) developed a method for generalized linear models with binary outcomes to integrate the ratios of coefficients from external regression models. The equality of ratio statistics across different studies is a relaxation of the assumption of homogeneous study populations, but it is still restrictive and subject to other assumptions, among which the coefficients need to be close to zero.

To be able to improve estimation efficiency without introducing estimation bias when integrating external summary information from possibly heterogeneous populations, Zhai and Han (2022) developed the penalized constrained maximum likelihood (PCML) method that simultaneously selects and incorporates the useful external information and discards the rest (see also Chen et al. 2021). The PCML method is based on the CML method (Chatterjee et al. 2016) for homogeneous study populations (see also Qin 2000; Han and Lawless 2019). The external information is formulated as moment constraints on the internal study model. The constraints corresponding to studies that target the same population as the internal study are valid and should be incorporated for efficiency improvement, and the rest constraints are invalid and should be discarded. A major assumption made by Zhai and Han (2022) is that the external study sample sizes are much larger than the internal sample size so that the uncertainty associated with the external summary information is negligible. Such an assumption is commonly made in the existing literature, including most of the aforementioned methods with exceptions such as Zhang et al. (2020). When the external information uncertainty is not properly accounted for, integrating external information may not improve the estimation efficiency for the internal study, and may even introduce estimation bias.

In this article, we consider the setting where (i) an internal study collects individual-level data to fit a parametric regression model for an outcome, (ii) some external studies have fitted less detailed regression models for the same outcome and the model fitting results are available as summary information, such as the estimated coefficients and standard errors, (iii) these external studies may target populations different from the internal study and their sample sizes may not be very large. Our goal is to incorporate only the external information that is useful to improve the efficiency of internal parameter estimation, even if the external sample sizes are not much larger than the internal one. Compared to Zhai and Han (2022), we properly take into account the uncertainty associated with the external summary information. Although we also formulate the data integration problem as a variable selection problem to deal with population heterogeneity, we quantify the difference in model parameter estimates between internal and external studies rather than the bias of moment constraints. This allows us to directly account for the uncertainty associated with the estimated coefficients from the external studies.

In addition, our proposed method allows incorporating partial summary information from external studies in cases where only some but not all estimates from external models are reported and/or certain parameters are known to be unequal between the internal and external studies. Furthermore, when stan-

dard errors of the external study parameter estimates are not available but the sample sizes are, our proposed method can still to a large degree account for the external information uncertainty. Our method covers some existing ones as special cases. In particular, it extends Zhang et al. (2020) by allowing differences between the internal and external study populations beyond only in the covariate distributions, and extends Zhai and Han (2022) by allowing the external studies to have limited sample sizes. Both extensions lead to a much wider applicability. The estimation method developed in Hu et al. (2022) also deals with both population heterogeneity and external information uncertainty, with certain computational advantages. However, for their procedure-selected functionals of the data distribution, their method requires both efficient estimators and efficient influence functions based on the internal study data, which may be difficult or impractical to construct. This is especially the case when the external study model is complex and only a subset of its parameters are selected for information integration where the relation between the selected subset and the rest parameters is unspecified, settings that are allowed by our development.

2. Method

2.1. Notation and setup

Let $(Y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$, $i = 1, \ldots, n$, denote the individual-level data from a random sample collected by the internal study, where Y is the outcome of interest, $\boldsymbol{X}$ is the vector of covariates that are routinely collected for different studies on Y, and $\boldsymbol{Z}$ is the vector of covariates that are only collected by the internal study. For example, $\boldsymbol{X}$ may include conventional covariates such as demographical variables and $\boldsymbol{Z}$ may include newly discovered biomarkers. We allow $\boldsymbol{Z}$ to be the null set if the internal study only collects $\boldsymbol{X}$. Our main interest is to fit a parametric regression model $f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ for the distribution $f(Y|\boldsymbol{X}, \boldsymbol{Z})$, where $\boldsymbol{\beta}$ is a q-dimensional vector of parameters with true value $\boldsymbol{\beta}_0$ such that $f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) = f(Y|\boldsymbol{X}, \boldsymbol{Z})$. With only the internal study data available, $\boldsymbol{\beta}_0$ can be estimated by the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}_{MLE}$ that maximizes the likelihood $\prod_{i=1}^{n} f(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta})$.

Suppose that there are K independent external studies on the same outcome Y that can potentially provide useful information to improve the efficiency of internal model parameter estimation. The kth external study, $k \in \{1, \ldots, K\}$, fits a regression model of Y on $\boldsymbol{X}_{(k)}$, where $\boldsymbol{X}_{(k)}$ is either $\boldsymbol{X}$ or a coarsened version of $\boldsymbol{X}$, such as a subset and/or a categorization of $\boldsymbol{X}$. In other words, the external study has a less detailed covariate measurement. Suppose that the fitted model can be formulated as the estimating equation $\mathbb{E}_{(k)}[\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)})] = \boldsymbol{0}$, where $\boldsymbol{\eta}_{(k)}$ is the vector of regression parameters, $\boldsymbol{h}_{(k)}\left(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}\right)$ is the estimating function determined by the external study regression model and has the same dimension as $\boldsymbol{\eta}_{(k)}$, and the expectation $\mathbb{E}_{(k)}(\cdot)$ is taken under the kth external study data distribution $f_{(k)}(Y, \boldsymbol{X}_{(k)})$. Let $\tilde{\boldsymbol{\eta}}_{(k)}^{E}$ denote the estimate of $\boldsymbol{\eta}_{(k)}$ provided by the kth external study based on its own sample with

sample size N_k, and $\boldsymbol{\eta}_{(k)}^{E*}$ the probability limit of $\tilde{\boldsymbol{\eta}}_{(k)}^{E}$ as $N_k \to \infty$ such that $\mathbb{E}_{(k)}[\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}^{E*})] = \mathbf{0}$. One example for the external study regression model is a parametric model $f_{(k)}(Y|\boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)})$ for $f_{(k)}(Y|\boldsymbol{X}_{(k)})$, in which case $\boldsymbol{h}_{(k)}\left(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}\right)$ is the corresponding score function and $\tilde{\boldsymbol{\eta}}_{(k)}^{E}$ is the solution to the score equation. Note that we allow $f_{(k)}(Y|\boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)})$ to be a misspecified model. Another example is that the kth external study provides stratified mean of Y with strata defined by the value of $\boldsymbol{X}_{(k)}$, in which case $\boldsymbol{h}_{(k)}\left(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}\right)$ is a vector of functions $(Y - \eta_{(k)}^{\mathcal{X}})I(\boldsymbol{X}_{(k)} \in \mathcal{X})$, where $\mathcal{X}$ is any stratum based on $\boldsymbol{X}_{(k)}$ and $\eta_{(k)}^{\mathcal{X}}$ is the mean of Y within this stratum under $f_{(k)}(Y|\boldsymbol{X}_{(k)})$.

The external study model can of course be fitted using the internal study data. Let $\tilde{\boldsymbol{\eta}}_{(k)}^{I}$ denote the parameter estimate from fitting the kth external study model to the internal study data such that $\sum_{i=1}^{n} \boldsymbol{h}_{(k)}(Y_i, \boldsymbol{X}_{(k)i}; \tilde{\boldsymbol{\eta}}_{(k)}^{I}) = \mathbf{0}$. Let $\boldsymbol{\eta}_{(k)}^{I*}$ denote the probability limit of $\tilde{\boldsymbol{\eta}}_{(k)}^{I}$ as $n \to \infty$ such that $\mathbb{E}[\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}^{I*})] = \mathbf{0}$, where $\mathbb{E}(\cdot)$ is the expectation under the internal data distribution $f(Y, \boldsymbol{X}, \boldsymbol{Z})$. The existence of $\boldsymbol{\eta}_{(k)}^{I*}$ follows White (1982). Assuming (i) $f_{(k)}(Y|\boldsymbol{X}_{(k)})$ is the same as $f(Y|\boldsymbol{X}_{(k)})$ such that $\boldsymbol{\eta}_{(k)}^{E*} = \boldsymbol{\eta}_{(k)}^{I*}$ and (ii) N_k is very large such that the uncertainty associated with $\tilde{\boldsymbol{\eta}}_{(k)}^{E}$ is negligible and thus $\boldsymbol{\eta}_{(k)}^{E*} = \tilde{\boldsymbol{\eta}}_{(k)}^{E}$, Chatterjee et al. (2016) proposed the CML estimator $\hat{\boldsymbol{\beta}}_{CML}$ for $\boldsymbol{\beta}_0$, defined through

$$\begin{aligned} &\max_{\boldsymbol{\beta}} \max_{p_1,\ldots,p_n} \log\left[\prod_{i=1}^{n} f(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) p_i\right] \\ &\text{subject to} \quad p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \boldsymbol{g}_{(k)}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \tilde{\boldsymbol{\eta}}_{(k)}^{E}) = \mathbf{0}, \quad k{=}1, \ldots, K \end{aligned} \tag{1}$$

where the p_i's are a discrete distribution on the internal study covariate data $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$, and

$$\boldsymbol{g}_{(k)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\eta}_{(k)}) = \int \boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}) f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) dY \tag{2}$$

such that $\mathbb{E}\left[\boldsymbol{g}_{(k)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_{(k)}^{I*})\right] = \mathbb{E}[\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}^{I*})] = \mathbf{0}$. Under Assumptions (i) and (ii), $\hat{\boldsymbol{\beta}}_{CML}$ has a higher efficiency compared to $\hat{\boldsymbol{\beta}}_{MLE}$ because of the incorporation of the external study information $\boldsymbol{\eta}_{(k)}^{E*}$.

Assumption (i) is very restrictive. For many problems it is known that certain components of $\boldsymbol{\eta}_{(k)}^{E*}$ and $\boldsymbol{\eta}_{(k)}^{I*}$ are not equal due to study population heterogeneity. For example, for an external case-control study that has a different disease prevalence, the intercept component of $\boldsymbol{\eta}_{(k)}^{E*}$ and $\boldsymbol{\eta}_{(k)}^{I*}$ is not equal, while the components corresponding to covariate effects can be the same. In the presence of a substantial population heterogeneity, there may not be any equal components between $\boldsymbol{\eta}_{(k)}^{E*}$ and $\boldsymbol{\eta}_{(k)}^{I*}$. Based on this consideration, without loss of generality, we write $\boldsymbol{\eta}_{(k)} = (\boldsymbol{\alpha}_{(k)}^{T}, \boldsymbol{\theta}_{(k)}^{T})^{T}$, where $\boldsymbol{\alpha}_{(k)}$ consists of the components known to have unequal values between the internal and external studies (i.e., $\boldsymbol{\alpha}_{(k)}^{E*} \neq \boldsymbol{\alpha}_{(k)}^{I*}$) and

$\boldsymbol{\theta}_{(k)}$ consists of the rest components. Note that the distinction between $\boldsymbol{\alpha}_{(k)}$ and $\boldsymbol{\theta}_{(k)}$ is based on prior knowledge. It is not necessary for the proposed method to have a non-null $\boldsymbol{\alpha}_{(k)}$. We allow the possibility of a non-null $\boldsymbol{\alpha}_{(k)}$ for the generality of the method. When some components of $\boldsymbol{\theta}_{(k)}^{I*}$ are indeed equal to the corresponding components of $\boldsymbol{\theta}_{(k)}^{E*}$, incorporating the value of those components provided by the external study into internal model fitting can improve the efficiency for internal model parameter estimation. Our goal is to develop methods to select these components of $\boldsymbol{\theta}_{(k)}^{E*}$ and incorporate their information to improve estimation efficiency.

Another consideration is that, in practice, an external study may report the estimated value for only some instead of all components of $\boldsymbol{\eta}_{(k)}$. For example, an study may only report estimated effect size for the risk factors of main interest even though there are additional covariates included as an effect adjustment. In this case, we will include the components of $\boldsymbol{\eta}_{(k)}$ whose estimated value is not available from the external study as part of $\boldsymbol{\alpha}_{(k)}$ as well. In other words, $\boldsymbol{\alpha}_{(k)}$ includes the components of $\boldsymbol{\eta}_{(k)}$ for which either the value is known to be unequal between the internal and external studies or the estimated value is not reported by the external study. The kth external study provides $\tilde{\boldsymbol{\theta}}_{(k)}^{E}$ as an estimate of $\boldsymbol{\theta}_{(k)}$. If $\boldsymbol{\theta}_{(k)}^{I*}$ and $\boldsymbol{\theta}_{(k)}^{E*}$ have certain equal components, then making use of the external estimate $\tilde{\boldsymbol{\theta}}_{(k)}^{E}$ may help improve estimation efficiency for internal model parameters. We will focus on the non-trivial case where $\boldsymbol{\theta}_{(k)}$ is not the null set, as otherwise we will simply exclude the kth external study from further consideration.

Assumption (ii) is also restrictive. The external study sample size N_k is not necessarily much larger than n, in which case $\tilde{\boldsymbol{\theta}}_{(k)}^{E} \neq \boldsymbol{\theta}_{(k)}^{E*}$ and the uncertainty associated with $\tilde{\boldsymbol{\theta}}_{(k)}^{E}$ needs to properly accounted for when integrating $\tilde{\boldsymbol{\theta}}_{(k)}^{E}$ into internal model fitting. The uncertainty is typically quantified by the variance $N_k^{-1}\tilde{\boldsymbol{\Sigma}}_{(k)}^{E}$ of $\tilde{\boldsymbol{\theta}}_{(k)}^{E}$, based on the asymptotic result $\sqrt{N_k}(\tilde{\boldsymbol{\theta}}_{(k)}^{E} - \boldsymbol{\theta}_{(k)}^{E*}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{(k)}^{E})$ with $\boldsymbol{\Sigma}_{(k)}^{E}$ being estimated by $\tilde{\boldsymbol{\Sigma}}_{(k)}^{E}$. Our goal is to account for the uncertainty in $\tilde{\boldsymbol{\theta}}_{(k)}^{E}$ by incorporating external information about $N_k^{-1}\tilde{\boldsymbol{\Sigma}}_{(k)}^{E}$ into the internal model fitting as well.

2.2. The dPCML method for heterogeneous populations

When some components of $\boldsymbol{\theta}_{(k)}^{I*}$ and $\boldsymbol{\theta}_{(k)}^{E*}$ are indeed equal, making use of the corresponding components of $\tilde{\boldsymbol{\theta}}_{(k)}^{E}$ provided by the external study to estimate $\boldsymbol{\beta}_0$ may help improve the estimation efficiency. To account for the fact that we do not know which components of $\boldsymbol{\theta}_{(k)}^{I*}$ and $\boldsymbol{\theta}_{(k)}^{E*}$ are equal and which ones are not as a result of study population heterogeneity, we introduce the nuisance parameters $\boldsymbol{\gamma}_{(k)}^{*}$ such that $\boldsymbol{\gamma}_{(k)}^{*} = \boldsymbol{\theta}_{(k)}^{I*} - \boldsymbol{\theta}_{(k)}^{E*}$ represents the difference between $\boldsymbol{\theta}_{(k)}^{I*}$ and $\boldsymbol{\theta}_{(k)}^{E*}$. The zero components of $\boldsymbol{\gamma}_{(k)}^{*}$ correspond to the part of the external information from study k that should be incorporated to improve the internal analysis. Since $\boldsymbol{\gamma}_{(k)}^{*}$ is unknown and needs to be estimated, it is desirable to estimate the zero components of $\boldsymbol{\gamma}_{(k)}^{*}$ to be exactly zero to select the corresponding external

information. To achieve this goal, we will impose an adaptive Lasso penalty (Zou 2006) that can consistently shrink the estimate of the zero components of $\boldsymbol{\gamma}^*_{(k)}$ to zero.

On the other hand, since the external study provides $\tilde{\boldsymbol{\theta}}^E_{(k)}$ instead of $\boldsymbol{\theta}^{E*}_{(k)}$ and the sample size N_k used to derive $\tilde{\boldsymbol{\theta}}^E_{(k)}$ is not necessarily much larger than the internal sample size n, the uncertainty associated with $\tilde{\boldsymbol{\theta}}^E_{(k)}$ needs to be properly accounted for when $\tilde{\boldsymbol{\theta}}^E_{(k)}$ is incorporated into the internal estimation of $\boldsymbol{\beta}_0$. Since $\boldsymbol{\theta}^{E*}_{(k)} = \boldsymbol{\theta}^{I*}_{(k)} - \boldsymbol{\gamma}^*_{(k)}$ is how $\boldsymbol{\theta}^{E*}_{(k)}$ and $\boldsymbol{\theta}^{I*}_{(k)}$ are connected, when the estimated variance of $\tilde{\boldsymbol{\theta}}^E_{(k)}$, i.e. $N_k^{-1}\tilde{\boldsymbol{\Sigma}}^E_{(k)}$, is also available from the external study in addition to $\tilde{\boldsymbol{\theta}}^E_{(k)}$, we can account for the uncertainty associated with $\tilde{\boldsymbol{\theta}}^E_{(k)}$ by shrinking the estimate of $\boldsymbol{\theta}^{I*}_{(k)} - \boldsymbol{\gamma}^*_{(k)}$ to the normal distribution $\mathcal{N}(\tilde{\boldsymbol{\theta}}^E_{(k)}, N_k^{-1}\tilde{\boldsymbol{\Sigma}}^E_{(k)})$.

Based on the above considerations, we propose the doubly penalized constrained maximum likelihood (dPCML) estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$ defined through

$$\begin{aligned}\max_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\gamma}} \max_{p_1,\ldots,p_n} \Bigg\{ & \log\left[\prod_{i=1}^n f_i(\boldsymbol{\beta}) p_i\right] \\ & - \sum_{k=1}^K \frac{N_k}{2} (\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)})^T \tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)} (\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) \\ & - n\lambda_n \sum_{k=1}^K \sum_{j=1}^{d_k} \frac{|\gamma_{(kj)}|}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w} \Bigg\} \\ \text{subject to} \quad & p_i \geq 0, \quad \sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i \boldsymbol{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \mathbf{0},\end{aligned} \tag{3}$$

where $f_i(\boldsymbol{\beta}) = f(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta})$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^T_{(1)}, \ldots, \boldsymbol{\alpha}^T_{(K)})^T$, $\boldsymbol{\theta} = (\boldsymbol{\theta}^T_{(1)}, \ldots, \boldsymbol{\theta}^T_{(K)})^T$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}^T_{(1)}, \ldots, \boldsymbol{\gamma}^T_{(K)})^T$, $\boldsymbol{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = [\boldsymbol{g}_{(1)}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}_{(1)}, \boldsymbol{\theta}_{(1)})^T, \ldots, \boldsymbol{g}_{(K)}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}_{(K)}, \boldsymbol{\theta}_{(K)})^T]^T$ with $\boldsymbol{g}_{(k)}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}_{(k)}, \boldsymbol{\theta}_{(k)}) = \boldsymbol{g}_{(k)}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\eta}_{(k)})$ given by (2), $|\gamma_{(kj)}||\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^{-w}$ is the adaptive Lasso (aLasso) penalty on $\gamma_{(kj)}$, the jth component of $\boldsymbol{\gamma}_{(k)}$, $j = 1, \ldots, d_k$, $\lambda_n > 0$ is the tuning parameter, and $w > 0$ is some user-specified positive number such as 1 or 2 (e.g., Zou 2006; Liao 2013). In the aLasso penalty, $\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}$ serves as a preliminary consistent estimator for $\gamma_{(kj)}$. When $\gamma_{(kj)}$ is a zero component, $\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}$ will be close to zero and thus $|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^{-w}$ imposes a heavy penalty to shrink the estimate of $\gamma_{(kj)}$ to zero in order to maximize (3).

Compared to the optimization in (1) that defines the CML estimator, the optimization in (3) is over $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ in addition to $\boldsymbol{\beta}$, with two penalties imposed. This optimization includes $\boldsymbol{\alpha}$ because $\boldsymbol{\alpha}$ consists of the values that are either known to be unequal between the internal and external studies or are not reported by the external studies. Note that $\boldsymbol{\alpha}$ is not necessarily nuisance parameters and may be parameter of interest. Information integration for components of $\boldsymbol{\theta}$ is achieved by optimizing over $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ while shrinking $\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)}$ towards

the information from external study k via the quadratic penalty and shrinking components of $\boldsymbol{\gamma}$ to zero via the aLasso penalty.

With the aLasso penalty and a properly chosen degree of shrinkage via the tuning parameter λ_n, all the zero components and only those components of $\boldsymbol{\gamma}^*$ are estimated exactly as zero, in which case the corresponding external information will be automatically incorporated into the estimation of $\boldsymbol{\beta}_0$ and the resulting dPCML estimator is consistent and has improved efficiency compared to the MLE. The aLasso penalty allows a simultaneous selection of useful external information and estimation of $\boldsymbol{\beta}_0$ incorporating that information. The uncertainty associated with the external estimate $\tilde{\boldsymbol{\theta}}^E_{(k)}$ is accounted for via the quadratic penalty on $\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)}$, adopting the idea in Zhang et al. (2020). This quadratic penalty is the kernel of the log-likelihood of a normal distribution for $\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)}$ with mean $\tilde{\boldsymbol{\theta}}^E_{(k)}$ and variance $N_k^{-1}\tilde{\boldsymbol{\Sigma}}^E_{(k)}$. When N_k is much larger compared to n, uncertainty in $\tilde{\boldsymbol{\theta}}^E_{(k)}$ is small, and the N_k factor in the quadratic penalty puts a heavy weight on the information from external study k to make $\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)}$ close to the very precise $\tilde{\boldsymbol{\theta}}^E_{(k)}$ during the optimization. On the contrary, when N_k is much smaller compared to n, uncertainty in $\tilde{\boldsymbol{\theta}}^E_{(k)}$ is big, and the N_k factor in the quadratic penalty puts a light weight on the information from external study k to diminish its contribution to the estimation of $\boldsymbol{\beta}_0$.

The proposed optimization in (3) covers some methods in the existing literature as special cases. By dropping $\boldsymbol{\gamma}$ and the aLasso penalty, the method essentially becomes the one proposed in Zhang et al. (2020) under the assumption that all study populations are the same. By dropping the quadratic penalty and replacing $\boldsymbol{\theta}$ with $\tilde{\boldsymbol{\theta}}^E + \boldsymbol{\gamma}$, the method becomes similar to the one proposed in Zhai and Han (2022) under the assumption that external information has no uncertainty. The major difference is that Zhai and Han (2022) introduced the nuisance parameters $\boldsymbol{\gamma}^*_{(k)} = \mathbb{E}[\boldsymbol{g}_{(k)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\eta}^{E*}_{(k)})]$ to represent the bias of the moment constraints resulted from the population difference, whereas the $\boldsymbol{\gamma}^*_{(k)}$ we introduced represents the difference in the population values $\boldsymbol{\theta}^{I*}_{(k)}$ and $\boldsymbol{\theta}^{E*}_{(k)}$. In the context of integrating external aggregate information into survival data analysis, Chen et al. (2021) also introduced nuisance parameters to represent the bias of the moment constraints that incorporate the external information. One major advantage of our approach, compared to Chen et al. (2021) and Zhai and Han (2022), is the flexibility in dealing with the case where only the estimate of a subset of components of $\boldsymbol{\eta}^{E*}_{(k)}$ is available instead of the whole vector. In this case, our approach can focus on the possible bias of the available subset estimate. On the contrary, it may not be possible to assess the bias of the moment constraints since it would require the availability of the estimate of the whole vector $\boldsymbol{\eta}^{E*}_{(k)}$. Another advantage of our approach is the straightforwardness in accounting for the external information uncertainty, since this uncertainty is directly for the parameter estimate.

In (3), to account for the uncertainty in $\tilde{\boldsymbol{\theta}}^E_{(k)}$, we assume that the variance matrix $N_k^{-1}\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ for $\tilde{\boldsymbol{\theta}}^E_{(k)}$ is available, which may not be the case for many external studies. In practice, oftentimes only the standard errors for the components

of $\tilde{\boldsymbol{\theta}}^E_{(k)}$, i.e. the square root of the diagonal elements of $N_k^{-1}\tilde{\boldsymbol{\Sigma}}^E_{(k)}$, are available from the external studies. In this case we can replace $N_k^{-1}\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ in (3) by the diagonal matrix with diagonal elements the squares of standard errors. There may also be situations where only the external study sample size N_k is available instead of any standard errors or variance matrix. In this case we can replace $\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ in (3) by the identity matrix. Our theoretical studies show that using these compromised solutions to account for external information uncertainty does not affect the estimation consistency of the dPCML estimator but only the efficiency (see next section for more discussion). Our numerical studies show that these compromised solutions still have clear efficiency improvement over the MLE by integrating the external information. Such observations are not surprising, since the amount of external information uncertainty to a large degree is determined by the external sample size N_k. Thus even if only N_k is available a large degree of uncertainty can be accounted for.

The aLasso penalty in (3) ensures that the integration of summary information from external study k is carried out in a component-wise manner for each component of $\tilde{\boldsymbol{\theta}}^E_{(k)}$. Such a choice of the penalty function is based on the consideration that not all components of $\boldsymbol{\theta}^{I*}_{(k)}$ are necessarily different from the corresponding components of $\boldsymbol{\theta}^{E*}_{(k)}$ even when the study populations are not the same. If one prefers to treat the information from an external study as a whole, a study-wise shrinkage can be easily achieved by replacing the aLasso penalty on $\gamma_{(kj)}$ with the adaptive group Lasso (agLasso) penalty (Wang and Leng 2008) on $\boldsymbol{\gamma}_{(k)}$, i.e. $n\lambda_n \sum_{k=1}^K \|\boldsymbol{\gamma}_{(k)}\| \|\tilde{\boldsymbol{\theta}}^I_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}\|^{-w}$, where $\|\cdot\|$ is the Euclidean norm. It is worth to point out that, the component-wise shrinkage allows us to make the maximum use of external information since the study-wise shrinkage may discard an external study completely if one component of $\boldsymbol{\theta}^{I*}_{(k)}$ and $\boldsymbol{\theta}^{E*}_{(k)}$ is different. The component-wise shrinkage can be particularly helpful when no external study information appears to be useful with a study-wise shrinkage. In this article, we will present the properties and the numerical implementation of the dPCML estimator based on component-wise shrinkage.

Using the Lagrange multiplier method, it is easy to show that the constrained optimization in (3) can be equivalently written as

$$\min_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\gamma}} \Bigg\{ -\sum_{i=1}^n \log f_i(\boldsymbol{\beta}) + \sum_{k=1}^K \frac{N_k}{2} (\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)})^T \tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(k)} (\boldsymbol{\theta}_{(k)} - \boldsymbol{\gamma}_{(k)} - \tilde{\boldsymbol{\theta}}^E_{(k)}) + n\lambda_n \sum_{k=1}^K \sum_{j=1}^{d_k} \frac{|\gamma_{(kj)}|}{|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w} + \max_{\boldsymbol{\rho}} \sum_{i=1}^n \log\{1 - \boldsymbol{\rho}^T [\boldsymbol{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})]\} \Bigg\}, \tag{4}$$

where $\boldsymbol{\rho}$ is the Lagrange multiplier. The expression in (4) is the so-called saddle-point representation in the empirical likelihood literature (e.g., Owen 2001; Newey and Smith 2004) and is used for both the derivation of the asymptotic properties in Section 2.3 and the numerical implementation in Section 3.

2.3. Asymptotic properties

This section provides some asymptotic properties of the proposed estimator and corresponding assumptions. When establishing these properties, we consider the setting where $N_k/n \to c_k \in (0,\infty)$ as $n \to \infty$, $k = 1,\ldots,K$, which means that N_k is of the same order as n and thus the uncertainty in the external summary information can not be ignored for data integration. If $c_k = 0$ then there is no need to integrate the external information, and if $c_k = \infty$ then there is no uncertainty associated with the external information, both of which are cases already considered in the existing literature.

Assumption 1. *(i)* $\mathcal{B}\times\mathcal{A}\times\mathcal{C}\times\mathcal{T}$, *the parameter space for* $(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\gamma})$, *is compact;*

(ii) $\mathbb{E}\left[\log f(Y|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})\right]$ *is uniquely maximized at* $\boldsymbol{\beta}_0 \in \mathcal{B}$*;*

(iii) $(\boldsymbol{\alpha}^{I*},\boldsymbol{\theta}^{I*})$ *is the unique solution to* $\mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\alpha},\boldsymbol{\theta})] = \mathbf{0}$*;*

(iv) $\log f(Y|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})$ *is continuous at each* $\boldsymbol{\beta}\in\mathcal{B}$ *with probability one;*

(v) $\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta})$ *is continuous at each* $(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta})\in\mathcal{B}\times\mathcal{A}\times\mathcal{C}$ *with probability one;*

(vi) $\mathbb{E}\left[\sup_{(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta})\in\mathcal{B}\times\mathcal{A}\times\mathcal{C}} \|\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta})\|^a\right] < \infty$ *for some* $a > 2$*;*

(vii) $\mathbb{E}\left[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\alpha}^{I*},\boldsymbol{\theta}^{I*})\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\alpha}^{I*},\boldsymbol{\theta}^{I*})^T\right]$ *is non-singular;*

(viii) $\sup_{\boldsymbol{\beta}\in\mathcal{B}} n^{-1/2}\sum_{i=1}^n\{\log f_i(\boldsymbol{\beta}) - \mathbb{E}[\log f(Y|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta})]\} = O_p(1)$*;*

(ix) $\sup_{(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta})\in(\mathcal{B}\times\mathcal{A}\times\mathcal{C})} n^{-1/2}\sum_{i=1}^n\{\boldsymbol{g}(\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta})]\} = O_p(1)$*;*

(x) $\lambda_n = O_p(n^{-\xi})$ *for some* ξ *with* $1/a < \xi < 1/2$.

Assumptions 1(i)–(vii) are standard ones commonly made in the literature on maximum likelihood estimator and empirical likelihood estimator (e.g., Newey and McFadden 1994; Qin and Lawless 1994; Newey and Smith 2004); (viii) and (ix) are functional Central Limit Theorem, which 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) and is a uniform version of the standard Central Limit Theorem that holds under the typical regularity conditions (e.g. Newey and McFadden 1994); (x) is an assumption on the turning parameter λ_n and ensures that the aLasso penalty function is small enough compared to the likelihood function and disappears as $n\to\infty$ to avoid introducing estimation bias.

Under Assumption 1, the consistency of $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}})$ is given by Theorem 1. The proof makes use of the saddle-point representation in (4). This proof, together with the proofs of all other theorems, is given in the Supplementary Material.

Theorem 1 (Consistency)**.** *Under Assumption 1, the estimator* $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}})$ *converges to* $(\boldsymbol{\beta}_0,\boldsymbol{\alpha}^{I*},\boldsymbol{\theta}^{I*},\boldsymbol{\gamma}^*)$ *in probability as* $n\to\infty$.

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

Assumption 2. *(i)* $(\boldsymbol{\beta}_0,\boldsymbol{\alpha}^{I*},\boldsymbol{\theta}^{I*},\boldsymbol{\gamma}^*)$ *is in the interior of* $\mathcal{B}\times\mathcal{A}\times\mathcal{C}\times\mathcal{T}$*;*

(ii) $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ *is continuously differentiable in some neighborhood* $\mathcal{B}_{\mathcal{N}} \times \mathcal{A}_{\mathcal{N}} \times \mathcal{C}_{\mathcal{N}}$ *of* $(\beta_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})$ *and* $\mathbb{E}[\sup_{(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\theta}) \in \mathcal{B}_{\mathcal{N}} \times \mathcal{A}_{\mathcal{N}} \times \mathcal{C}_{\mathcal{N}}} \|\partial \boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})/\partial \boldsymbol{\mu}\|] < \infty$, *where* $\boldsymbol{\mu}^T = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\theta}^T)$;

(iii) $\log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ *is twice continuously differentiable in some neighborhood* $\mathcal{B}_{\mathcal{N}}$ *of* $\boldsymbol{\beta}_0$ *and* $\mathbb{E}[\sup_{\boldsymbol{\beta} \in \mathcal{B}_{\mathcal{N}}} \|\partial \boldsymbol{s}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\|] < \infty$, *where* $\boldsymbol{s}(\boldsymbol{\beta}) = \partial \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})/\partial \boldsymbol{\beta}$;

(iv) $\mathbb{E}\left[\partial^2 \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T\right]$ *is non-singular;*

(v) $\mathbb{E}\left[\partial \boldsymbol{g}\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}\right)/\partial \boldsymbol{\eta}\right]$ *is non-singular, where* $\boldsymbol{\eta}^T = (\boldsymbol{\alpha}^T, \boldsymbol{\theta}^T)$;

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

Assumption 2(i)-(v) are similar to those made in Newey and McFadden (1994), Newey and Smith (2004) and Liao (2013). The $\sqrt{n}$-convergence requires that the tuning parameter converges to zero fast enough so that the aLasso penalty is asymptotically small compared to the likelihood, and (vi) specifies the convergence rate.

Theorem 2 ($\sqrt{n}$-Consistency)**.** *Under Assumptions 1 and 2, we have (i)* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$; *(ii)* $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{I*}\| = O_p(n^{-1/2}), \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{I*}\| = O_p(n^{-1/2})$, *and* $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\| = O_p(n^{-1/2})$; *and (iii)* $\hat{\boldsymbol{\rho}} = \arg\max_{\boldsymbol{\rho}} \sum_{i=1}^n \log[1 - \boldsymbol{\rho}^T \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})]$, *the Lagrange multiplier as in (4), exists with probability approaching one and* $\|\hat{\boldsymbol{\rho}}\| = O_p(n^{-1/2})$.

Consistency and $\sqrt{n}$-consistency of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$ does not imply the consistency of selection of external information that is compatible with the internal study population. Let $\mathcal{K}_{=0} = \{(k, j) : \gamma^*_{(kj)} = 0, k = 1, \ldots, K, j = 1, \ldots, d_k\}$ and $\mathcal{K}_{\neq 0} = \{(k, j) : \gamma^*_{(kj)} \neq 0, k = 1, \ldots, K, j = 1, \ldots, d_k\}$ denote the index sets for the zero and nonzero components of $\boldsymbol{\gamma}^*$, corresponding to the coefficients provided by external studies that are the same as the corresponding coefficients of the internal study and those that are different, respectively. Let $\hat{\mathcal{K}}_{=0} = \{(k, j) : \hat{\gamma}_{(kj)} = 0, k = 1, \ldots, K, j = 1, \ldots, d_k\}$ and $\hat{\mathcal{K}}_{\neq 0} = \{(k, j) : \hat{\gamma}_{(kj)} \neq 0, k = 1, \ldots, K, j = 1, \ldots, d_k\}$ denote the index sets for the zero and nonzero components of $\hat{\boldsymbol{\gamma}}$, corresponding to the external study coefficients that are selected by the dPCML method for information integration and those that are not selected, respectively. Then selection consistency means that $\hat{\mathcal{K}}_{=0}$ is the same as $\mathcal{K}_{=0}$ asymptotically.

To ensure the selection consistency, we impose the following condition on the convergence rate of the tuning parameter λ_n, which ensures that λ_n does not converge to zero too fast so that the aLasso penalty can shrink $\hat{\gamma}_{(kj)}$ to exactly zero for those $\gamma^*_{(kj)} = 0$.

Assumption 3. $n^{1/2+w/2}\lambda_n \to \infty$ *as* $n \to \infty$.

We have the following result regarding the selection consistency of external information.

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

To derive the asymptotic distribution of the proposed estimator, rewrite $\boldsymbol{\gamma}^*$ as $\boldsymbol{\gamma}^{*T} = (\boldsymbol{\gamma}^*_{\neq 0}{}^T, \boldsymbol{\gamma}^*_{=0}{}^T)$ without loss of generality, where $\boldsymbol{\gamma}^*_{\neq 0}$ contains those $\gamma^*_{(kj)}$ that $\gamma^*_{(kj)} \neq 0$ and $\boldsymbol{\gamma}^*_{=0}$ contains those $\gamma^*_{(kj)}$ that $\gamma^*_{(kj)} = 0$. Denote the dimension of $\boldsymbol{\gamma}^*_{\neq 0}$ as $d_{\neq 0}$ and the dimension of $\boldsymbol{\gamma}^*_{=0}$ as $d_{=0}$. Correspondingly, write $\boldsymbol{\theta}$ as $\boldsymbol{\theta}^T = (\boldsymbol{\theta}^T_{\neq 0}, \boldsymbol{\theta}^T_{=0})$, $\boldsymbol{\gamma}$ as $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}^T_{\neq 0}, \boldsymbol{\gamma}^T_{=0})$, and $\hat{\boldsymbol{\gamma}}$ as $\hat{\boldsymbol{\gamma}}^T = (\hat{\boldsymbol{\gamma}}^T_{\neq 0}, \hat{\boldsymbol{\gamma}}^T_{=0})$. Let $\boldsymbol{V}^E = \mathrm{diag}(c_1 \boldsymbol{\Sigma}^{E^{-1}}_{(1)}, \ldots, c_K \boldsymbol{\Sigma}^{E^{-1}}_{(K)})$, and then rearrange the rows/columns of $\boldsymbol{V}^E$ according to $\boldsymbol{\gamma}^* = (\boldsymbol{\gamma}^*_{\neq 0}{}^T, \boldsymbol{\gamma}^*_{=0}{}^T)^T$. Define $\boldsymbol{\nu}^T = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\theta}^T, \boldsymbol{\gamma}^T_{\neq 0})$, $\boldsymbol{\nu}^T_0 = (\boldsymbol{\beta}^T_0, \boldsymbol{\alpha}^{I*T}, \boldsymbol{\theta}^{I*T}, \boldsymbol{\gamma}^*_{\neq 0}{}^T)$, and $\hat{\boldsymbol{\nu}}^T = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\gamma}}^T_{\neq 0})$. Because $\hat{\boldsymbol{\gamma}}_{=0} = \boldsymbol{0}$ with probability approaching one based on Theorem 3, we just need to derive the asymptotic distribution of $\hat{\boldsymbol{\nu}}$. The result is given by the following theorem.

Theorem 4 (Asymptotic Normality)**.** *Under Assumptions 1, 2 and 3, we have*

$$\sqrt{n}(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}_0) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \left\{ \begin{bmatrix} \boldsymbol{S}_0 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}^T \boldsymbol{V}^E \boldsymbol{A} \end{bmatrix} + \begin{bmatrix} \boldsymbol{G}^T_{\boldsymbol{\mu}} \boldsymbol{\Omega}^{-1} \boldsymbol{G}_{\boldsymbol{\mu}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \right\}^{-1}),$$

where $\boldsymbol{S}_0 = \mathbb{E}[\boldsymbol{s}(\boldsymbol{\beta}_0)\boldsymbol{s}(\boldsymbol{\beta}_0)^T]$, $\boldsymbol{A} = \begin{bmatrix} \mathcal{I}_{d_{\neq 0}} & \boldsymbol{0} & -\mathcal{I}_{d_{\neq 0}} \\ \boldsymbol{0} & \mathcal{I}_{d_{=0}} & \boldsymbol{0} \end{bmatrix}$, $\boldsymbol{G}_{\boldsymbol{\mu}} = \mathbb{E}[\partial \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}) / \partial \boldsymbol{\mu}]$, $\boldsymbol{\mu}^T = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\theta}^T)$, *and* $\boldsymbol{\Omega} = \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*}) \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})^T]$.

Based on Theorem 4 we have the following corollary.

Corollary 1. *Under Assumptions 1, 2 and 3, (i)* $\hat{\boldsymbol{\beta}}$ *is asymptotically more efficient than* $\hat{\beta}_{MLE}$*, the MLE based on the internal study data alone; (ii)* $\hat{\boldsymbol{\beta}}$ *is asymptotically as efficient as the estimator for* $\boldsymbol{\beta}_0$ *that knows which components of* $\boldsymbol{\theta}^{I*}$ *and* $\boldsymbol{\theta}^{E*}$ *are equal, i.e., the estimator for* $\boldsymbol{\beta}_0$ *defined through*

$$\max_{\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\gamma}_{\neq 0}} \max_{p_1, \ldots, p_n} \left\{ \log \left[\prod_{i=1}^n f_i(\boldsymbol{\beta}) p_i \right] - \frac{1}{2} \begin{bmatrix} \boldsymbol{\theta}_{\neq 0} - \boldsymbol{\gamma}_{\neq 0} - \tilde{\boldsymbol{\theta}}^E_{\neq 0} \\ \boldsymbol{\theta}_{=0} - \tilde{\boldsymbol{\theta}}^E_{=0} \end{bmatrix}^T \tilde{\boldsymbol{V}}^E_N \begin{bmatrix} \boldsymbol{\theta}_{\neq 0} - \boldsymbol{\gamma}_{\neq 0} - \tilde{\boldsymbol{\theta}}^E_{\neq 0} \\ \boldsymbol{\theta}_{=0} - \tilde{\boldsymbol{\theta}}^E_{=0} \end{bmatrix} \right\}$$

$$\text{subject to} \quad p_i \geq 0, \quad \sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i \boldsymbol{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \boldsymbol{0},$$

where $\tilde{\boldsymbol{V}}^E_N$ *is* $diag(N_1 \tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(1)}, \ldots, N_K \tilde{\boldsymbol{\Sigma}}^{E^{-1}}_{(K)})$ *with rows/columns rearranged according to* $\boldsymbol{\gamma}^* = (\boldsymbol{\gamma}^*_{\neq 0}{}^T, \boldsymbol{\gamma}^*_{=0}{}^T)^T$.

All the above results are established by using $\tilde{\boldsymbol{\Sigma}}^E_{(k)}$, a consistent estimate of $\boldsymbol{\Sigma}^E_{(k)}$ provided by the external studies in addition to the estimate $\tilde{\boldsymbol{\theta}}^E_{(k)}$, to account for the uncertainty associated with $\tilde{\boldsymbol{\theta}}^E_{(k)}$. It turns out that Theorems 1, 2 and 3 still hold even if $\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ is not consistent for $\boldsymbol{\Sigma}^E_{(k)}$. These three theorems remain valid if $\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ is replaced by any positive definite matrix with dimension equal to that of $\boldsymbol{\theta}_{(k)}$. In particular, when only the standard errors for the components of $\tilde{\boldsymbol{\theta}}^E_{(k)}$ are available, $\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ can be replaced by a diagonal matrix based on the standard errors. When only the external study sample size N_k is available instead of any standard errors, $\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ can be replaced with the identity matrix. Consistency of estimation and information selection remains valid. The

asymptotic distribution in Theorem 4 will, however, be different. It is hard to establish a clear comparison as in Corollary 1 in this case, but our simulation studies show that the proposed estimator still has efficiency improvement over the MLE by integrating the external information.

3. Implementation

3.1. Implementation based on saddle-point representation

The numerical implementation of the proposed dPCML method is based on the saddle-point representation (4) 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 $\boldsymbol{\rho}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ at a given value of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$, and the outer loop updates $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\gamma})$.

Specifically, the inner loop is $\max_{\boldsymbol{\rho}} \sum_{i=1}^{n} \log\left\{1 - \boldsymbol{\rho}^T[\boldsymbol{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})]\right\}$ as in (4). When the given value $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is close to the true value $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^{I*}, \boldsymbol{\theta}^{I*})$, which is indeed the case during the implementation if the initial value of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is taken to be the consistent estimator $(\hat{\boldsymbol{\beta}}_{MLE}, \tilde{\boldsymbol{\alpha}}^I, \tilde{\boldsymbol{\theta}}^I)$, 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 $\boldsymbol{\rho} = \mathbf{0}$ because of Theorem 2.

To present the outer loop, let $\hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ denote the computed Lagrange multiplier from the inner loop at a given $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$. The outer loop computes the dPCML estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$ in the following steps.

Step 0. Take the initial value $(\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)}, \hat{\boldsymbol{\gamma}}^{(0)}) = (\hat{\boldsymbol{\beta}}_{MLE}, \tilde{\boldsymbol{\alpha}}^I, \tilde{\boldsymbol{\theta}}^I, \tilde{\boldsymbol{\theta}}^I - \tilde{\boldsymbol{\theta}}^E)$.

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

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

$$\left| \frac{N_k}{n} \left[\tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}}\right]_{j.} \left[\hat{\boldsymbol{\theta}}_{(k)}^{(l)} - \hat{\boldsymbol{\gamma}}_{(k)}^{\left(l+\frac{j}{d_k}\right)}(0) - \tilde{\boldsymbol{\theta}}_{(k)}^{E}\right] \right| < \frac{\lambda_n}{|\tilde{\theta}_{(kj)}^I - \tilde{\theta}_{(kj)}^E|^w} \tag{5}$$

and equal to the root of the equation

$$\frac{\lambda_n}{|\tilde{\theta}_{(kj)}^I - \tilde{\theta}_{(kj)}^E|^w} \frac{\hat{\gamma}_{(kj)}}{|\hat{\gamma}_{(kj)}|} - \frac{N_k}{n} \left[\tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}}\right]_{j.} \left[\hat{\boldsymbol{\theta}}_{(k)}^{(l)} - \hat{\boldsymbol{\gamma}}_{(k)}^{\left(l+\frac{j}{d_k}\right)}(\gamma_{(kj)}) - \tilde{\boldsymbol{\theta}}_{(k)}^{E}\right] = 0 \tag{6}$$

as an equation for $\gamma_{(kj)}$ if (5) does not hold, where $\left[\tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}}\right]_{j.}$ denotes the jth row of $\tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}}$, and $\hat{\boldsymbol{\gamma}}_{(k)}^{\left(l+\frac{j}{d_k}\right)}(\gamma_{(kj)}) = \left[\hat{\gamma}_{(k,1)}^{(l+1)}, \ldots, \hat{\gamma}_{(k,j-1)}^{(l+1)}, \gamma_{(kj)}, \hat{\gamma}_{(k,j+1)}^{(l)}, \ldots, \hat{\gamma}_{(k,d_k)}^{(l)}\right]^T$.

Step 2. Set $(\hat{\boldsymbol{\alpha}}^{(l+1)}, \hat{\boldsymbol{\theta}}^{(l+1)})$ equal to the root of the equation

$$\begin{cases} \sum_{i=1}^{n} \frac{\{\partial \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})/\partial \boldsymbol{\alpha}\}^T \hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})}{1 - \left[\hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})\right]^T \left[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})\right]} = \mathbf{0} \\ -N_1 \tilde{\boldsymbol{\Sigma}}_{(1)}^{E^{-1}} \left\{ \boldsymbol{\theta}_{(1)} - (\tilde{\boldsymbol{\theta}}_{(1)}^{E} + \hat{\boldsymbol{\gamma}}_{(1)}^{(l+1)}) \right\} + \sum_{i=1}^{n} \frac{\{\partial \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}_{(1)}\}^T \hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})}{1 - \left[\hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})\right]^T \left[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})\right]} = \mathbf{0} \\ \vdots \\ -N_K \tilde{\boldsymbol{\Sigma}}_{(K)}^{E^{-1}} \left\{ \boldsymbol{\theta}_{(K)} - (\tilde{\boldsymbol{\theta}}_{(K)}^{E} + \hat{\boldsymbol{\gamma}}_{(K)}^{(l+1)}) \right\} + \sum_{i=1}^{n} \frac{\{\partial \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}_{(K)}\}^T \hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})}{1 - \left[\hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})\right]^T \left[\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^{(l)}, \boldsymbol{\alpha}, \boldsymbol{\theta})\right]} = \mathbf{0} \end{cases} \tag{7}$$

as an equation for $(\boldsymbol{\alpha}, \boldsymbol{\theta})$.

Step 3. Set $\hat{\boldsymbol{\beta}}^{(l+1)}$ equal to the root of the equation

$$\sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\beta}) + \sum_{i=1}^{n} \frac{\{\partial \boldsymbol{g}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}^{(l+1)}, \hat{\boldsymbol{\theta}}^{(l+1)})/\partial \boldsymbol{\beta}\}^T \hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}^{(l+1)}, \hat{\boldsymbol{\theta}}^{(l+1)})}{1 - [\hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}^{(l+1)}, \hat{\boldsymbol{\theta}}^{(l+1)})]^T [\boldsymbol{g}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}^{(l+1)}, \hat{\boldsymbol{\theta}}^{(l+1)})]} = \mathbf{0}. \tag{8}$$

as an equation for $\boldsymbol{\beta}$.

Step 4. Repeat **Steps 1**–**3** until convergence such that $\|\hat{\boldsymbol{\beta}}^{(l+1)} - \hat{\boldsymbol{\beta}}^{(l)}\|$, $\|\hat{\boldsymbol{\alpha}}^{(l+1)} - \hat{\boldsymbol{\alpha}}^{(l)}\|$, $\|\hat{\boldsymbol{\theta}}^{(l+1)} - \hat{\boldsymbol{\theta}}^{(l)}\|$, and $\|\hat{\boldsymbol{\gamma}}^{(l+1)} - \hat{\boldsymbol{\gamma}}^{(l)}\|$ are smaller than some pre-specified small number and $\hat{\mathcal{K}}_{=0}^{(l+1)} = \hat{\mathcal{K}}_{=0}^{(l)}$, where $\hat{\mathcal{K}}_{=0}^{(l)} = \{(k, j) : \hat{\gamma}_{(kj)}^{(l)} = 0, k = 1, \ldots, K, j = 1, \ldots, d_k\}$.

Equations (6), (7) and (8) are the first-order condition of the saddle-point representation (4) with respect to $\gamma_{(kj)}$ when $\gamma_{(kj)} \neq 0$, $(\boldsymbol{\alpha}, \boldsymbol{\theta})$ and $\boldsymbol{\beta}$, respectively, treating $\hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ as an implicit function of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$. 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 (7) and (8) needs to again treat $\hat{\boldsymbol{\rho}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ as an implicit function of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$. The expression of the Jacobian matrix for (8) is the same as that in Han and Lawless (2019) and the expression for (7) can be similarly derived. Details are omitted here due to their lengthy expressions.

3.2. Tuning parameter selection

The rate of convergence of the tuning parameter λ_n is crucial when deriving the asymptotic properties of the dPCML estimator, and Assumptions 2(vi) and 3 specify some sufficient conditions on the convergence rate that guarantee the $\sqrt{n}$-convergence of the dPCML estimator and the information selection consistency. For practical implementation, however, we need an effective way of selecting a concrete value for the tuning parameter.

Note from (5) that $\gamma_{(kj)}^{*}$ is estimated exactly as zero if

$$\left| \frac{N_k}{\sqrt{n}} \left[\tilde{\boldsymbol{\Sigma}}_{(k)}^{E^{-1}} \right]_{j\cdot} \left[\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\gamma}}_{(k,-j)} - \tilde{\boldsymbol{\theta}}_{(k)}^{E} \right] \right| < \frac{\sqrt{n} \lambda_n}{|\tilde{\theta}_{(kj)}^{I} - \tilde{\theta}_{(kj)}^{E}|^w}, \tag{9}$$

where $\hat{\boldsymbol{\gamma}}_{(k,-j)} = (\hat{\gamma}_{(k,1)}, \ldots, \hat{\gamma}_{(k,j-1)}, 0, \hat{\gamma}_{(k,j+1)}, \ldots, \hat{\gamma}_{(k,d_k)})^T$.

For any $\gamma^*_{(kj)} \neq 0$, the left-hand side of (9) is asymptotically bounded away from zero, in which case to avoid estimating $\gamma^*_{(kj)}$ to be zero $\sqrt{n}\lambda_n$ needs to converge to zero as fast as possible, since $|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}|^w$ converges to a non-zero constant. With all $\gamma^*_{(kj)} \neq 0$ estimated as non-zeros, for any $\gamma^*_{(kj)} = 0$, the left-hand side of (9) is of order $O_p(1)$, and in addition $|\tilde{\theta}^I_{(kj)} - \tilde{\theta}^E_{(kj)}| = O_p(n^{-1/2})$. Therefore, to estimate $\gamma^*_{(kj)} = 0$ exactly as zero $n^{1/2+w/2}\lambda_n$ needs to diverge to infinity as fast as possible. These considerations agree with Assumptions 2(vi) and 3. To balance these rate requirements on λ_n, we choose $\lambda_n = Cn^{-1/2-w/4}$, where C is a positive constant. We did an exploration of the idea in Liao (2013) to select C and found that the numerical performance with selected C was similar to that with $C = 1$ when the covariance matrix for $\tilde{\boldsymbol{\theta}}^E_{(k)}$ or the standard errors for the components of $\tilde{\boldsymbol{\theta}}^E_{(k)}$ are available as a quantification of the uncertainty, but was worse when only the sample size N_k is available. Thus we recommend to take $C = 1$ in implementation, which also avoids the complex procedure of selecting C.

4. Simulation studies

4.1. Simulation setup

The internal study has covariates $X_1, X_2, \ldots, X_5$ and Z_1, Z_2, where $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{125})$ with unit variances, correlation coefficients $\rho_{12} = \rho_{25} = 0.3$ and $\rho_{15} = 0.2$, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(1), $X_4 \sim$ Bernoulli(0.4), and $\boldsymbol{Z}|\boldsymbol{X} \sim \mathcal{N}((X_1 + X_3, X_1 - X_3), \boldsymbol{\Sigma}_{\boldsymbol{Z}})$ with unit variances and correlation coefficient 0.2. Given $\boldsymbol{X}$ and $\boldsymbol{Z}$, Y is generated from a Bernoulli distribution with $\text{logit}\{P(Y = 1|\boldsymbol{X}, \boldsymbol{Z})\} = (1, X_1, \ldots, X_5, Z_1, Z_2, X_1Z_1)\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0^T = (1, 0.5, -1.5, 1, -1, 0.5, -0.5, 0.5, 1)$. The internal study model is the logistic regression $\text{logit}\{P(Y = 1|\boldsymbol{X}, Z)\} = \beta_c + \beta_{X_1}X_1 + \cdots + \beta_{X_5}X_5 + \beta_{Z_1}Z_1 + \beta_{Z_2}Z_2 + \beta_{X_1Z_1}X_1Z_1$ with $\boldsymbol{\beta}^T = (\beta_c, \beta_{X_1}, \ldots, \beta_{X_5}, \beta_{Z_1}, \beta_{Z_2}, \beta_{X_1Z_1})$ having true value $\boldsymbol{\beta}_0$.

We consider three external studies. For Study 1 the data are generated as $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((-0.5, -0.5, 0), \boldsymbol{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, $X_3 \sim$ Exponential(1.25), X_4 and $\boldsymbol{Z}|\boldsymbol{X}$ follow the same distributions as in the internal study, Y follows a Bernoulli distribution with $\text{logit}\{P(Y = 1|\boldsymbol{X}, \boldsymbol{Z})\} = (1, X_1, \ldots, X_5, Z_1, Z_2, X_1Z_1)\boldsymbol{\beta}_{1*}$ and $\boldsymbol{\beta}_{1*}^T = (0.75, 1, -1, 0.75, -1, 0.8, -0.6, 0.75, 0.75)$. Study 1 measures only Y, X_4 and X_5 to fit the logistic regression model $\text{logit}\{P(Y = 1|X_4, X_5)\} = \theta_{(1,1)} + \theta_{(1,2)}X_4 + \theta_{(1,3)}X_5$. Some numerical calculation based on a large sample size 10^6 for both the internal study and Study 1 shows that $\boldsymbol{\gamma}^*_{(1)} = (\gamma^*_{(1,1)}, \gamma^*_{(1,2)}, \gamma^*_{(1,3)})^T = (0.622, 0.001, -0.212)^T$, with the second component almost zero.

Study 2 has the same data distribution as the internal study and measures only Y, X_1, X_2 and X_5 to fit the logistic regression model $\text{logit}\{P(Y = 1|X_1, X_2, X_5)\} = \theta_{(2,1)} + \theta_{(2,2)}X_1 + \theta_{(2,3)}X_2 + \theta_{(2,4)}X_5$. It is clear that $\boldsymbol{\gamma}^*_{(2)} = (\gamma^*_{(2,1)}, \gamma^*_{(2,2)}, \gamma^*_{(2,3)}, \gamma^*_{(2,4)})^T = (0, 0, 0, 0)^T$.

For Study 3 the data are generated as $(X_1, \tilde{X}_2, X_5) \sim \mathcal{N}((0, 0.5, 0.5), \mathbf{\Sigma}_{125})$, $X_2 = I(\tilde{X}_2 > 0)$, X_3 and X_4 follow the same distributions as the internal study, Y follows a Bernoulli distribution with $\text{logit}\{P(Y = 1|\boldsymbol{X})\} = (1, X_1, X_2, \ldots, X_5)$ $(\alpha^{I*}_{(3,1)} - 0.5, \alpha^{I*}_{(3,2)} + 0.5, \theta^{I*}_{(3,1)} - 0.5, \theta^{I*}_{(3,2)}, \theta^{I*}_{(3,3)}, \theta^{I*}_{(3,4)})^T$, where $(\boldsymbol{\alpha}^{I*^T}_{(3)}, \boldsymbol{\theta}^{I*^T}_{(3)})^T = (\alpha^{I*}_{(3,1)}, \alpha^{I*}_{(3,2)}, \theta^{I*}_{(3,1)}, \theta^{I*}_{(3,2)}, \theta^{I*}_{(3,3)}, \theta^{I*}_{(3,4)})^T$ is derived by fitting the corresponding logistic regression model to a data set with sample size 10^6 generated under the internal data distribution. Study 3 measures Y and $X_1, X_2, \ldots, X_5$ to fit the logistic regression model $\text{logit}\{P(Y = 1|\boldsymbol{X})\} = \alpha_{(3,1)} + \alpha_{(3,2)}X_1 + \theta_{(3,1)}X_2 + \theta_{(3,2)}X_3 + \theta_{(3,3)}X_4 + \theta_{(3,4)}X_5$. After model fitting, Study 3 provides information about $\theta_{(3,1)}$, $\theta_{(3,2)}$, $\theta_{(3,3)}$ and $\theta_{(3,4)}$, but not $\alpha_{(3,1)}$ and $\alpha_{(3,2)}$. It is clear that $\boldsymbol{\gamma}^*_{(3)} = (\gamma^*_{(3,1)}, \gamma^*_{(3,2)}, \gamma^*_{(3,3)}, \gamma^*_{(3,4)})^T = (-0.5, 0, 0, 0)^T$.

For the three external studies, $\boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\eta}_{(k)}) = \boldsymbol{h}_{(k)}(Y, \boldsymbol{X}_{(k)}; \boldsymbol{\alpha}_{(k)}, \boldsymbol{\theta}_{(k)})$ is the score function for the corresponding external logistic regression model, where $\boldsymbol{X}_{(1)} = (X_4, X_5)$, $\boldsymbol{X}_{(2)} = (X_1, X_2, X_5)$ and $\boldsymbol{X}_{(3)} = (X_1, X_2, X_3, X_4, X_5)$. Here both $\boldsymbol{\alpha}_{(1)}$ and $\boldsymbol{\alpha}_{(2)}$ are the null set, while $\boldsymbol{\alpha}_{(3)} = (\alpha_{(3,1)}, \alpha_{(3,2)})^T$, for which Study 3 does not provide any information. The three external studies provide the estimates $\tilde{\boldsymbol{\theta}}^E_{(k)}$. For the uncertainty associated with $\tilde{\boldsymbol{\theta}}^E_{(k)}$, we consider three scenarios: (i) the variance matrices $N_k^{-1}\tilde{\mathbf{\Sigma}}^E_{(k)}$ for $\tilde{\boldsymbol{\theta}}^E_{(k)}$ are available from external studies, (ii) only the standard errors for the components of $\tilde{\boldsymbol{\theta}}^E_{(k)}$ are available, and (iii) only N_k are available.

We consider two sample sizes, $n = 300$ and 800, for the internal study. The external study sample sizes are set as $N_1 = 3n$, $N_2 = 2n$ and $N_3 = n$ for Studies 1, 2, and 3, respectively, in order to be consistent with our assumption that $N_k/n \to c_k > 0$ as $n \to \infty$ and the consideration that studies which collect more covariates may have smaller sample sizes due to budget or technical constraints. We summarize the results based on 1000 replications. Each replication regenerates both the internal and the external data. We take $w = 2$ in (3) for the aLasso penalty.

To make comparisons, in addition to the MLE using internal study data alone, we also include the CML estimator of Chatterjee et al (2016), the generalized integration method (GIM) estimator of Zhang et al. (2020), the optimal covariance weighted (OCW) estimator and the selective coefficient learner (SCL) of Gu et al. (2021), and the component-wise PCML estimator of Zhai and Han (2022). Since the CML, OCW, SCL and PCML estimators do not deal with cases where only the information about some subset of external regression coefficients is available, Study 3 is discarded when computing these estimators. The CML and PCML estimators do not account for the uncertainty of external information. For the OCW and SCL estimators we make use of $N_k^{-1}\tilde{\mathbf{\Sigma}}^E_{(k)}$ to account for the uncertainty in $\tilde{\boldsymbol{\theta}}^E_{(k)}$. The GIM method makes use of N_k only since it assumes population homogeneity and computes the covariance matrix. The OCW and SCL estimators are computed by the R package "MetaIntegration" (Gu et al. 2021) and the GIM estimator is computed by the R package "gim" (Zhang and Yu 2020).

4.2. Simulation observations

From Tables 1 and 2, it is seen that our proposed estimator (dPCML) has substantial efficiency improvement without introducing bias, compared to the MLE, by integrating external study information and properly accounting for the associated uncertainty. When only the standard errors for the components of $\tilde{\boldsymbol{\theta}}^E_{(k)}$ are available from external studies instead of the variance matrices $N_k^{-1}\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ as a quantification of the uncertainty, the performance stays almost the same. When only N_k is available, the improvement over MLE becomes smaller but is still substantial. The observation that the proposed estimator remains unbiased even if the external uncertainty can only be quantified by the sample size is in full agreement with the discussion at the end of Section 2.3.

As a comparison, the CML estimator has a substantial bias because of the heterogeneity between Study 1 and the internal study data distributions. Moreover, compared to the MLE, the CML estimator may even have larger empirical standard errors since it does not account for the uncertainty in the external information. The OCW and SCL estimators are unbiased but the reduction in empirical standard errors compared to the MLE is not as impressive as our proposed estimator. The PCML estimator has no clear-cut improvement over the MLE since its bias is not negligible when $n = 300$ and its empirical standard errors are not necessarily smaller, due to ignoring the external information uncertainty. The GIM estimator is clearly biased although it has a substantial reduction of empirical standard errors compared to the MLE, due to the study population heterogeneity.

Following Zhai and Han (2022), we recommend the bootstrap method for standard error calculation for the proposed method. As an assessment, in Tables 1 and 2 we include the mean of bootstrap standard errors for dPMCL-i based on 200 bootstrap resamples for each replication. From a comparison to the empirical standard errors, the bootstrap standard errors overall seem to have a reasonable performance.

Table 3 presents the percentage of estimating $\gamma^*_{(kj)}$ exactly as zero by our proposed method. It is seen that, as n increases from 300 to 800, the percentage of estimating the $\gamma^*_{(kj)} = 0$ as zero increases and the percentage of estimating $\gamma^*_{(kj)} \neq 0$ as zero decreases, in full agreement with the selection consistency of external information. The selection rate stays overall the same when either only the standard errors for the components of $\tilde{\boldsymbol{\theta}}^E_{(k)}$ are available from external studies or the variance matrices $N_k^{-1}\tilde{\boldsymbol{\Sigma}}^E_{(k)}$ are available. When only the sample sizes N_k are available, the selection rate becomes lower, in agreement with the previous observation that in this case the efficiency improvement over MLE is smaller.

5. Data application

We apply the proposed dPCML 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 demographic and clinical variables such as age, race, the prostate specific

TABLE 1
Simulation results summarized based on 1000 *replications with internal sample size* $n = 300$ *and external sample sizes* $N_1 = 3n$, $N_2 = 2n$, $N_3 = n$.

		β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_{X_4}	β_{X_5}	β_{Z_1}	β_{Z_2}	$\beta_{X_1Z_1}$
	Bias	0.054	0.019	-0.064	0.038	-0.069	0.030	-0.027	0.013	0.063
MLE	ESE	0.351	0.304	0.358	0.335	0.342	0.178	0.180	0.171	0.182
	RMSE	0.355	0.305	0.364	0.337	0.349	0.180	0.182	0.172	0.193
	Bias	0.035	0.026	-0.099	0.040	-0.045	0.077	-0.026	0.013	0.063
dPCML-i	ESE	0.283	0.279	0.274	0.323	0.270	0.130	0.180	0.171	0.182
	BSE	0.243	0.279	0.257	0.333	0.242	0.117	0.192	0.182	0.202
	RMSE	0.286	0.281	0.292	0.326	0.274	0.151	0.182	0.172	0.193
	Bias	0.020	0.026	-0.093	0.038	-0.042	0.081	-0.026	0.013	0.063
dPCML-ii	ESE	0.284	0.279	0.274	0.324	0.270	0.130	0.180	0.171	0.182
	RMSE	0.285	0.280	0.290	0.326	0.273	0.153	0.182	0.172	0.193
	Bias	0.055	0.039	-0.115	0.035	-0.069	0.065	-0.026	0.013	0.063
dPCML-iii	ESE	0.307	0.282	0.305	0.334	0.286	0.142	0.180	0.171	0.183
	RMSE	0.312	0.285	0.326	0.336	0.294	0.156	0.182	0.172	0.193
	Bias	0.262	0.402	-0.161	0.036	-0.375	0.059	-0.027	0.015	-0.051
CML	ESE	0.332	0.350	0.344	0.347	0.326	0.157	0.185	0.178	0.215
	RMSE	0.423	0.533	0.380	0.348	0.497	0.168	0.187	0.179	0.221
	Bias	0.050	0.018	-0.061	0.038	-0.069	0.029	-0.026	0.013	0.063
OCW	ESE	0.332	0.291	0.321	0.335	0.342	0.160	0.180	0.171	0.182
	RMSE	0.336	0.291	0.327	0.337	0.348	0.162	0.182	0.172	0.193
	Bias	0.038	0.018	-0.062	0.038	-0.067	0.034	-0.026	0.013	0.063
SCL	ESE	0.339	0.291	0.320	0.335	0.333	0.166	0.180	0.171	0.182
	RMSE	0.342	0.291	0.326	0.337	0.340	0.169	0.182	0.172	0.193
	Bias	-0.131	0.143	-0.232	0.043	-0.067	0.101	-0.026	0.013	0.061
GIM	ESE	0.255	0.279	0.257	0.320	0.250	0.116	0.180	0.171	0.183
	RMSE	0.287	0.314	0.347	0.323	0.259	0.154	0.182	0.172	0.193
	Bias	0.173	0.022	-0.068	0.038	-0.432	0.081	-0.026	0.013	0.062
PCML	ESE	0.383	0.288	0.318	0.335	0.574	0.158	0.180	0.171	0.182
	RMSE	0.421	0.289	0.325	0.337	0.718	0.178	0.182	0.172	0.193

[1] ESE: empirical standard error. RMSE: root mean squared error. BSE: mean of bootstrap standard error over 1000 replications based on 200 bootstrap resamples for each replication.
[2] CML: constrained maximum likelihood (Chatterjee et al. 2016). OCW: optimal covariance weighted (Gu et al. 2021). SCL: selective coefficient learner (Gu et al. 2021). GIM: generalized integration method (Zhang et al. 2020). PCML: the PCML method (Zhai and Han 2022).
[3] −i, −ii, −iii: using $\tilde{\mathbf{\Sigma}}_{(k)}$, diag($\tilde{\mathbf{\Sigma}}_{(k)}$) and $\mathcal{I}_{d_k}$ in (3).

TABLE 2
Simulation results summarized based on 1000 *replications with internal sample size* $n = 800$ *and external sample sizes* $N_1 = 3n$, $N_2 = 2n$, $N_3 = n$.

		β_c	β_{X_1}	β_{X_2}	β_{X_3}	β_{X_4}	β_{X_5}	β_{Z_1}	β_{Z_2}	$\beta_{X_1Z_1}$
	Bias	0.003	0.000	-0.015	0.021	-0.012	0.004	-0.005	0.009	0.024
MLE	ESE	0.192	0.164	0.207	0.208	0.191	0.105	0.107	0.104	0.110
	RMSE	0.193	0.164	0.208	0.209	0.192	0.105	0.107	0.105	0.113
	Bias	0.007	0.004	-0.044	0.020	-0.003	0.043	-0.005	0.009	0.024
dPCML-i	ESE	0.149	0.152	0.165	0.201	0.141	0.079	0.107	0.104	0.110
	BSE	0.132	0.154	0.139	0.185	0.129	0.063	0.107	0.103	0.110
	RMSE	0.149	0.152	0.171	0.202	0.141	0.090	0.107	0.105	0.113
	Bias	-0.003	0.003	-0.034	0.019	-0.005	0.046	-0.005	0.009	0.024
dPCML-ii	ESE	0.151	0.153	0.167	0.201	0.141	0.079	0.107	0.104	0.110
	RMSE	0.151	0.153	0.171	0.202	0.141	0.092	0.107	0.105	0.113
	Bias	0.006	0.004	-0.037	0.016	-0.011	0.029	-0.005	0.009	0.024
dPCML-iii	ESE	0.168	0.156	0.183	0.205	0.156	0.083	0.107	0.104	0.110
	RMSE	0.168	0.156	0.187	0.206	0.156	0.088	0.107	0.105	0.113
	Bias	0.226	0.367	-0.101	0.030	-0.362	0.047	-0.015	0.015	-0.072
CML	ESE	0.207	0.219	0.216	0.216	0.183	0.095	0.108	0.110	0.138
	RMSE	0.307	0.428	0.239	0.218	0.406	0.106	0.109	0.111	0.155
	Bias	0.001	-0.001	-0.013	0.021	-0.012	0.005	-0.005	0.009	0.024
OCW	ESE	0.181	0.158	0.187	0.208	0.191	0.094	0.107	0.104	0.110
	RMSE	0.181	0.158	0.188	0.209	0.192	0.094	0.107	0.105	0.113
	Bias	-0.004	-0.001	-0.013	0.021	-0.012	0.006	-0.005	0.009	0.024
SCL	ESE	0.186	0.158	0.187	0.208	0.189	0.098	0.107	0.104	0.110
	RMSE	0.186	0.158	0.188	0.209	0.190	0.099	0.107	0.105	0.113
	Bias	-0.170	0.123	-0.181	0.023	-0.021	0.084	-0.004	0.009	0.022
GIM	ESE	0.145	0.156	0.155	0.199	0.139	0.071	0.107	0.104	0.110
	RMSE	0.223	0.199	0.238	0.201	0.141	0.109	0.107	0.105	0.112
	Bias	0.019	-0.006	-0.015	0.021	-0.077	0.028	-0.005	0.009	0.023
PCML	ESE	0.203	0.159	0.179	0.208	0.309	0.096	0.107	0.104	0.110
	RMSE	0.204	0.159	0.180	0.209	0.318	0.100	0.107	0.105	0.113

[1] ESE: empirical standard error. RMSE: root mean squared error. BSE: mean of bootstrap standard error over 1000 replications based on 200 bootstrap resamples for each replication.
[2] CML: constrained maximum likelihood (Chatterjee et al. 2016). OCW: optimal covariance weighted (Gu et al. 2021). SCL: selective coefficient learner (Gu et al. 2021). GIM: generalized integration method (Zhang et al. 2020). PCML: the PCML method (Zhai and Han 2022).
[3] −i, −ii, −iii: using $\tilde{\mathbf{\Sigma}}_{(k)}$, $\text{diag}(\tilde{\mathbf{\Sigma}}_{(k)})$ and $\mathcal{I}_{d_k}$ in (3).

TABLE 3
The percentage (%) of estimating $\gamma^*_{(kj)}$ as zero, summarized based on 1000 replications with external sample sizes $N_1 = 3n$, $N_2 = 2n$, $N_3 = n$.

	$\gamma^*_{(kj)} \neq 0$			$\gamma^*_{(kj)} = 0$							
	$\gamma_{(1,1)}$	$\gamma_{(1,3)}$	$\gamma_{(3,1)}$	$\gamma_{(1,2)}$	$\gamma_{(2,1)}$	$\gamma_{(2,2)}$	$\gamma_{(2,3)}$	$\gamma_{(2,4)}$	$\gamma_{(3,2)}$	$\gamma_{(3,3)}$	$\gamma_{(3,4)}$
						$n = 300$					
dPCML-i	0.9	56.9	49.2	83.9	85.7	89.8	82.2	89.4	92.1	81.5	88.5
dPCML-ii	1.2	57.7	50.1	84.3	89.1	90.7	86.4	89.9	91.9	80.4	88.5
dPCML-iii	0.5	38.9	26.9	70.3	76.8	85.4	65.9	75.0	82.5	56.7	66.8
						$n = 800$					
dPCML-i	0.0	27.0	27.5	91.3	92.5	96.4	86.9	94.0	95.2	89.3	92.6
dPCML-ii	0.0	27.8	28.9	91.7	94.8	97.0	91.1	94.6	94.8	89.8	93.1
dPCML-iii	0.0	13.4	10.3	78.7	87.4	92.0	78.9	82.1	88.5	67.4	73.9

[1] −i, −ii, −iii: using $\tilde{\boldsymbol{\Sigma}}_{(k)}$, diag($\tilde{\boldsymbol{\Sigma}}_{(k)}$) and $\mathcal{I}_{d_k}$ in (3).

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) in the United States. This PCPT risk calculator is the first online prostate cancer risk assessment tool and is among the most widely used ones. The model behind this risk calculator, together with the parameter estimates and 95% confidence intervals, is provided in Thompson et al. (2006) as follows: $\text{logit}(P(Y = 1)) = -6.25 + 1.29\log(X_1) + 0.03X_2 + 1.00X_3 - 0.36X_4 + 0.96X_5$, where 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_4 is a binary indicator of negative previous biopsies, and X_5 is a binary indicator of being African American.

Previous studies have also shown that the prostate volume is related to PSA level (e.g., Bohnen et al. 2007), and should be taken into account when assessing men for prostate cancer risk (e.g., Al-Azab et al. 2007). The European Randomized Study of Screening for Prostate Cancer Risk Calculator 3 (ERSPC-RC3) (Roobol et al. 2012) is one of the validated tools for prostate cancer risk assessment that include transrectal ultrasound prostate volume (TRUS-PV) as a predictor. Developed based on data from 3616 men, the ERSPC-RC3 is modeled as $\text{logit}(P(Y = 1)) = \log(0.03) + \log(3.24)\overline{\log_2(X_1)} + \log(6.13)X_3 + \log(0.22)\overline{\log_2(X_6)}$, where X_6 is TRUS-PV reclassified in three categories (25, 40, and 60 cm^3), and the lines over $\log_2(X_1)$ and $\log_2(X_6)$ imply that they are centered. The 95% confidence intervals for all the coefficient estimates are also reported in Roobol et al. (2012).

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 risk factors $(X_1, \ldots, X_6)$ and the new biomarkers on the risk of prostate cancer after adjusting for each other. We use part of the sample collected in Tomlins et al. (2016) as the internal data, which consists of 1218 men presenting for diagnostic prostate biopsy at seven community clinics throughout the United States. We fit the logistic regression model $\text{logit}(P(Y=1)) = \beta_c + \beta_1 \log_2(X_1) + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 \overline{\log_2(X_6)} + \beta_7 \log_2(Z_1+1) + \beta_8 Z_2$, where 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). The final sample size of the internal study is $n = 1174$ after removing subjects with missing TRUS-PV.

When fitting the internal study model, we will incorporate the information from the two aforementioned external risk calculators. Note that the sample sizes for both external studies are not very large compared to the internal one, and thus the uncertainty associated with the external parameter estimates should be properly accounted for. Moreover, there are some apparent differences between the internal study data distribution and the distribution reported in Thompson et al. (2006) (see Zhai and Han 2022). The covariance matrices of the parameter estimates are not reported by the external studies, but we can obtain the standard errors from the reported 95% confidence intervals. Note that, due to the centering of $\log_2(X_1)$, the intercept of the ERSPC-RC3 model is different from that of the internal study model, and thus we discard the information of the intercept from the ERSPC-RC3 model.

The external study estimates are $\tilde{\boldsymbol{\theta}}^E_{(1)} = (-6.25, 1.29, 0.03, 1.00, -0.36, 0.96)^T$ for PCPT and $\tilde{\boldsymbol{\theta}}^E_{(2)} = (\log(3.24), \log(6.13), \log(0.22))^T$ for ERSPC-RC3, which leads to $\tilde{\boldsymbol{\gamma}}_{(1)} = \tilde{\boldsymbol{\theta}}^I_{(1)} - \tilde{\boldsymbol{\theta}}^E_{(1)} = (-0.11, -0.39, 0.02, -0.37, -0.62, -1.03)^T$ and $\tilde{\boldsymbol{\gamma}}_{(2)} = \tilde{\boldsymbol{\theta}}^I_{(2)} - \tilde{\boldsymbol{\theta}}^E_{(2)} = (-0.40, -1.03, 0.24)^T$. The non-zero components of $\tilde{\boldsymbol{\gamma}}_{(1)}$ and $\tilde{\boldsymbol{\gamma}}_{(2)}$ clearly indicate study population heterogeneity. On the other hand, some components of $\tilde{\boldsymbol{\gamma}}$ are small, implying that part of the external information may be useful to improve the internal estimation. Indeed, in our analysis the first, third and fourth components of $\boldsymbol{\gamma}_{(1)}$ and the last component of $\boldsymbol{\gamma}_{(2)}$ are estimated exactly as zero.

Table 4 contains the analysis results. The dPCML estimates are close to the MLE since the dPCML method does not introduce estimation bias by discarding the external information that is incompatible with the internal data. The efficiency improvement of dPCML over MLE is apparent from the substantially reduced standard errors. Both the MLE and the proposed method show that, while both negative previous biopsies and larger prostate volume are associated with significantly decreased risk of high-grade prostate cancer, higher PSA level, older age, abnormal DRE results, and higher PCA3 and T2:REG scores are all associated with significantly increased risk. The information integration leads to substantially reduced standard errors for the estimates of the effects of abnormal DRE and prostate volume.

TABLE 4
Analysis results for the prostate cancer data with $n = 1174$.

	MLE			dPCML		
	Estimate	Std. Err	P-value	Estimate	Std. Err	P-value
Intercept	−8.124	0.739	< 0.001	−7.973	0.429	< 0.001
PSA	0.733	0.094	< 0.001	0.899	0.080	< 0.001
Age	0.045	0.011	< 0.001	0.035	0.007	< 0.001
DRE	0.617	0.198	0.002	0.877	0.121	< 0.001
Biopsy	−0.793	0.240	0.001	−0.671	0.231	0.004
Race	−0.297	0.375	0.429	−0.205	0.349	0.557
TRUS-PV	−1.351	0.203	< 0.001	−1.491	0.108	< 0.001
PCA3	0.307	0.061	< 0.001	0.307	0.057	< 0.001
T2:ERG	0.630	0.180	< 0.001	0.632	0.192	0.001

[1] Std. Err: standard error. The standard errors for the dPCML estimates are calculated based on 200 bootstrap samples.

6. Discussion

We proposed a doubly penalized constrained maximum likelihood (dPCML) method for using summary-level information from external studies while building a refined regression model based on individual-level data collected in an internal study. For existing methods, incorporating external information increases efficiency of the parameter estimates for the internal model, without introducing biases, under either one or both of the assumptions that (1) the internal and external studies are conducted for the same population and (2) the external datasets are very big such that the uncertainty associated with external information is negligible. These two assumptions are both restrictive. The proposed dPCML method is robust to departures from both these two assumptions. It can simultaneously select and incorporate the external information that agrees with the internal study while properly accounting for the uncertainty associated with the external information.

The dPCML method is very flexible in several aspects. First, it allows incorporating partial summary information from external studies in cases where only some but not all estimates from external models are reported and/or certain parameters are known to be unequal between the internal and external studies. Second, it allows different covariate transformations for different external models (e.g., in data application the PCPT calculator uses log(PSA) while the ERSPC-RC3 uses $\overline{\log_2(\text{PSA})}$). Third, even with only the external sample sizes available, the dPCML method can still to a large degree account for external information uncertainty and improve efficiency over the MLE.

Some extensions of the proposed method are of possible interest as future work. For example, when the new covariates collected by the internal study are high-dimensional, variable selection may be needed for internal model fitting, which can be achieved by adding an additional penalty. Another possible extension is to take into account the design of studies. In this paper we considered a random sample for the internal study. However, in practice, biased sampling is often used for data collection, such as case-control sampling, and it is of vi-

tal importance to take these study designs into consideration. In addition, our method requires that the external studies use less detailed covariates than the internal study. But some external studies may use variables that are not collected by the internal study, and it is worthwhile to explore methods that can address such situations. Another question of interest is whether the proposed estimator achieves the efficiency bound associated with the model class restricted by the assumptions considered in this paper. This investigation may be done by following Zhang et al. (2020) and Hu et al. (2022) but is beyond the scope of this current paper.

Acknowledgments

The authors would like to thank the Editor, Associate Editor, and two referees for their helpful comments that improved the quality of this work.

References

[1] Al-Azab, R., Toi, A., Lockwood, G., Kulkarni, G. S. and Fleshner, N. (2007). Prostate volume is strongest predictor of cancer diagnosis at transrectal ultrasound-guided prostate biopsy with prostate-specific antigen values between 2.0 and 9.0 ng/mL. *Urology* **69** 103–107.

[2] Andrews, D. W. K. (1994). Chapter 37 Empirical process methods in econometrics. In *Handbook of Econometrics*, **4** 2247–2294. Elsevier, Amsterdam. MR1315972

[3] Bohnen, A. M., Groeneveld, F. P. and Bosch, J. L. H. R. (2007). Serum prostate-specific antigen as a predictor of prostate volume in the community: the Krimpen study. *European Urology* **51** 1645–1653.

[4] 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. MR3494641

[5] 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 (Statistical Methodology)* **70** 311–328. MR2424755

[6] 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. MR1888365

[7] Chen, Z., Ning, J., Shen, Y. and Qin, J. (2021). Combining primary cohort data with external aggregate information without assuming comparability. *Biometrics* **77** 1024–1036. https://doi.org/10.1111/biom.13356. MR4320675

[8] Cheng, W., Taylor, J. M. G., Gu, T., Tomlins, S. A. and Mukherjee, B. (2019). Informing a risk prediction model for binary outcomes with

external coefficient information. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **68** 121–139. https://doi.org/10.1111/rssc.12306. MR3902985

[9] Choi, K., Taylor, J. M. and Han, P. (2023). Robust data integration from multiple external sources for generalized linear models with binary outcomes. *Biometrics* To Appear. MR4565447

[10] Estes, J. P., Mukherjee, B. and Taylor, J. M. G. (2018). Empirical Bayes estimation and prediction using summary-level information from external big data sources adjusting for violations of transportability. *Statistics in Biosciences* **10** 568–586.

[11] Gu, T. and Mukherjee, B. (2021). MetaIntegration: Ensemble Meta-Prediction Framework R package version 0.1.2.

[12] Gu, T., Taylor, J. M. G. and Mukherjee, B. (2021). A meta-inference framework to integrate multiple external models into a current study. *Biostatistics*. kxab017. https://doi.org/10.1093/biostatistics/kxab017. MR4578352

[13] Gu, T., Taylor, J. M. G., Cheng, W. and Mukherjee, B. (2019). Synthetic data method to incorporate external information into a current study. *Canadian Journal of Statistics* **47** 580–603. MR4035790

[14] Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association* **109** 1159–1173. MR3265688

[15] Han, P. and Lawless, J. F. (2019). Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statistica Sinica* **29** 1321–1342. MR3932520

[16] Han, P., Taylor, J. M. G. and Mukherjee, B. (2023). Integrating information from existing risk prediction models with no model details. *Canadian Journal of Statistics* **51** 355–374. MR4595233

[17] Hu, W., Wang, R., Li, W. and Miao, W. (2022). Paradoxes and resolutions for semiparametric fusion of individual and summary data. *arXiv preprint arXiv:2210.00200*.

[18] Huang, C.-Y., Qin, J. and Tsai, H.-T. (2016). Efficient estimation of the Cox model with auxiliary subgroup survival information. *Journal of the American Statistical Association* **111** 787–799. MR3538705

[19] Huang, C.-Y. and Qin, J. (2020). A unified approach for synthesizing population-level covariate effect information in semiparametric estimation with survival data. *Statistics in Medicine* **39** 1573–1590. MR4098508

[20] Imbens, G. W. and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *Review of Economic Studies* **61** 655–680. MR1299309

[21] Kitamura, Y. (2007). Empirical likelihood methods in econometrics: Theory and practice. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, (R. Blundell, W. Newey and T. Persson, eds.). *Econometric Society Monographs* **3** 174–237. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511607547.008. MR2352814

[22] Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference. Springer Series in Statistics*. Springer New York, New York, NY. MR2724368

[23] Kundu, P., Tang, R. and Chatterjee, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika* **106** 567–585. https://doi.org/10.1093/biomet/asz030. MR3992390

[24] Liao, Z. (2013). Adaptive GMM shrinkage estimation with consistent moment selection. *Econometric Theory* **29** 857–904. https://doi.org/10.1017/S0266466612000783. MR3148818

[25] Newey, W. K. and McFadden, D. (1994). Chapter 36 Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, **4** 2111–2245. Elsevier, Amsterdam. MR1315971

[26] Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72** 219–255. MR2031017

[27] Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall/CRC, Boca Raton, Florida.

[28] Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika* **87** 484–490. MR1782493

[29] Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22** 300–325. MR1272085

[30] Qin, J., Zhang, H., Li, P., Albanes, D. and Yu, K. (2015). Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika* **102** 169–180. MR3335103

[31] Roobol, M. J., van Vugt, H. A., Loeb, S., Zhu, X., Bul, M., Bangma, C. H., van Leenders, A. G. L. J. H., Steyerberg, E. W. and Schröder, F. H. (2012). Prediction of prostate cancer risk: the role of prostate volume and digital rectal examination in the ERSPC risk calculators. *European Urology* **61** 577–583.

[32] Sheng, Y., Sun, Y., Huang, C. and Kim, M. (2021). Synthesizing external aggregated information in the presence of population heterogeneity: A penalized empirical likelihood approach. *Biometircs* 1–12. MR4450586

[33] Taylor, J. M., Choi, K. and Han, P. (2023). Data integration: Exploiting ratios of parameter estimates from a reduced external model. *Biometrika* **110** 119–134. https://doi.org/10.1093/biomet/asac022. MR4565447

[34] Thompson, I. M., Ankerst, D. P., Chi, C., Goodman, P. J., Tangen, C. M., Lucia, M. S., Feng, Z., Parnes, H. L. and Coltman Jr, C. A. (2006). Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. *Journal of the National Cancer Institute* **98** 529–534.

[35] Tomlins, S. A., Day, J. R., Lonigro, R. J., Hovelson, D. H., Siddiqui, J., Kunju, L. P., Dunn, R. L., Meyer, S., Hodge, P., Groskopf, J., Wei, J. T. and Chinnaiyan, A. M. (2016). Urine TMPRSS2:ERG Plus PCA3 for individualized prostate cancer risk assessment.

European Urology **70** 45–53.

[36] VAN DER VAART, A. W. (2000). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. MR1652247

[37] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes. Springer Series in Statistics*. Springer New York, New York, NY. MR1385671

[38] WANG, H. and LENG, C. (2008). A note on adaptive group lasso. *Computational Statistics and Data Analysis* **52** 5277–5286. MR2526593

[39] WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163

[40] WU, C. and SITTER, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association* **96** 185–193. MR1952731

[41] ZHAI, Y. and HAN, P. (2022). Data integration with oracle use of external information from heterogeneous populations. *Journal of Computational and Graphical Statistics* **31** 1001–1012. https://doi.org/10.1080/10618600.2022.2050248. MR4513365

[42] ZHANG, H. and YU, K. (2020). gim: Generalized Integration Model R package version 0.33.1.

[43] ZHANG, H., DENG, L., SCHIFFMAN, M., QIN, J. and YU, K. (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika* **107** 689–703. MR4138984

[44] ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469