Statistica Sinica Preprint No: SS-2025-0225	
Title	Integrating External Summary Information Via James-stein Shrinkage
Manuscript ID	SS-2025-0225
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0225
Complete List of Authors	Peisong Han, Haoyue Li and Jeremy M G Taylor
Corresponding Authors	Peisong Han
E-mails	peisong@umich.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

Integrating External Summary Information via James-Stein Shrinkage

Peisong Han, Haoyue Li, Jeremy M.G. Taylor

Biostatistics Innovation Group, Gilead Sciences

Department of Statistics, Pennsylvania State University

Department of Biostatistics, University of Michigan

Abstract:

Consider fitting a general parametric regression model, such as a generalized linear model, with individual data. It is common to have summary information, such as parameter estimates, available from external studies that use similar regression models. Many methods have been developed to incorporate this external information into internal model fitting to improve parameter estimation. Some of these methods aim to reduce estimation variance without introducing estimation bias that could result from study population heterogeneity. Others allow introduction of bias in exchange for substantial variance reduction, based on the bias-variance trade-off consideration. We take the latter approach and develop James-Stein shrinkage estimators to integrate the external information. These estimators can reduce the asymptotic risk compared to not using the external information, regardless of the degree of heterogeneity between internal and external populations. This is a highly desirable property as it provides a safe passage for

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

the utility of external information. Few existing methods provide such a guaranteed improvement. We also conduct simulation studies and apply the method to a prostate cancer dataset to illustrate the numerical performance.

Key words and phrases: data integration, efficiency, mean squared error, population heterogeneity, risk

1. Introduction

Data integration has become increasingly important in recent years due to the need to leverage information from various sources to provide better answers to scientific questions. We consider a data integration problem where (i) a study, referred to as the internal study, collects individual data to fit a general parametric regression model, such as a generalized linear model, and (ii) an external study provides some estimated parameters under a regression model with possibly less detailed covariates with no individual data available. The goal is to integrate the external summary information to improve the internal model estimation. This setting is of substantial interest in practice, as it is common for summary information to be found from existing studies.

Many methods have been developed to deal with the aforementioned data integration problem, including Qin (2000),Chatterjee et al. (2016),

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

Huang et al. (2016), Cheng et al. (2018), Han and Lawless (2019), Kundu et al. (2019), Zhang et al. (2020), Chen et al. (2021), Taylor et al. (2023), Zhai and Han (2022), Zhai and Han (2024). Most of these methods aim to improve the efficiency of the internal study estimation without introducing estimation bias when integrating external information. Alternatively, some methods have been developed to allow for the introduction of estimation bias in exchange for a reduction of estimation variance, based on the bias-variance trade-off consideration. These methods include Estes et al. (2018), Gu et al. (2019), Hector and Martin (2024), Han et al. (2024), among others.

In particular, Han et al. (2024) developed a James-Stein shrinkage estimator (James and Stein, 1961) for the data integration problem for linear regression models. The estimator guarantees improvement in terms of prediction mean squared errors even if some estimation bias may be introduced when integrating external information due to study population heterogeneity. In addition, this guarantee is regardless of the degree of heterogeneity, although severe heterogeneity can result in a negligible improvement. This guaranteed improvement is highly desirable, since study population heterogeneity is almost inevitable when data are integrated from different sources. Existing methods, including those based on empirical Bayes such as Estes et al. (2018) and Gu et al. (2019), do not provide theoretical results of

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

such a guaranteed improvement. Numerical evidence in Han et al. (2024) for Gaussian linear regression shows the superior performance compared to these competitors.

In this paper, we develop James-Stein shrinkage estimators for data integration for a general parametric regression model. This development covers the generalized linear model for categorical outcomes as a special case. This is a significant extension of Han et al. (2024) because of the much broader applicability of a generalized linear model. We focus on the positive-part James-Stein estimators, as they have been shown to dominate the original James-Stein estimator in terms of mean squared error for Gaussian distributions (Baranchik, 1964). Our theoretical treatment adopts the framework in Hansen (2016).

We show that, when the shrinkage dimension is greater than 2, the proposed estimators are guaranteed to have a reduced asymptotic risk (e.g., mean squared error) after integrating the external information, regardless of the degree of heterogeneity between internal and external study populations. This property is appealing because it provides a safe passage to utilize available external information to improve internal model estimation, despite that the improvement may not be substantial in the presence of severe population heterogeneity. To the best of our knowledge, there have

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

been no existing methods in the literature that provide such a guaranteed improvement when integrating external information.

The setting we consider is different from those in the literature of transfer learning and federated learning, although all of these are about integrating data from multiple studies/sources, with or without the assumption of transportability that some data distribution characteristics are shared across studies. Transfer learning in general focuses on how to transfer common information shared between studies. In doing so, transfer learning often considers settings where individual data are accessible from all studies and the same model is fitted to all study data (e.g. Li et al. (2022); Tian and Feng (2023)). Federated learning, on the other hand, does not require the sharing of individual data. However, the summary statistics computed and shared across studies are under the command of a central site, which dictates what summary statistics should be computed and shared (e.g. Guo et al. (2025); Han et al. (2025)). In contrast, in our setting only estimated regression coefficients from the external study are needed, and the external model can be different from the internal model with fewer covariates. Oftentimes, the external estimates are provided by historical studies, from which the internal study takes whatever is available to integrate into its model fitting without the capability of demanding new computations from

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

the external study. This is different from those settings considered by transfer learning and federated learning, and thus the existing methods under those learnings do not directly apply to our setting.

2. The Proposed James-Stein Estimators

2.1 Notation and Setup

The main interest is to study the association between the response Y and certain covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$. We purposely separate $\boldsymbol{X}$ and $\boldsymbol{Z}$ so that, for example, $\boldsymbol{X}$ includes those demographical covariates that are also measured by the external study and $\boldsymbol{Z}$ includes biomarkers only measured by the internal study. We allow $\boldsymbol{Z}$ to be a null set.

Let $f(Y \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\gamma})$ denote a parametric model for the density $f(Y \mid \boldsymbol{X}, \boldsymbol{Z})$, which corresponds to a regression model in this paper. Here $\boldsymbol{\beta}$ is the vector of regression coefficients and is of main interest and $\boldsymbol{\gamma}$ is the vector of nuisance parameters, with true values $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ for the internal study population such that $f(Y \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = f(Y \mid \boldsymbol{X}, \boldsymbol{Z})$. For example, if Y is continuous and $f(Y \mid \boldsymbol{X}, \boldsymbol{Z})$ is a normal distribution corresponding to a linear regression, then $\boldsymbol{\beta}$ contains the regression coefficients and $\boldsymbol{\gamma}$ contains the variance parameters. If Y is binary and is modeled by a logistic regression, then $\boldsymbol{\beta}$ contains the regression coefficients and $\boldsymbol{\gamma}$ is a

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

null vector. If Y is count and is modeled by a Poisson regression, then $\boldsymbol{\beta}$ contains the regression coefficients and $\boldsymbol{\gamma}$ is again a null vector. With the independent and identically distributed data $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$, collected by the internal study, the MLE $\hat{\boldsymbol{\beta}}_{mle}$ for $\boldsymbol{\beta}_0$ is defined through maximizing $\prod_{i=1}^{n} f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\gamma})$ over both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

The external study collected data on the same response Y and some covariates $\boldsymbol{X}^*$ and built a regression model for Y on $\boldsymbol{X}^*$, where $\boldsymbol{X}^*$ is a coarsened version of $\boldsymbol{X}$, such as a subset or a categorization of the continuous components of $\boldsymbol{X}$. This is based on the consideration that the internal study usually builds a more detailed regression model, sometimes by taking a finer measurement on some covariates. The external study produced an estimate $\boldsymbol{\theta}_*$ for some parameter $\boldsymbol{\theta}$ introduced when building the regression model for Y on $\boldsymbol{X}^*$. In this paper we do not assume the availability of the standard error of $\boldsymbol{\theta}_*$. In general, $\boldsymbol{\theta}_*$ is derived by solving a set of estimating equations that are the sample version of $E^*\{\boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta})\} = \mathbf{0}$ based on the external study data that we do not have access to. Here the expectation E^* is taken under the external joint distribution for $(Y, \boldsymbol{X}^*)$, and the estimating function $\boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta})$ is specified by the external regression model and is known to us. In this paper we consider $\boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta})$ to be differentiable. For example, for linear regression $\boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta}) = \boldsymbol{X}^*(Y - \boldsymbol{X}^{*\mathrm{T}}\boldsymbol{\theta})$, for logistic re-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

gression $\boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta}) = \boldsymbol{X}^*\{Y - \text{expit}(\boldsymbol{X}^{*\text{T}}\boldsymbol{\theta})\}$, and for Poisson regression $\boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta}) = \boldsymbol{X}^*\{Y - \exp(\boldsymbol{X}^{*\text{T}}\boldsymbol{\theta})\}$, all of which are the corresponding score functions. In this paper we treat $\boldsymbol{\theta}_*$ as if it were derived based on an infinite external sample size so that $E^*\{\boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta}_*)\} = \mathbf{0}$. This will make the exposition easier without changing the theoretical conclusions. We will discuss this point in the Discussion Section.

2.2 The Shrinkage Target

To construct a James-stein estimator, we need a good target that incorporates the external study information and towards which the MLE will be shrunk. The derivation of a good target requires a connection between the internal model parameters $\boldsymbol{\beta}$ and the external model parameters $\boldsymbol{\theta}$. Note that although our main focus is on estimation in the presence of study population heterogeneity, the connection between $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ needs to be established when the internal and external distributions are the same, because otherwise it depends on the unknown heterogeneity and thus varies case by case. With shrinkage targets constructed assuming no heterogeneity, the resulting JS estimators still improve over the MLE even when heterogeneity exists.

We first establish a connection between $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Under the same

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

data distribution, the external information becomes $E\{\boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta}_*)\} = \mathbf{0}$, where the expectation E is under the internal joint distribution. Define

$$\boldsymbol{U}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \int \boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta}) f(Y \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\gamma}) dY,$$

whose dependence on $\boldsymbol{X}^*$ is through $\boldsymbol{X}$ since $\boldsymbol{X}^*$ is a coarsened version of $\boldsymbol{X}$, then $\boldsymbol{U}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{\theta}) = E\{\boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta}) \mid \boldsymbol{X}, \boldsymbol{Z}\}$. For example, with linear regression such that $\boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta}) = \boldsymbol{X}^*(Y - \boldsymbol{X}^{*\mathrm{T}}\boldsymbol{\theta})$, we have $\boldsymbol{U}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \boldsymbol{X}^*\{(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}})\boldsymbol{\beta} - \boldsymbol{X}^{*\mathrm{T}}\boldsymbol{\theta}\}$ with $\boldsymbol{\gamma}$ being the variance parameter that does not appear in this function. With logistic regression such that $\boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta}) = \boldsymbol{X}^*\{Y - \mathrm{expit}(\boldsymbol{X}^{*\mathrm{T}}\boldsymbol{\theta})\}$, we have $\boldsymbol{U}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \boldsymbol{X}^*[\mathrm{expit}\{(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}})\boldsymbol{\beta}\} - \mathrm{expit}(\boldsymbol{X}^{*\mathrm{T}}\boldsymbol{\theta})]$, where $\boldsymbol{\gamma}$ is the null vector. With Poisson regression such that $\boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta}) = \boldsymbol{X}^*\{Y - \exp(\boldsymbol{X}^{*\mathrm{T}}\boldsymbol{\theta})\}$, we have $\boldsymbol{U}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \boldsymbol{X}^*[\exp\{(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}})\boldsymbol{\beta}\} - \exp(\boldsymbol{X}^{*\mathrm{T}}\boldsymbol{\theta})]$, where $\boldsymbol{\gamma}$ is the null vector. It is then easy to see that

$$E_{(\boldsymbol{X}, \boldsymbol{Z})}\{\boldsymbol{U}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{\theta}_*)\} = \mathbf{0}, \tag{2.1}$$

where the expectation $E_{(\boldsymbol{X}, \boldsymbol{Z})}$ is taken under the internal covariate distribution.

Equation (2.1) represents the connection between $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Based on this connection, to fix a shrinkage target when constructing the James-Stein estimator we consider two constrained maximum likelihood (CML)

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

estimators. One is $\hat{\boldsymbol{\beta}}_{cmle-sp}$ defined through

$$\max_{\boldsymbol{\beta},\boldsymbol{\gamma}} \prod_{i=1}^{n} f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}) \quad \text{subject to} \quad \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}_*) = \mathbf{0}, \tag{2.2}$$

where "sp" stands for "simple" sample average in the above constraints. Another one is $\hat{\boldsymbol{\beta}}_{cmle-el}$ defined through

$$\max_{\boldsymbol{\beta},\boldsymbol{\gamma},q_1,\ldots,q_n} \prod_{i=1}^{n} \{f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}) q_i\} \quad \text{subject to}$$
$$q_i > 0, \sum_{i=1}^{n} q_i = 1, \sum_{i=1}^{n} q_i \boldsymbol{U}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}_*) = \mathbf{0}, \tag{2.3}$$

where $q_i = dF(\boldsymbol{X}_i, \boldsymbol{Z}_i)$ is a discrete distribution on the internal study covariate data, and "el" stands for the "empirical likelihood" nature of the above constraints.

It is clear that (2.2) is a special case of (2.3) by taking $q_i \equiv n^{-1}$. The formulation (2.3) considers a likelihood based on the joint distribution for $(Y, \boldsymbol{X}, \boldsymbol{Z})$ where $f(Y \mid \boldsymbol{X}, \boldsymbol{Z})$ is modeled parametrically and $F(\boldsymbol{X}, \boldsymbol{Z})$ is modeled nonparametrically. Such a formulation has been considered in Qin (2000), Chatterjee et al. (2016), Han and Lawless (2019), among others. We use both $\hat{\boldsymbol{\beta}}_{cmle-sp}$ and $\hat{\boldsymbol{\beta}}_{cmle-el}$ as shrinkage targets when constructing the JS estimators. When the internal and external study populations are the same, both estimators are consistent for $\boldsymbol{\beta}_0$ with $\hat{\boldsymbol{\beta}}_{cmle-el}$ being more efficient since (2.2) is obtained by taking $q_i \equiv 1/n$ in (2.3) instead of being

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

maximized (Han and Lawless, 2016). In the presence of study population heterogeneity, $\hat{\boldsymbol{\beta}}_{cmle-sp}$ and $\hat{\boldsymbol{\beta}}_{cmle-el}$ may no longer be consistent for $\boldsymbol{\beta}_0$, but consistency of the shrinkage target is not necessary for the construction of the JS estimators.

2.3 The Proposed Estimators

The JS estimators we propose shrink the MLE $\hat{\boldsymbol{\beta}}_{mle}$ towards the constrained MLE $\hat{\boldsymbol{\beta}}_{cmle}$ (i.e. $\hat{\boldsymbol{\beta}}_{cmle-sp}$ or $\hat{\boldsymbol{\beta}}_{cmle-el}$) and are defined as

$$\hat{\boldsymbol{\beta}}_{JS} = \hat{w}\hat{\boldsymbol{\beta}}_{mle} + (1-\hat{w})\hat{\boldsymbol{\beta}}_{cmle}$$

with the shrinkage weight

$$\hat{w} = \left(1 - \frac{\hat{\tau}}{n(\hat{\boldsymbol{\beta}}_{mle} - \hat{\boldsymbol{\beta}}_{cmle})^{\mathrm{T}}\hat{\boldsymbol{V}}_{\boldsymbol{\beta},mle}^{-1}(\hat{\boldsymbol{\beta}}_{mle} - \hat{\boldsymbol{\beta}}_{cmle})}\right)_{+}. \tag{2.4}$$

Here $(a)_+ = a1(a \geq 0)$ is the positive part of any quantity a, and $\hat{\tau}$ is the computed value for a scalar shrinkage parameter τ that controls the degree of shrinkage. The detailed expression for $\hat{\tau}$ is given by (3.4) after the derivation of the asymptotic risk of $\hat{\boldsymbol{\beta}}_{JS}$. The $\hat{\boldsymbol{V}}_{\boldsymbol{\beta},mle}$ is a consistent estimator for the asymptotic variance $\boldsymbol{V}_{\boldsymbol{\beta},mle}$ of the MLE $\hat{\boldsymbol{\beta}}_{mle}$ with $\sqrt{n}(\hat{\boldsymbol{\beta}}_{mle} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta},mle})$.

The shrinkage weight $\hat{w}$ is always between 0 and 1 and makes intuitive sense. A large distance between $\hat{\boldsymbol{\beta}}_{mle}$ and $\hat{\boldsymbol{\beta}}_{cmle}$ would lead to a heavy

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

weight on $\hat{\boldsymbol{\beta}}_{mle}$, which should be the case intuitively because a large distance between $\hat{\boldsymbol{\beta}}_{mle}$ and $\hat{\boldsymbol{\beta}}_{cmle}$ implies the constraints on $\boldsymbol{\beta}$ constructed by integrating the external information may be incorrect and thus more trust should be given to $\hat{\boldsymbol{\beta}}_{mle}$. However, regardless of how large the distance between $\hat{\boldsymbol{\beta}}_{mle}$ and $\hat{\boldsymbol{\beta}}_{cmle}$ is, the weight $\hat{w}$ on $\hat{\boldsymbol{\beta}}_{mle}$ is always less than 1 so that the reduced variation in $\hat{\boldsymbol{\beta}}_{cmle}$ due to the parameter constraints can help reduce the variance of $\hat{\boldsymbol{\beta}}_{JS}$ compared to $\hat{\boldsymbol{\beta}}_{mle}$, even if a non-zero weight on $\hat{\boldsymbol{\beta}}_{cmle}$ may introduce some estimation bias. This bias-variance trade off leads to a better overall numerical performance for $\hat{\boldsymbol{\beta}}_{JS}$ in terms of the (weighted) mean squared error that is detailed in the next section. With a small distance between $\hat{\boldsymbol{\beta}}_{mle}$ and $\hat{\boldsymbol{\beta}}_{cmle}$, the weight on $\hat{\boldsymbol{\beta}}_{mle}$ is small and may become exactly zero, the benefit of which can again be seen from a bias-variance trade off. The distance between $\hat{\boldsymbol{\beta}}_{mle}$ and $\hat{\boldsymbol{\beta}}_{cmle}$ is scaled by $\hat{\boldsymbol{V}}^{-1}_{\boldsymbol{\beta},mle}$ to prevent it from being dominated by any particular components.

3. Reduction of Asymptotic Risk

3.1 Asymptotic Risk and Framework

To make comparison of the proposed JS estimators to the MLE, by closely following Hansen (2016), we calculate and compare their asymptotic risks. The asymptotic risk for an estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$ with a loss function $l(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

is defined as

$$R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) = \lim_{\zeta \to \infty} \liminf_{n \to \infty} E\left[\min\left\{nl(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0), \zeta\right\}\right].$$

Here the expectation is taken under the internal distribution and is for the scaled loss function trimmed at ζ that is negligible ($\zeta \to \infty$). The loss function we will consider is the weighted quadratic loss

$$l(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}} \boldsymbol{V}_{\boldsymbol{\beta},mle}^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{3.1}$$

so that the coefficients in $\boldsymbol{\beta}$ are scaled to have roughly the same importance.

To deal with population heterogeneity when calculating the asymptotic risk, we follow Hansen (2016) and adopt the framework of asymptotically local alternatives (e.g. Newey and McFadden (1994)). Specifically, for any fixed internal data sample size n, we consider the population heterogeneity that results in

$$E_{(\boldsymbol{X},\boldsymbol{Z})}\{\boldsymbol{U}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{\theta}_{n*})\} = n^{-1/2}\boldsymbol{\delta}. \tag{3.2}$$

in contrast to the restrictions in (2.1) without heterogeneity, where $\boldsymbol{\delta}$ is a localizing parameter and may take any arbitrary value, and $\boldsymbol{\theta}_{n*}$ is the value of $\boldsymbol{\theta}$ obtained from an infinitely large external data. The magnitude of the population heterogeneity is reflected by $\boldsymbol{\delta}$ and n. For any fixed $\boldsymbol{\delta}$ the heterogeneity disappears as n increases. However, since there is no restrictions on

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

the magnitude of $\boldsymbol{\delta}$, (3.2) can represent any realistic heterogeneity between the internal and external study populations. The $\boldsymbol{\theta}_{n*}$ in (3.2) depends on n because we consider (3.2) to be the result of a sequence of external study distributions indexed by n that are local to the internal study distribution (e.g., Chapter 7 in van der Vaart (1998)), and in this case for any fixed n the external study model yields $\boldsymbol{\theta}_{n*}$. We have $\boldsymbol{\theta}_{n*} \to \boldsymbol{\theta}_*$ as $n \to \infty$.

3.2 Asymptotic Risk Comparison

In calculating the asymptotic risk of $\hat{\boldsymbol{\beta}}_{JS}$, we consider the shrinkage weight

$$\hat{w} = \left[1 - \frac{\tau}{n(\hat{\boldsymbol{\beta}}_{mle} - \hat{\boldsymbol{\beta}}_{cmle})^{\mathrm{T}} \hat{\boldsymbol{V}}_{\boldsymbol{\beta},mle}^{-1} (\hat{\boldsymbol{\beta}}_{mle} - \hat{\boldsymbol{\beta}}_{cmle})}\right]_+$$

that has a general shrinkage parameter $\tau > 0$ compared to (2.4). This allows us to find the optimal shrinkage parameter value that minimizes an upper bound for the asymptotic risk. To present the asymptotic risk of $\hat{\boldsymbol{\beta}}_{JS}$, some notation is needed. Let $\boldsymbol{P} = (\boldsymbol{I}, \boldsymbol{0})$ be the matrix that selects the $\boldsymbol{\beta}$-block from $(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{S}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \partial \log f(Y \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\gamma}) / \partial(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is the score function for the internal study model, $\boldsymbol{\Omega} = E\{\boldsymbol{S}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\boldsymbol{S}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)^{\mathrm{T}}\}$, $\boldsymbol{\Sigma} = E\{\boldsymbol{U}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{\theta}_*)\boldsymbol{U}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{\theta}_*)^{\mathrm{T}}\}$, $\boldsymbol{G} = E\{\partial \boldsymbol{U}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{\theta}_*) / \partial(\boldsymbol{\beta}, \boldsymbol{\gamma})\}$,

$$\boldsymbol{J}_1 = \boldsymbol{V}_{\boldsymbol{\beta},mle}^{-1/2} \boldsymbol{P} \boldsymbol{\Omega}^{-1} \boldsymbol{G}^{\mathrm{T}} (\boldsymbol{G} \boldsymbol{\Omega}^{-1} \boldsymbol{G}^{\mathrm{T}})^{-1} \boldsymbol{G} \boldsymbol{\Omega}^{-1} \boldsymbol{P}^{\mathrm{T}} \boldsymbol{V}_{\boldsymbol{\beta},mle}^{-1/2},$$

$$\boldsymbol{J}_2 = \boldsymbol{V}_{\boldsymbol{\beta},mle}^{-1/2} \boldsymbol{P} \boldsymbol{\Omega}^{-1} \boldsymbol{G}^{\mathrm{T}} (\boldsymbol{\Sigma} + \boldsymbol{G} \boldsymbol{\Omega}^{-1} \boldsymbol{G}^{\mathrm{T}})^{-1} \boldsymbol{G} \boldsymbol{\Omega}^{-1} \boldsymbol{P}^{\mathrm{T}} \boldsymbol{V}_{\boldsymbol{\beta},mle}^{-1/2}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

For $\boldsymbol{J} = \boldsymbol{J}_1$ or $\boldsymbol{J}_2$, let $\mathrm{tr}(\boldsymbol{J})$ denote the trace of $\boldsymbol{J}$ and $\| \boldsymbol{J} \|$ the largest eigenvalue of $\boldsymbol{J}$, and define $d = \mathrm{tr}(\boldsymbol{J})/ \| \boldsymbol{J} \|$. Under certain regularity conditions, we have the following result, the proof of which is provided in the Supplementary Materials.

Theorem 1. (i). The asymptotic risk for $\hat{\boldsymbol{\beta}}_{mle}$ is

$$R(\hat{\boldsymbol{\beta}}_{mle}, \boldsymbol{\beta}_0) = E(\boldsymbol{\Delta}_{\boldsymbol{S}}^{\mathrm{T}} \boldsymbol{\Omega}^{-1} \boldsymbol{P}^{\mathrm{T}} \boldsymbol{V}_{\boldsymbol{\beta},mle}^{-1} \boldsymbol{P} \boldsymbol{\Omega}^{-1} \boldsymbol{\Delta}_{\boldsymbol{S}}),$$

where $\boldsymbol{\Delta}_{\boldsymbol{S}}$ is a random variable such that $\boldsymbol{\Delta}_{\boldsymbol{S}} \sim N(\boldsymbol{0}, \boldsymbol{\Omega})$. (ii) If $d > 2$, then for any τ such that $0 < \tau \leq 2\,[\mathrm{tr}(\boldsymbol{J}) - 2\|\boldsymbol{J}\|]$ and for any $\boldsymbol{\delta}$, the asymptotic risk for $\hat{\boldsymbol{\beta}}_{JS}$ satisfies

$$R(\hat{\boldsymbol{\beta}}_{JS}, \boldsymbol{\beta}_0) < R(\hat{\boldsymbol{\beta}}_{mle}, \boldsymbol{\beta}_0) - \tau \frac{2\,[\mathrm{tr}(\boldsymbol{J}) - 2 \| \boldsymbol{J} \|] - \tau}{E\{(\boldsymbol{\Delta}_* + \boldsymbol{\delta})^{\mathrm{T}} \boldsymbol{B} (\boldsymbol{\Delta}_* + \boldsymbol{\delta})\}}, \tag{3.3}$$

where $\boldsymbol{J} = \boldsymbol{J}_1$ for $\hat{\boldsymbol{\beta}}_{JS}$ based on $\hat{\boldsymbol{\beta}}_{cmle-sp}$ and $\boldsymbol{J} = \boldsymbol{J}_2$ for $\hat{\boldsymbol{\beta}}_{JS}$ based on $\hat{\boldsymbol{\beta}}_{cmle-el}$. Here $\boldsymbol{\Delta}_* = (\boldsymbol{G}\boldsymbol{\Omega}^{-1}, \boldsymbol{I})\boldsymbol{\Delta}$, $\boldsymbol{\Delta} \equiv (\boldsymbol{\Delta}_{\boldsymbol{S}}^{\mathrm{T}}, \boldsymbol{\Delta}_{\boldsymbol{U}}^{\mathrm{T}})^{\mathrm{T}}$ where $\boldsymbol{\Delta}_{\boldsymbol{U}}$ is a random variable such that $\boldsymbol{\Delta} \sim N(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\Delta}})$ with $\boldsymbol{V}_{\boldsymbol{\Delta}} = \mathrm{diag}(\boldsymbol{\Omega}, \boldsymbol{\Sigma})$, and $\boldsymbol{B} = \boldsymbol{L}^{\mathrm{T}} \boldsymbol{P}^{\mathrm{T}} \boldsymbol{V}_{\boldsymbol{\beta},mle}^{-1} \boldsymbol{P} \boldsymbol{L}$ with $\boldsymbol{L} = \boldsymbol{\Omega}^{-1} \boldsymbol{G}^{\mathrm{T}} (\boldsymbol{\Sigma} + \boldsymbol{G}\boldsymbol{\Omega}^{-1}\boldsymbol{G}^{\mathrm{T}})^{-1}$.

From Theorem 1, with $d > 2$ and for all shrinkage parameter values between 0 and $2\,[\mathrm{tr}(\boldsymbol{J}) - 2 \| \boldsymbol{J} \|]$, the asymptotic risk of $\hat{\boldsymbol{\beta}}_{JS}$ is strictly smaller than that of the MLE, and this result holds for any fixed but arbitrarily large $\boldsymbol{\delta}$, which makes the dominance of $\hat{\boldsymbol{\beta}}_{JS}$ over the MLE meaningful in practice. The d plays the same role as the shrinkage dimension for the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

original JS estimator, and $d > 2$ is needed for $\hat{\boldsymbol{\beta}}_{JS}$ to achieve a uniform reduction in asymptotic risk compared to the MLE regardless of the degree of population heterogeneity. A necessary condition for $d > 2$ is that both $\boldsymbol{V}_{\boldsymbol{\beta},mle}$ and $\boldsymbol{G}$ have ranks exceeding 2, which implies that both the dimension of $\boldsymbol{\beta}$ and the dimension of $\boldsymbol{\theta}$ need to exceed 2.

Theorem 1 depends on the invertibility of $\boldsymbol{\Omega}$, which is typically assumed for general maximum likelihood theory (e.g. Newey and McFadden (1994), Theorem 3.3). When $\boldsymbol{\Omega}$ is invertible, the invertibility of $\boldsymbol{G}\boldsymbol{\Omega}^{-1}\boldsymbol{G}^T$ and $\boldsymbol{\Sigma} + \boldsymbol{G}\boldsymbol{\Omega}^{-1}\boldsymbol{G}^T$, which is needed by Theorem 1, is anticipated in general, because (i) $\boldsymbol{G}$ has fewer rows than columns since $\boldsymbol{\theta}$ has fewer dimension than $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, (ii) $\boldsymbol{G}$ should have full row rank when components of $\boldsymbol{Q}(Y, \boldsymbol{X}^*; \boldsymbol{\theta})$ are not linearly dependent, and (iii) $\boldsymbol{\Sigma}$ is positive semi-definite.

The upper bound on the asymptotic risk of $\hat{\boldsymbol{\beta}}_{JS}$ in (3.3) is a quadratic function of τ and is minimized at $\tau_* = \mathrm{tr}(\boldsymbol{J}) - 2 \parallel \boldsymbol{J} \parallel$, which gives the optimal value for the shrinkage parameter. Therefore, $\hat{\tau}$ in (2.4) is taken to be

$$\hat{\tau} = \mathrm{tr}(\hat{\boldsymbol{J}}) - 2 \parallel \hat{\boldsymbol{J}} \parallel, \tag{3.4}$$

where $\hat{\boldsymbol{J}}$ is a consistent estimator of $\boldsymbol{J}$. Specifically, let $\hat{E}(\cdot)$ denote the sample average $n^{-1}\sum_{i=1}^{n}(\cdot)$, we have

$$\hat{\boldsymbol{J}}_1 = \hat{\boldsymbol{V}}_{\boldsymbol{\beta},mle}^{-1/2}\boldsymbol{P}\hat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{G}}^{\mathrm{T}}(\hat{\boldsymbol{G}}\hat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{G}}^{\mathrm{T}})^{-1}\hat{\boldsymbol{G}}\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{P}^{\mathrm{T}}\hat{\boldsymbol{V}}_{\boldsymbol{\beta},mle}^{-1/2},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

$$\hat{\boldsymbol{J}}_2 = \hat{\boldsymbol{V}}_{\boldsymbol{\beta},mle}^{-1/2}\boldsymbol{P}\hat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{G}}^{\mathrm{T}}(\hat{\boldsymbol{\Sigma}}+\hat{\boldsymbol{G}}\hat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{G}}^{\mathrm{T}})^{-1}\hat{\boldsymbol{G}}\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{P}^{\mathrm{T}}\hat{\boldsymbol{V}}_{\boldsymbol{\beta},mle}^{-1/2},$$

where $\hat{\boldsymbol{\Omega}} = \hat{E}\{\boldsymbol{S}_i(\hat{\boldsymbol{\beta}}_{mle},\hat{\boldsymbol{\gamma}}_{mle})\boldsymbol{S}_i(\hat{\boldsymbol{\beta}}_{mle},\hat{\boldsymbol{\gamma}}_{mle})^{\mathrm{T}}\}$, $\hat{\boldsymbol{\Sigma}} = \hat{E}\{\boldsymbol{U}_i(\hat{\boldsymbol{\beta}}_{mle},\hat{\boldsymbol{\gamma}}_{mle},\boldsymbol{\theta}_*)\boldsymbol{U}_i(\hat{\boldsymbol{\beta}}_{mle},\hat{\boldsymbol{\gamma}}_{mle},\boldsymbol{\theta}_*)^{\mathrm{T}}\}$, $\hat{\boldsymbol{G}} = \hat{E}\{\partial\boldsymbol{U}_i(\hat{\boldsymbol{\beta}}_{mle},\hat{\boldsymbol{\gamma}}_{mle},\boldsymbol{\theta}_*)/\partial(\boldsymbol{\beta},\boldsymbol{\gamma})\}$, and $\hat{\boldsymbol{V}}_{\boldsymbol{\beta},mle}$ is the upper left block matrix of $\hat{\boldsymbol{\Omega}}^{-1}$ corresponding to $\boldsymbol{\beta}$.

3.3 Other Loss Functions

The focus so far has been on the loss (3.1). One may consider a more general weighted quadratic loss $l(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}_0) = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^{\mathrm{T}}\boldsymbol{W}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$ with any constant weighting matrix $\boldsymbol{W}$. One particular example is when the interest is in the prediction accuracy. In this case one may take the loss to be the following integrated squared error as the measure of prediction accuracy

$$\begin{aligned} l(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}_0) &= \int\{(1,\boldsymbol{X}^{\mathrm{T}},\boldsymbol{Z}^{\mathrm{T}})\hat{\boldsymbol{\beta}}-(1,\boldsymbol{X}^{\mathrm{T}},\boldsymbol{Z}^{\mathrm{T}})\boldsymbol{\beta}_0\}^2 f(\boldsymbol{X},\boldsymbol{Z})d\boldsymbol{X}\,d\boldsymbol{Z} \\ &= (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^{\mathrm{T}}E\left\{\begin{pmatrix}1\\ \boldsymbol{X}\\ \boldsymbol{Z}\end{pmatrix}(1,\boldsymbol{X}^{\mathrm{T}},\boldsymbol{Z}^{\mathrm{T}})\right\}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0). \qquad (3.5) \end{aligned}$$

Another example is the unweighted quadratic loss $l(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}_0) = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^{\mathrm{T}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$ where $\boldsymbol{W}$ is the identity matrix. It turns out that results similar to Theorem 1 on the dominance of JS estimators over the MLE can still be established by following the same proof, with the weight $\hat{w}$ in (2.4) modified

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

as

$$\hat{w} = \left[1 - \frac{\hat{\tau}}{n(\hat{\boldsymbol{\beta}}_{mle} - \hat{\boldsymbol{\beta}}_{cmle})^{\mathrm{T}} \hat{\boldsymbol{W}} (\hat{\boldsymbol{\beta}}_{mle} - \hat{\boldsymbol{\beta}}_{cmle})}\right]_+,$$

where $\hat{\boldsymbol{W}}$ is a consistent estimator of $\boldsymbol{W}$, and $\hat{\tau}$ is still given by (3.4) but with all $\hat{\boldsymbol{V}}_{\boldsymbol{\beta},mle}^{-1/2}$ in $\hat{\boldsymbol{J}}$ replaced by $\hat{\boldsymbol{W}}^{1/2}$.

4. Simulation Studies

4.1 Simulation Setup

We evaluate the performance of our proposed estimators by simulation studies. The internal study has covariates X_1, X_2, X_3, X_4 and Z, where $X_1 \sim Exp(1)$, (X_2, X_3, X_4) follows a multivariate normal distribution with mean (μ, μ, μ), unit variances and correlation 0.3, and $Z \sim N(\alpha \log(X_1), 1)$. The binary outcome Y is generated from the logistic regression model $\text{logit}\{P(Y = 1|\boldsymbol{X}, Z)\} = \beta_c + \beta_X(X_1 + X_2 + X_3 + X_4) + \beta_Z Z + \beta_{XZ} X_3 Z$. We set $\mu = 0$, $\alpha = 0.2$, $\beta_c = 0.1$, $\beta_X = 0.5$, $\beta_Z = 0.2$ and $\beta_{XZ} = 0.2$ for the internal study.

The data in the external study is generated from the same logistic regression model. To introduce heterogeneity between internal and external studies, we vary the value of each of μ, α, β_c, β_X, β_Z and β_{XZ} while holding the rest at the internal study values when generating the external study data. The external study collects Y, X_1, X_2 and $\tilde{X}_3$ and fits a logistic re-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

gression model $\text{logit}\{P(Y = 1|\boldsymbol{X}^*)\} = \theta_c + \theta_1 X_1 + \theta_2 X_2 + \theta_3 \tilde{X}_3$, where $\tilde{X}_3 = I(X_3 > 0.7\bar{X}_3)$ with $\bar{X}_3$ being the sample mean.

For each estimator $\hat{\boldsymbol{\beta}}$ in our comparison, we evaluate the performance by calculating the risk under the weighted quadratic loss $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}} \boldsymbol{V}_{\hat{\boldsymbol{\beta}},mle}^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ based on 1000 replications. Because the theoretical framework adopted in this paper treats the external information as coming from a fixed external data set, we calculate the risk by replicating the internal data while using the same fixed external data. We consider four combinations of sample sizes of 300 and 1000 for the two studies. Because of the similar scale of internal and external sample sizes, the calculated risk value depends on the particular external data set used to generate the external information. In other words, the value of the risk varies if a new external data set is used. While this does not qualitatively affect the risk reduction of our proposed estimators in comparison to the MLE, it does introduce variation when demonstrating the quantitative impact of study heterogeneity on the magnitude of risk reduction if at each level of heterogeneity a completely new external data set is generated.

To mitigate this variation, we "recycle" the external study data set when varying the level of study heterogeneity. We first generate a set of external covariates using parameter values that are all the same as the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

internal study, under the external sample size. Then, when varying each of β_c, β_X, β_Z and β_{XZ} for the external study, all covariates are held fixed at the above generated values, while when varying each of μ and α the correct constant shift for the means of the corresponding covariates is added to the above generated values. Conditional on the covariates at each level of heterogeneity, the outcome Y is set to be 1 for subject i if $P(Y = 1|\boldsymbol{X}_i, Z_i) \geq U_i$ and 0 otherwise, where $U_i \sim \text{Uniform}(0, 1)$. The same set of U_i's are used for all 1,000 replications and for all levels of heterogeneity. Generating external data as described ensures that the difference in the external study data across different levels of heterogeneity is to largely the result of the magnitude of heterogeneity rather than the randomness from regenerating the covariates and the outcome. As a result, all the curves in the figures we will present are relatively smooth to facilitate an illustration of the quantitative impact of study heterogeneity on the magnitude of risk reduction.

4.2 Observations from the Simulation Results

Figures 1-4 contain our simulation results corresponding to the four sample size combinations. In addition to the MLE, the two CML estimators, and the two proposed JS estimators, we also include two empirical Bayes esti-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

mators constructed based on the two CML estimators (Estes et al. 2018). To facilitate the comparison, we plot the ratio of the risk of each estimator relative to the risk of the MLE, so that a risk ratio below one implies the estimator reduces the risk after integrating external information. For each plot, the magnitude of study heterogeneity starts from zero and increases as each parameter value increases from left to right on the horizontal axis. Overall, the two CML estimators perform very similarly, and the same is true for the respective two JS estimators and two empirical Bayes estimators constructed based on the CML estimators.

For combinations (300,300), (1000,300) and (1000,1000), where the two numbers inside the parentheses represent the internal and external sample sizes, respectively, the risk ratio plots for CML estimators have a jagged nature, which to a much lesser degree is also presents for the JS and EB estimators. This jagged nature is due to the non-negligible randomness when generating the external study data at varying values of external model parameters, despite that we have already designed the data generating procedure to reduce such randomness by "recycling" the external study data set. Furthermore, although the risk ratios for the CML estimators tend to be below one when study heterogeneity does not exist or is mild, this is not guaranteed as can be seen from the combination (1000,1000) for which the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

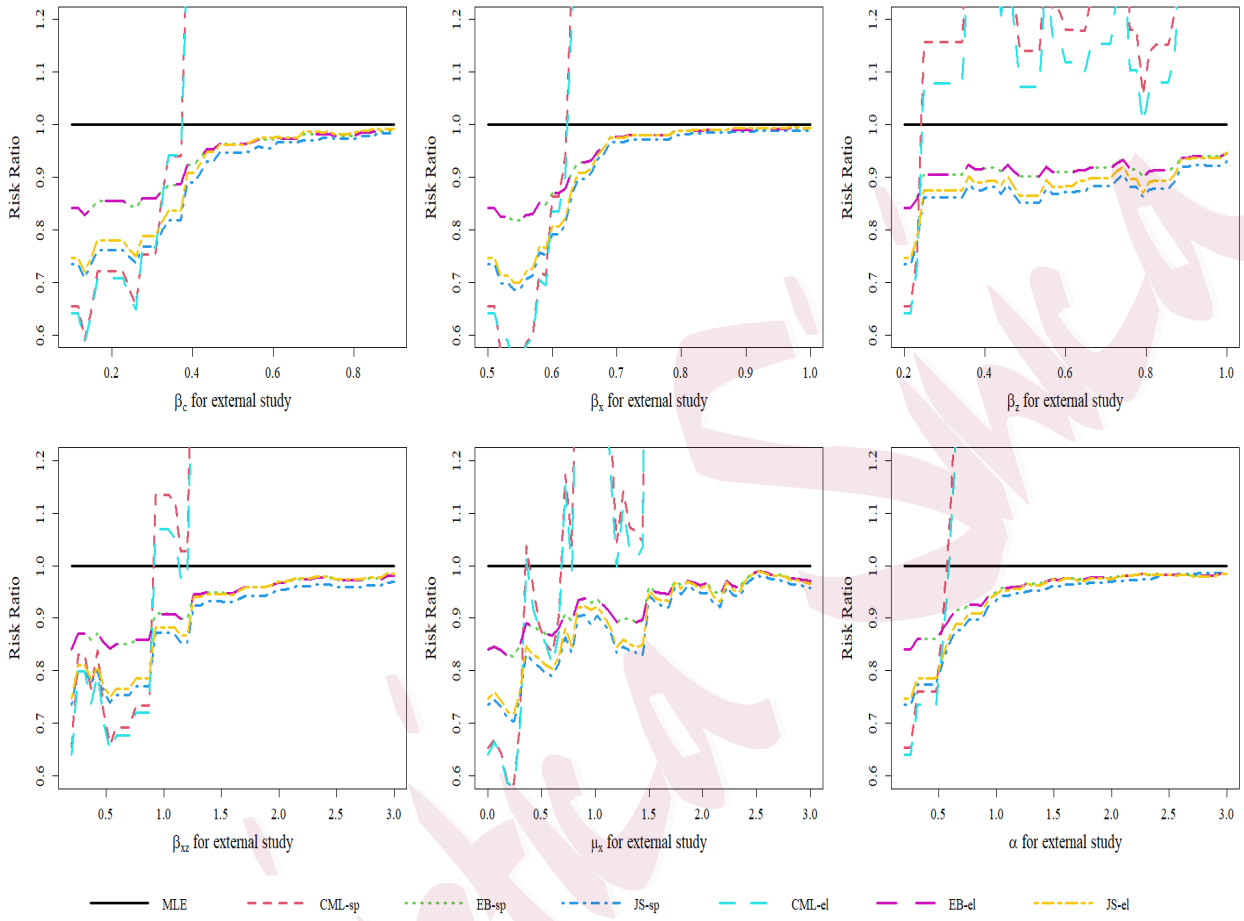

Figure 1: Comparison of different estimators relative to the risk of the MLE. The internal and external sample sizes are 300 and 300, respectively. MLE: maximum likelihood estimator; CML: constrained maximum likelihood; JS: James-Stein; EB: empirical Bayes; -sp: based on CML-sp as in (2.2); -el: based on CML-el as in (2.3)

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

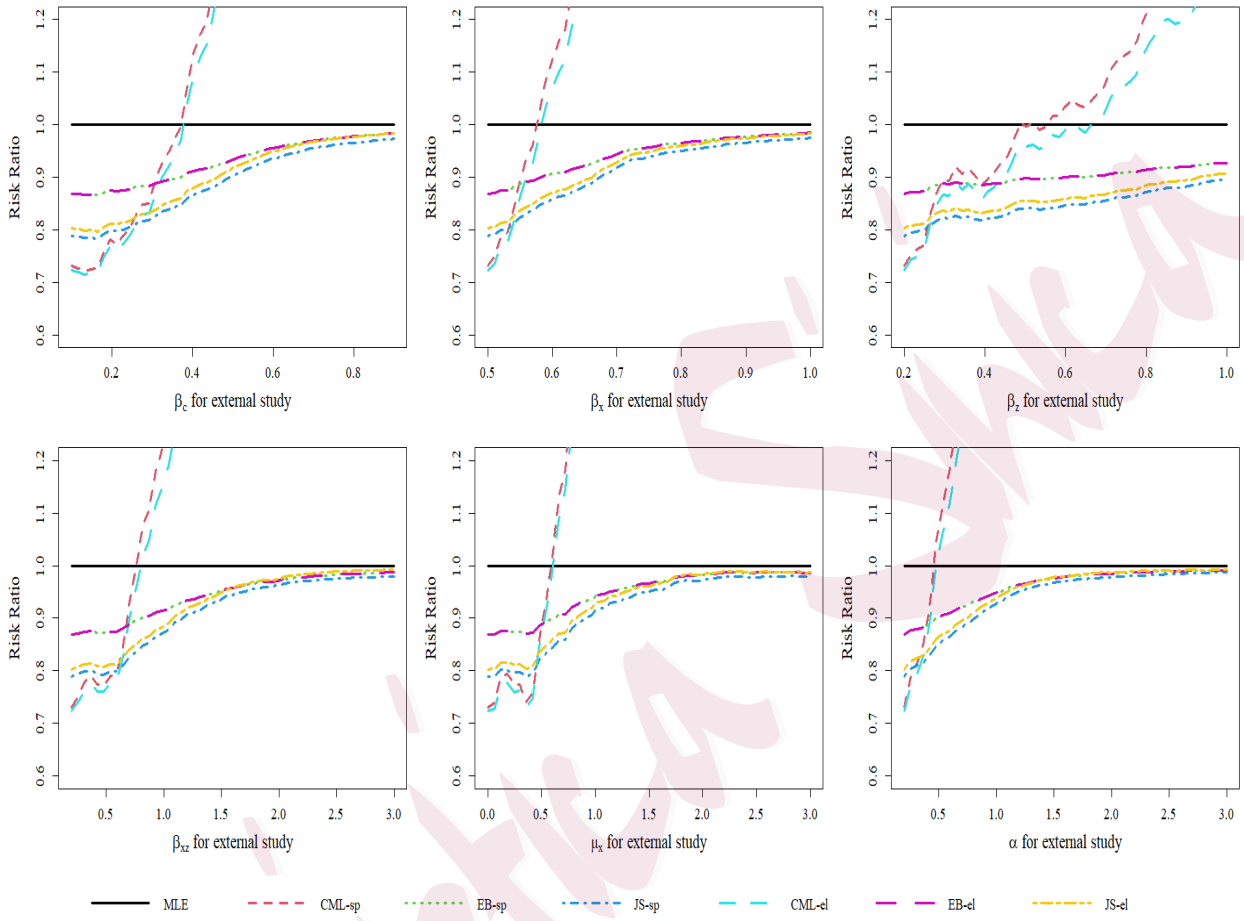

Figure 2: Comparison of different estimators relative to the risk of the MLE. The internal and external sample sizes are 300 and 1000, respectively. MLE: maximum likelihood estimator; CML: constrained maximum likelihood; JS: James-Stein; EB: empirical Bayes; -sp: based on CML-sp as in (2.2); -el: based on CML-el as in (2.3)

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

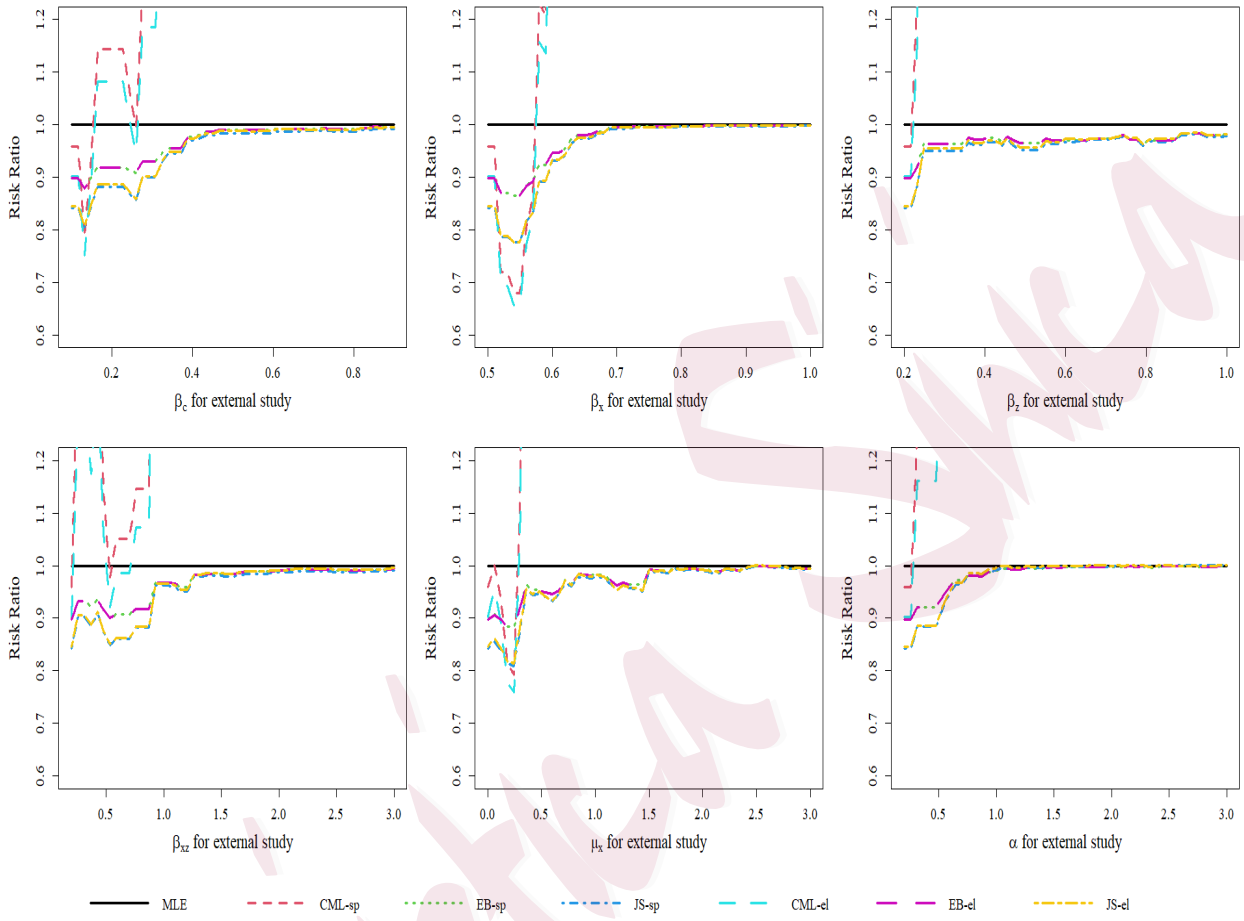

Figure 3: Comparison of different estimators relative to the risk of the MLE. The internal and external sample sizes are 1000 and 300, respectively. MLE: maximum likelihood estimator; CML: constrained maximum likelihood; JS: James-Stein; EB: empirical Bayes; -sp: based on CML-sp as in (2.2); -el: based on CML-el as in (2.3)

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

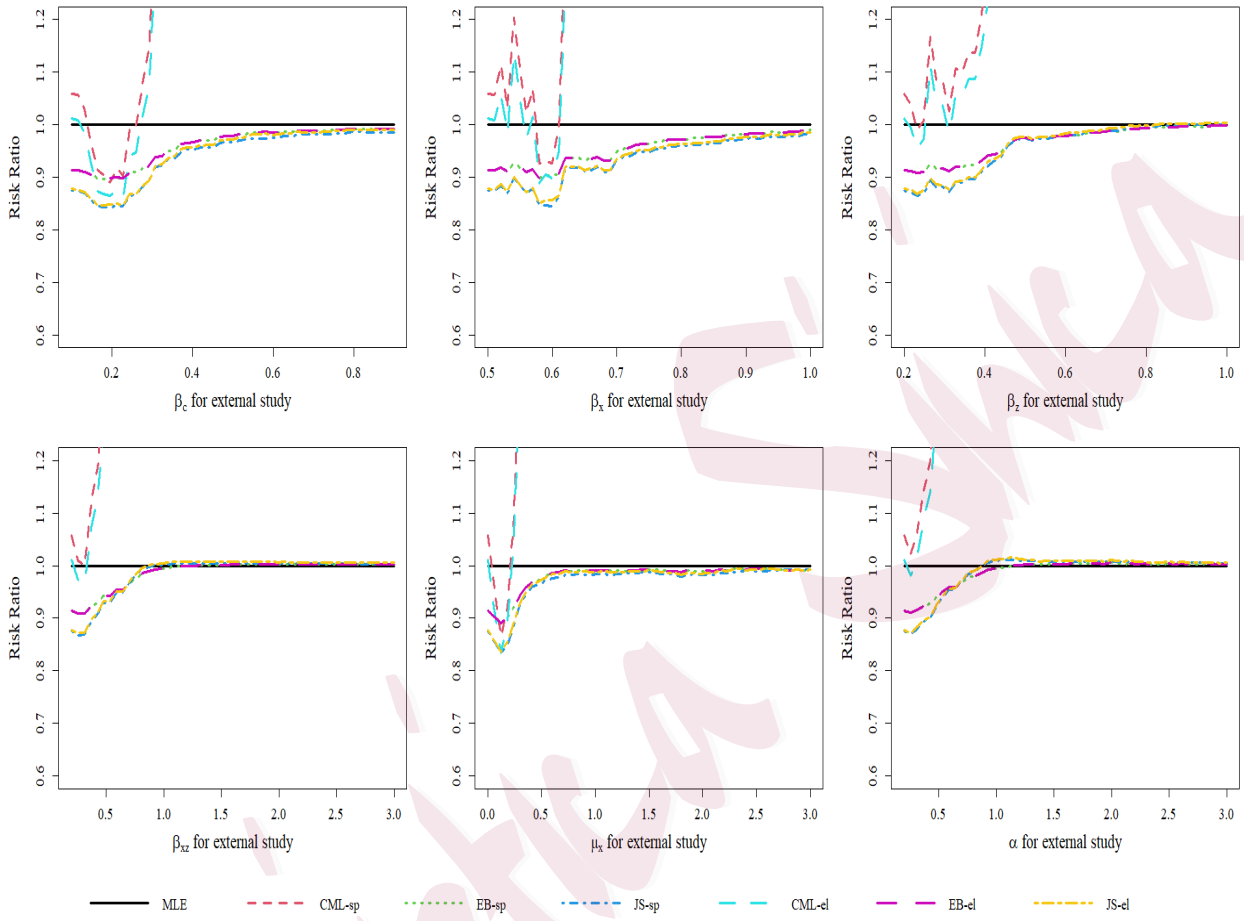

Figure 4: Comparison of different estimators relative to the risk of the MLE. The internal and external sample sizes are 1000 and 1000, respectively. MLE: maximum likelihood estimator; CML: constrained maximum likelihood; JS: James-Stein; EB: empirical Bayes; -sp: based on CML-sp as in (2.2); -el: based on CML-el as in (2.3)

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

risk ratios exceed one at the left end. This observation again is because of the non-negligible randomness in the external data that could make the parameter estimates produced by the external study inconsistent with the internal data even when there is no heterogeneity at the population level. When the external sample size is substantially larger than the internal size such that its relative randomness becomes greatly reduced, the CML risk ratio plots become smoother, and further when there is no heterogeneity the CML risk ratios become below one, as can be seen from the combination (300,1000) and a new combination (1000,5000) in Figure 5 which we included to confirm this point.

Another major drawback of the CML estimators is that their risk ratios can be much larger than one as the level of heterogeneity increases. This observation is not due to randomness in the external data and remains true regardless of the external sample size. In other words, the CML estimators can have a much worse performance after integrating the external information, compared to the MLE, in the presence of severe heterogeneity.

In contrast, both the JS and EB estimators have a substantial improvement over the CML estimators in all aforementioned aspects. In particular, their risk ratios remain below one regardless of the degree of study heterogeneity, except for some occasional numerical values that exceed one by a

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

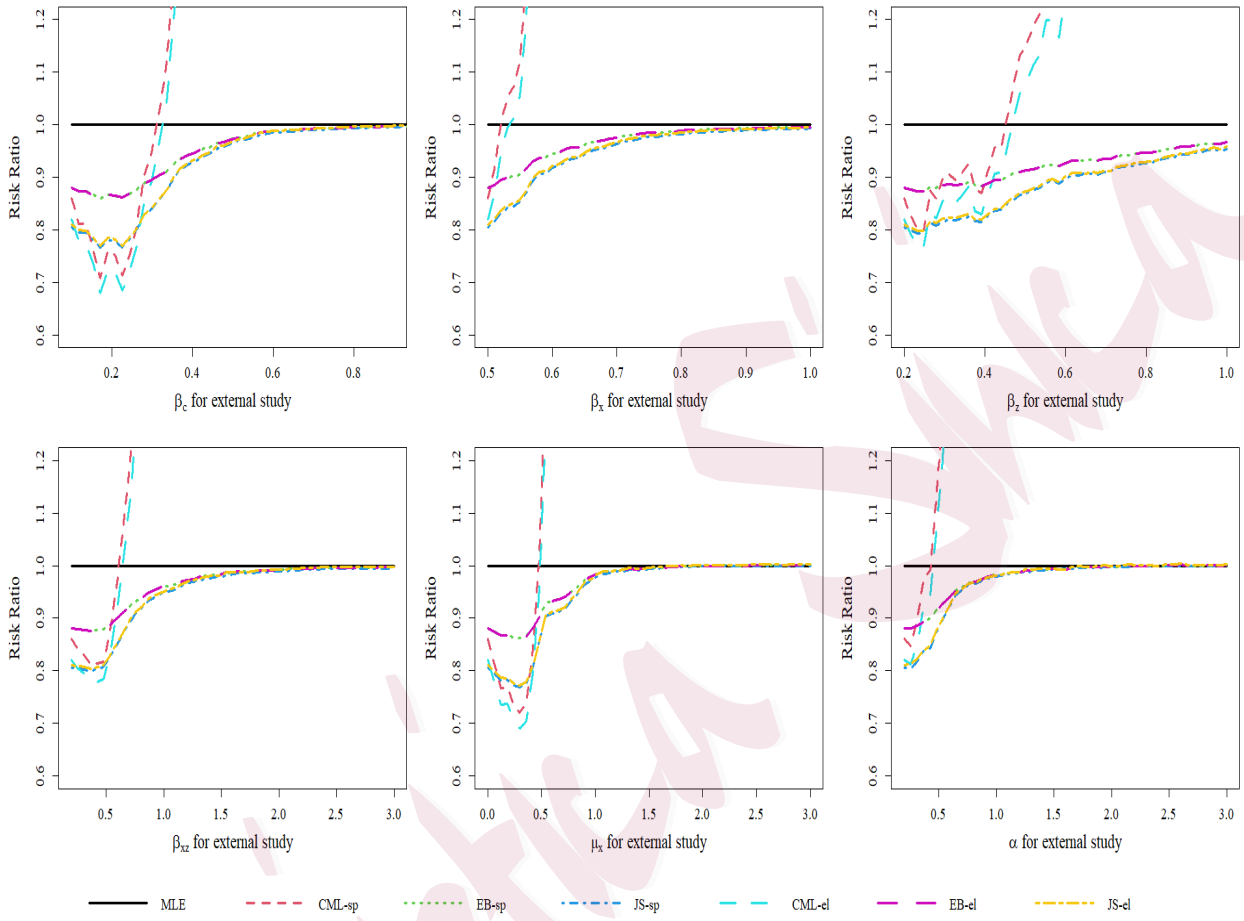

Figure 5: Comparison of different estimators relative to the risk of the MLE. The internal and external sample sizes are 1000 and 5000, respectively. MLE: maximum likelihood estimator; CML: constrained maximum likelihood; JS: James-Stein; EB: empirical Bayes; -sp: based on CML-sp as in (2.2); -el: based on CML-el as in (2.3)

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

tiny bit in the presence of severe heterogeneity. When comparing the JS and EB estimators, the risk reduction for the JS estimators by integrating external information is more substantial. In addition, unlike the theory established in this paper for the JS estimators, there has been no theoretical guarantee established for the EB estimators in the literature.

In summary, the proposed JS estimators have the best overall numerical performance among all the estimators under our comparison.

4.3 Impact of Signal Strength

We also examine the dependence of risk ratios on the signal strength. To avoid possible complications from interaction effects, the data for both the internal and external studies are generated from $\text{logit}\{P(Y=1)\} = \beta_c + \beta_X(X_1+X_2+X_3+X_4) + \beta_Z Z$ under the same parameter values. We vary β_X from 0 to 1 as a change of signal strength, while fixing $\beta_Z = 0.2$ and adjusting β_c accordingly so that the marginal event rate $\Pr(Y=1)$ remains approximately 0.6. We then fit logistic regression with intercept and main effects for X's and Z for the internal study while the external study excludes the main effect for Z. We take the sample size 300 for the internal study and 1000 for the external study and create Figure 6. It is seen that the risk ratios of both JS estimators are not dramatically affected by

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

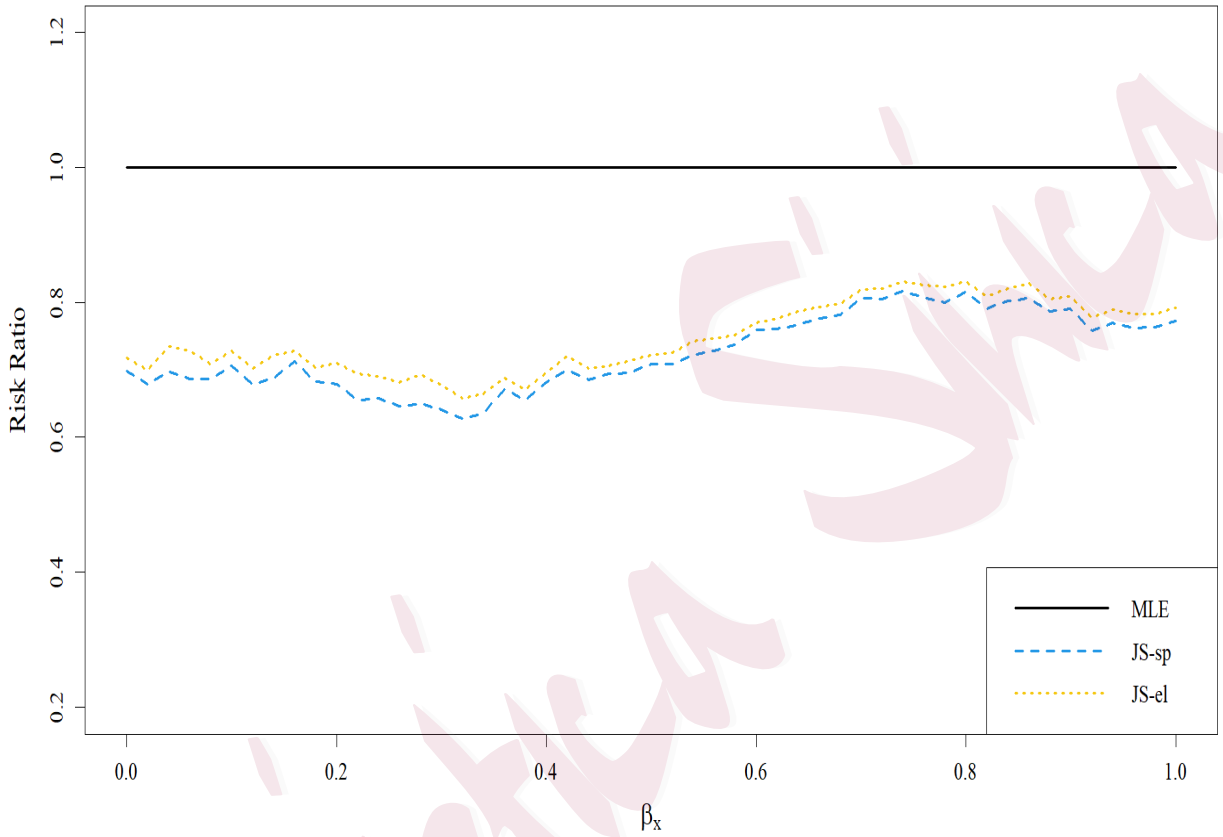

Figure 6: Impact of signal strength on the improvement of the JS estimators over the maximum likelihood estimator. The internal and external sample sizes are 300 and 1000, respectively. MLE: maximum likelihood estimator; JS: James-Stein; -sp: based on CML-sp as in (2.2); -el: based on CML-el as in (2.3)

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

the signal strength, indicating that the improvement over the MLE without using external information is robust regardless of the signal strength.

5. Data Application

We apply the proposed James-Stein method to study the risk of developing high-grade prostate cancer (Gleason grade ≥ 7) based on certain risk factors. Conventional risk factors such as age, race, prostate-specific antigen (PSA), digital rectal examination (DRE) result, and prior biopsy status have been extensively analyzed in previous studies, most notably in the Prostate Cancer Prevention Trial (PCPT). Thompson et al. (2006) reported a logistic regression model based on 5519 men in the PCPT placebo arm, and the published coefficient estimates are used as external summary information in our data application.

In recent years, research has shown that some molecular biomarkers can serve as additional tools for early risk stratification. Two such markers, TMPRSS2:ERG (T2:ERG) and prostate cancer antigen 3 (PCA3), have shown to be very helpful in improving disease detection performance beyond standard clinical variables (Tomlins et al., 2016). Motivated by this biological evidence, our internal study aims to include both traditional risk factors and these new biomarkers within a risk model based on logistic

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

regression.

The data we use is part of the sample from Tomlins et al. (2016), which consists of 1218 men undergoing diagnostic prostate biopsy at seven U.S. clinics. The covariates include PSA level (ng/ml), age, a binary indicator of an abnormal DRE result, a binary indicator of negative previous biopsies, a binary indicator of being African American, PCA3 score, and a binary indicator dichotomized at the sample median of the T2:ERG score (Cheng et al. 2019). The biomarker measurements on PCA3 score and T2:ERG are not present in the PCPT cohort.

For the analysis, we randomly split the available data into an internal set (1/3) for estimation and an evaluation set (2/3) for performance assessment. We compute the MLE based on logistic regression, two CML estimators, and the James-Stein estimators based on the two CML estimators. For the JS estimators, we consider two loss functions (3.1) and (3.5). Performance is compared on the evaluation set using Brier score. The analysis results are summarized in Table 1. It is seen that the two CML estimates are similar to each other, but they are different from the MLE. Both have a worse Brier score compared to the MLE on the evaluation data. This worse performance after using the external information is because of the study population heterogeneity between the internal set and

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

Table 1: Analysis results for the prostate cancer data

	$\hat{\boldsymbol{\beta}}_{mle}$	$\hat{\boldsymbol{\beta}}_{cml-sp}$	$\hat{\boldsymbol{\beta}}_{cml-el}$	$\hat{\boldsymbol{\beta}}^{wq}_{js-sp}$	$\hat{\boldsymbol{\beta}}^{pr}_{js-sp}$	$\hat{\boldsymbol{\beta}}^{wq}_{js-el}$	$\hat{\boldsymbol{\beta}}^{pr}_{js-el}$
Intercept	-8.21	-6.85	-7.29	-8.12	-8.16	-8.17	-8.20
PSA	0.60	1.32	1.25	0.65	0.63	0.63	0.61
Age	0.04	0.01	0.00	0.04	0.04	0.04	0.04
DRE	0.07	1.05	0.89	0.13	0.11	0.11	0.08
Biopsy	-0.63	-0.21	-0.21	-0.60	-0.62	-0.61	-0.62
Race	-0.19	0.88	0.77	0.04	0.02	0.02	0.00
PCA3	0.47	0.37	0.49	0.46	0.47	0.47	0.47
T2:ERG	0.49	0.42	0.56	0.49	0.49	0.50	0.49
Brier score	0.1309	0.1475	0.1442	0.1305	0.1306	0.1307	0.1308
$\hat{w}$ on $\hat{\boldsymbol{\beta}}_{mle}$				0.93	0.95	0.96	0.98

js-sp: JS estimator based on $\hat{\boldsymbol{\beta}}_{cml-sp}$; js-el: JS estimator based on $\hat{\boldsymbol{\beta}}_{cml-el}$; wq: JS estimator based on weighted quadratic loss (3.1); pr: JS estimator based on the prediction loss (3.5).

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

the PCPT set (e.g. Zhai and Han (2022)). In contrast, the JS estimates are much more similar to the MLE, with a very slight improvement in the Brier score. The weight $\hat{w}$ on the MLE when constructing all JS estimates indicates that the external information was considerably downweighted. This substantial downweighting represents a safeguard of the proposed method on integrating external information in the presence of population heterogeneity in real-world applications.

6. Discussion

We treated $\boldsymbol{\theta}_*$ as if it were derived based on an infinite external sample size. In practice, $\boldsymbol{\theta}_*$ is derived under a finite sample size with an associated uncertainty. We did not assume the availability of quantities associated with this uncertainty, such as standard errors. Instead, we treated $\boldsymbol{\theta}_*$ as a fixed and deterministic value. This did not affect our derivations and results because mathematically $\boldsymbol{\theta}_*$ could be imagined to coincide with the fixed and deterministic value that was derived under an infinite external sample size from another appropriate data distribution. Since our goal is to ensure an improvement by integrating external information regardless of the heterogeneity of the study population, our treatment of $\boldsymbol{\theta}_*$ still serves the purpose. It would be interesting to develop ways to incorporate the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

uncertainty associated with $\boldsymbol{\theta}_*$, when available, and to study if further improvement could be achieved after incorporating this uncertainty.

We considered information from one external study. When there are multiple external studies, one could compute a JS estimator using information from each external study and then construct the final estimator as the weighted average. The weights could follow those developed in George (1986), which have been used in the linear regression case (Ki, 1992; Han et al., 2024). However, it is not clear whether applying those weights in our setting would still ensure improvement over the MLE in terms of asymptotic risk. This would be an interesting future research topic.

We focused on conventional regression settings with a fixed number of covariates. The CML method we made use of was also investigated under the conventional settings instead of high-dimensional cases. It is anticipated that, when the number of constraints becomes large, the CML method becomes computationally intensive and can be numerically unstable, which will affect our proposed estimator. There may be possible ways to mitigate such impact, such as selecting the most informative constraints through principal component analysis. But there have not been formal investigations of this in existing literature. We think the high-dimensional settings deserve future research efforts.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

In addition, we considered settings where estimating functions are differentiable and loss functions are smooth (quadratic). Extensions of the method to settings where functions are nondifferentiable or loss functions are nonsmooth, such as quantile regression or absolute loss, may be possible and deserve future investigations.

Supplementary Material

The online Supplementary material contains some useful facts about the two constrained maximum likelihood estimators defined in (2.2) and (2.3) and the proof of Theorem 1.

Acknowledgments

We would like to thank the Editor, Associate Editor, and two referees for their helpful comments that improved the quality of this work. This research was partially supported by National Institutes of Health grants CA-129102 and CA-46592 to Taylor.

References

Baranchik, A. (1964). Multiple regression and estimation of the mean of the multivariate normal distribution. pp. Technical Report No. 51, Department of Statistics, Stanford University.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

https://purl.stanford.edu/hq883mw7116.

Chatterjee, N., Y.-H. Chen, P. Maas, and R. J. Carroll (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.

Chen, Z., J. Ning, Y. Shen, and J. Qin (2021). Combining primary cohort data with external aggregate information without assuming comparability. *Biometrics 77*, 1024–1036.

Cheng, W., J. M. G. Taylor, T. Gu, S. A. Tomlins, and B. Mukherjee (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.

Cheng, W., J. M. G. Taylor, P. S. Vokonas, S. K. Park, and B. Mukherjee (2018). Improving estimation and prediction in linear regression incorporating external information from an established reduced model. *Statistics in medicine 37*, 1515–1530.

Estes, J. P., B. Mukherjee, and J. M. G. Taylor (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.

George, E. I. (1986). Minimax multiple shrinkage estimation. *The Annals of Statistics 14*, 188–205.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

REFERENCES

Guo, Z., X. Li, L. Han, and T. Cai (2025). Robust inference for federated meta-learning. *Journal of the American Statistical Association 120*, 1695–1710.

Han, L., J. Hou, K. Cho, R. Duan, and T. Cai (2025). Federated adaptive causal estimation (face) of target treatment effects. *Journal of the American Statistical Association 120*, 1503–1516.

Han, P. and J. F. Lawless (2016). Discussion of "constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources". *Journal of the American Statistical Association 111*, 118–121.

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

Han, P., H. Li, S. Park, B. Mukherjee, and J. Taylor (2024). Improving prediction of linear regression models by integrating external information from heterogeneous populations: James–stein estimators. *Biometrics 80*, 10.1093/biomtc/ujae072.

Hansen, B. (2015). Shrinkage efficiency bounds. *Econometric Theory 31*, 860–879.

Hansen, B. E. (2016). Efficient shrinkage in parametric models. *Journal of Econometrics 190*, 115–132.

Hector, E. C. and R. Martin (2024). Turning the information-sharing dial: efficient inference from different data sources. *Electronic Journal of Statistics 18*, 2974–3020.

Huang, C.-Y., J. Qin, and H.-T. Tsai (2016). Efficient estimation of the cox model with auxiliary

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

subgroup survival information. *Journal of the American Statistical Association 111*, 787–799.

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

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings fo the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 361–379. University of California Press.

Ki, Y.-C. F. (1992). Multiple shrinkage estimators in multiple linear regression. *Communications in Statistics - Theory and Methods 21*, 111–136.

Kundu, P., R. Tang, and N. Chatterjee (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika 106*, 567–585.

Li, S., T. Cai, and H. Li (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*, 149–173.

Newey, W. K. and D. L. McFadden (1994). *Large Sample Estimation and Hypothesis Testing*. Handbook of Econometrics, Vol 4. Amsterdam, The Netherlands: Elsevier Science.

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

Taylor, J. M. G., K. Choi, and P. Han (2023). Data integration - exploiting ratios of parameter estimates from a reduced external model. *Biometrika 110*, 119–134.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

REFERENCES

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

Tian, Y. and Y. Feng (2023). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association 118*, 2684–2697.

Tomlins, S. A., J. R. Day, R. J. Lonigro, D. H. Hovelson, J. Siddiqui, L. P. Kunju, R. L. Dunn, S. Meyer, P. Hodge, J. Groskopf, J. T. Wei, and A. M. Chinnaiyan (2016). Urine tmprss2:erg plus pca3 for individualized prostate cancer risk assessment. *European Urology 70*, 45–53.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Zhai, Y. and P. Han (2022). Data integration with oracle use of external information from heterogeneous populations. *Journal of Computational and Graphical Statistics 31*, 1001–1012.

Zhai, Y. and P. Han (2024). Integrating external summary information under population heterogeneity and information uncertainty. *Electronic Journal of Statistics 18*, 5304–5329.

Zhang, H., L. Deng, M. Schiffman, J. Qin, and K. Yu (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika 107*, 689–703.

Biostatistics Innovation Group, Gilead Sciences

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0225

E-mail: hanpeisong@gmail.com

Department of Statistics, Pennsylvania State University

E-mail: hpl5434@psu.edu

Department of Biostatistics, University of Michigan

E-mail: jmgt@umich.edu

Statistica Sinica