The Canadian Journal of Statistics
Vol. 47, No. 4, 2019, Pages 580–603
La revue canadienne de statistique

Synthetic data method to incorporate external information into a current study

Tian GU[1], Jeremy M. G. TAYLOR[1], Wenting CHENG[1], and Bhramar MUKHERJEE[1]

[1]*Department of Biostatistics, University of Michigan, Ann Arbor, MI 48019, U.S.A.*

Key words and phrases: Constrained maximum likelihood; data integration; prediction models; synthetic data.

MSC 2010: Primary 62F30; secondary 62J12.

Abstract: We consider the situation where there is a known regression model that can be used to predict an outcome, Y, from a set of predictor variables $\mathbf{X}$. A new variable B is expected to enhance the prediction of Y. A dataset of size n containing $Y, \mathbf{X}$ and B is available, and the challenge is to build an improved model for $Y|\mathbf{X}, B$ that uses both the available individual level data and some summary information obtained from the known model for $Y|\mathbf{X}$. We propose a synthetic data approach, which consists of creating m additional synthetic data observations, and then analyzing the combined dataset of size $n+m$ to estimate the parameters of the $Y|\mathbf{X}, B$ model. This combined dataset of size $n+m$ now has missing values of B for m of the observations, and is analyzed using methods that can handle missing data (e.g., multiple imputation). We present simulation studies and illustrate the method using data from the Prostate Cancer Prevention Trial. Though the synthetic data method is applicable to a general regression context, to provide some justification, we show in two special cases that the asymptotic variances of the parameter estimates in the $Y|\mathbf{X}, B$ model are identical to those from an alternative constrained maximum likelihood estimation approach. This correspondence in special cases and the method's broad applicability makes it appealing for use across diverse scenarios. *The Canadian Journal of Statistics* 47: 580–603; 2019 © 2019 Statistical Society of Canada

Résumé: Les auteurs considèrent la situation où un modèle de régression connu peut être utilisé pour prédire une réponse Y à partir des prédicteurs $\mathbf{X}$. Une nouvelle variable B devrait permettre d'améliorer les prévisions de Y. Un jeu de données de taille n comportant les variables Y, $\mathbf{X}$ et B est disponible, et le défi consiste à construire un modèle amélioré pour $Y|\mathbf{X}, B$ qui s'appuie sur les données individuelles, mais aussi sur de l'information sommaire du modèle connu $Y|\mathbf{X}$. Les auteurs proposent de générer m nouvelles données synthétiques, puis d'analyser le jeu de données combiné de $n+m$ données pour estimer les paramètres du modèle $Y|\mathbf{X}, B$. Le jeu de données combiné compte $n+m$ données dont m comportant des valeurs manquantes pour B, et son analyse fait donc appel à des méthodes appropriées (comme l'imputation). Les auteurs présentent des études de simulation et illustrent leur méthode avec des données réelles du Prostate Cancer Prevention Trial. Même si l'approche par données synthétiques s'applique dans un contexte général de régression, les auteurs décrivent deux cas spéciaux pour lesquels la variance asymptotique des estimateurs des paramètres dans le modèle $Y|\mathbf{X}, B$ est identique à celle d'une approche au maximum de vraisemblance sous contraintes. Cette correspondance dans des cas spéciaux et la vaste applicabilité de la méthode en font un choix intéressant pour divers scénarios. *La revue canadienne de statistique* 47: 580–603; 2019 © 2019 Société statistique du Canada

Additional Supporting Information may be found in the online version of this article at the publisher's website.

**Author to whom correspondence may be addressed.*
E-mail: gutian@umich.edu

© 2019 Statistical Society of Canada / Société statistique du Canada

1. INTRODUCTION

In clinical biomedicine, many well-known models are used to predict a measure of disease from patient characteristics. Examples include but are not limited to the breast cancer risk calculator (Gail et al., 1989), and the colorectal cancer risk assessment tool (Freedman et al., 2009). These models are usually constructed from large datasets using principled statistical methods to predict a measure of risk or disease state, treating the patient characteristics as predictors. The patient characteristics, denoted as $\mathbf{X}$, can range from traditional epidemiologic, behavioural variables to well-known imaging, genetic and other molecular biomarkers. The predicted outcome variable Y, and the predictors $\mathbf{X}$, are often assumed to be connected through a regression model of the form $Y|\mathbf{X}$. The individual level original data that were used to construct this model are usually not available to the public but what are accessible are certain forms of summary-level information. This information can be available in the form of coefficient estimates for the fitted model, individual prediction probabilities or multiple prediction probabilities from competing models for the same outcome. The equations underlying the existing model may or may not be known.

While these existing models are often based on traditional epidemiologic and behavioural risk factors and well-established biomarkers, wider availability of high throughput data and novel assay technologies are generating new candidate biomarkers, say B, for possible inclusion in existing risk prediction models. Due to the potential improvement of prediction accuracy of the current model, it is ideal to incorporate B into the well-established model $Y|\mathbf{X}$, and construct an expanded prediction model of interest $Y|\mathbf{X}, B$. However, it is very likely that B and $\mathbf{X}$ are assessed only on participants in a study of moderate size and cannot be retrospectively measured on the much larger population used for $Y|\mathbf{X}$ model. It is natural to consider using the information from the well-established model to increase the accuracy of the expanded model. This represents a general statistical challenge to build a good model for $Y|\mathbf{X}, B$ that uses both the known external information from the $Y|\mathbf{X}$ model and the individual level data from a small sample dataset of Y, $\mathbf{X}$ and B.

There exist proposals in the literature to incorporate external information into regression estimation. Imbens & Lancaster (1994) investigate how aggregate data (e.g., the population average of the response) could be used to improve ML estimates in a regression model. More recently, Grill et al. (2015) proposed a simple method of incorporating new markers into an existing calculator via Bayes theorem. Chatterjee et al. (2016) developed a constrained semi-parametric maximum likelihood (CSPML) method for incorporating external coefficients to calibrate the current regression model. The performance of various approaches was assessed in a simulation study by Grill et al. (2017). Cheng et al. (2018, 2019) proposed Bayes and constrained ML methods to incorporate information obtained from external sources into regression estimation. In general, the constrained ML approaches require a specific form for the external information, for example, estimated coefficients from a correctly specified mean model and assumptions regarding the transportability of the distribution of $Y, \mathbf{X}, B$ across the internal and external sample. The constrained maximum likelihood (CML) approach proposed in Cheng et al. (2018) also requires the specification of a model for $B|\mathbf{X}$ and relies on some parametric assumptions. Although the CSPML approach does not require the $Y|\mathbf{X}$ model to be correctly specified or a model for $B|\mathbf{X}$, it does require the transportability of the $\mathbf{X}$ distribution, unless it is known in the external sample. Estes, Mukherjee & Taylor (2018) and Cheng et al. (2018) have found that the violations of this assumption and the small sample size in the internal data will cause unstable estimation.

In this article, we propose a synthetic data framework as a more flexible solution to this genre of problems, motivated by methods developed in the survey methodology literature (Reiter, 2002; Raghunathan, Reiter & Rubin, 2003; Reiter & Kinney, 2012). In this approach synthetic data for Y and $\mathbf{X}$ are generated from the $Y|\mathbf{X}$ model and added to the observed data, then from this combined dataset a model for $Y|\mathbf{X}, B$ is built. Our method relaxes the requirement on the

information that is available from the external model such that the only requirement is the ability to generate predictions of Y given $\mathbf{X}$.

The following is the structure of the remainder of this article: In Section 2, we introduce the notation, assumptions and implementation of the proposed synthetic data method. In Section 3, under various simulation scenarios, we evaluate the performance of the synthetic data method. We demonstrate the proposed method through an application to the Prostate Cancer Prevention Trial (PCPT) data in Section 4. We provide some theoretical justification and insight for the synthetic data method in Section 5. In two special cases we show that with a very large number of synthetic observations, our approach gives identical asymptotic variances for the parameters of the $Y|\mathbf{X}, B$ model as the constrained maximum likelihood (CML) estimation approach that exists in the literature. Because the CML is a maximum likelihood estimator, it is optimal if the models are correctly specified. Since the synthetic data method has the same asymptotic variance, it can also be considered optimal. Concluding remarks are presented in Section 6.

2. METHOD

2.1. General Description of the Problem

Let Y denote the outcome of interest, which can be either continuous or binary. Let $\mathbf{X}$ be a set of p standard variables and let B denote a new biomarker. There are two populations, an external population for which we do not have individual level data and an internal population for which we do have a dataset of size n with subject level data. We will assume that the distributions of $Y|\mathbf{X}, B$ are the same in the two populations, and likewise for the $Y|\mathbf{X}$ distribution. Our target of interest is the mean structure of $Y|\mathbf{X}, B$:

$$g(E(Y|\mathbf{X}, B)) = \gamma_0 + \gamma_{X_1} X_1 + \cdots + \gamma_{X_p} X_p + \gamma_B B, \tag{1}$$

where g is the known link function. We assume that a small dataset of size n with variables $Y, \mathbf{X}$ and a new covariate B is available to us for building the model of interest.

We assume a large, well-characterized previous study from the external population describes the provided information on the calculated distribution of $Y|\mathbf{X}$. This information can come in various forms, including partial or full knowledge of the distribution of the $Y|\mathbf{X}$ model.

2.2. Synthetic Data Method

We propose an algorithmic approach that can produce synthetic data on $(Y, \mathbf{X}, B)$, by using the combination of the available information from the established model and the observations from the current dataset. The synthetic data would incorporate the external information as well as enlarge the sample size, and thus it helps improve the inference about coefficients $\boldsymbol{\gamma}$ in model (1), compared to just analyzing the small dataset based on the observed data.

The synthetic data approach consists of creating m additional synthetic data observations, and then analyzing the combined dataset of size $n + m$ to estimate the parameters of model (1). The synthetic data are created in two steps as shown in Figure 1. In step 1, we replicate $\mathbf{X}$ a large number (say S) times in blocks of n rows to create $m = nS$ additional records. In step 2, we generate pseudo data called Y^* from the known $Y|\mathbf{X}$ distribution for these new m records. Finally, we combine the synthetic observations with the original dataset, and we note that the combined data will now have missing values of B for m observations. The combined data are then analyzed to give an estimate of $\boldsymbol{\gamma}$.

There may be different ways in which the combined data can be analyzed. In Section 5 we present two special cases for which a closed-form maximum likelihood estimate $\boldsymbol{\gamma}$ exists for the combined dataset of size $n + m$. In other cases, like the simulation study settings in Section 3,

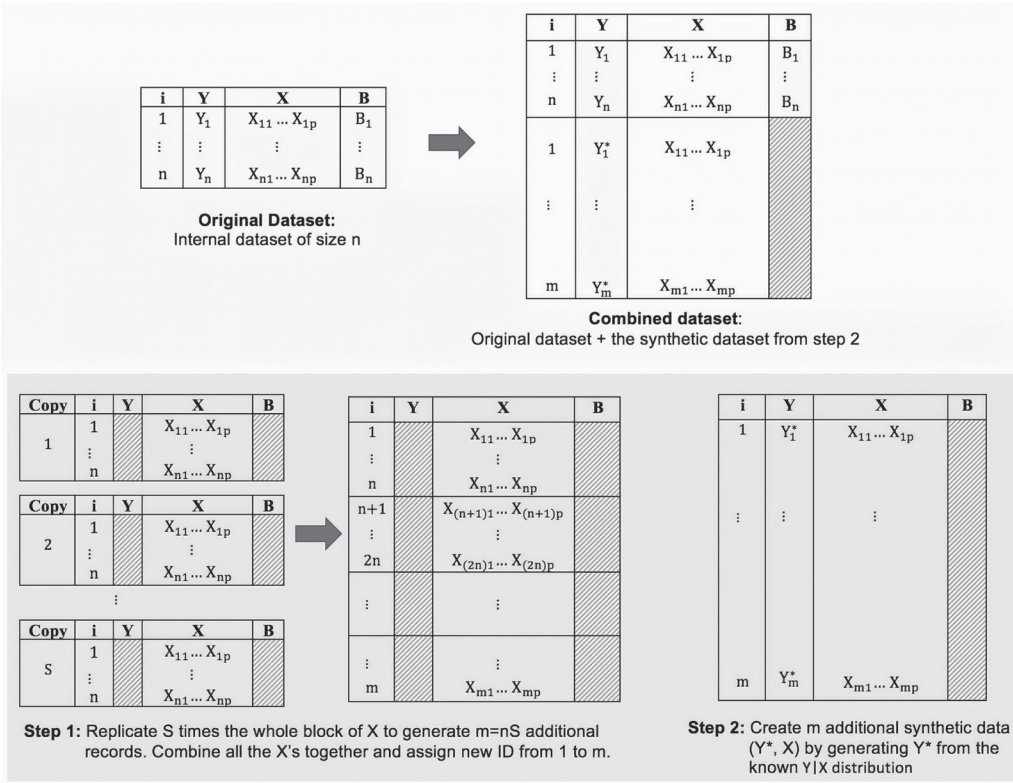

FIGURE 1: Two steps to create the synthetic data.

no closed-form solution for the maximum likelihood estimate of $\boldsymbol{\gamma}$ exists, and our proposed approach to deal with missing data is to use multiple imputation to impute the m missing values of B. Multiple imputation is a general procedure for analyzing datasets with missing values. It consists if defining a procedure to fill in the missing values, then applying that procedure many times to create many separate complete datasets. Each completed dataset is then analyzed and the results of these separate analyses are combined to give final estimates. In this particular case the multiple imputation approach requires us to specify a parametric model $(B|\mathbf{X}, Y)$, from which we draw 50 values of B to give 50 completed datasets. Then we fit model (1) for each complete data $(Y, \mathbf{X}, B)$ of size $n + m$. We then average the estimates of $\boldsymbol{\gamma}$ from the 50 complete datasets, and compute the total variance using Rubin's rules (Rubin, 1987). We then proceed with inference.

Multiple imputation has the additional advantage of being able to handle multiple biomarkers in B, some of which may be discrete and some continuous. It also allows for flexible structure for the conditional mean model for each biomarker in B given all other variables in the dataset, such as the possibility to incorporate non-linearity and interactions. For implementing multiple imputation, we use the R package MICE (Van Buuren & Groothuis-Oudshoorn, 2000). We use the function `mice` with imputation algorithm `logreg` (the Bayesian logistic regression model with flat prior) for the imputation of a binary B and the imputation algorithm `norm` (the Bayesian linear regression model) for the imputation of a continuous B. In the situation in which there are multiple Bs, say B_1 and B_2, imputations are done sequentially. That is, first draw B_1 from the $B_1|\mathbf{X}, Y, B_2$ distribution, then draw B_2 from the $B_2|\mathbf{X}, Y, B_1$ distribution and iterate between B_1 and B_2.

3. SIMULATION STUDY

To assess the performance of the proposed synthetic data method for both estimation and prediction, we conduct simulation studies under four different scenarios. Each scenario has a different true distribution for $Y|\mathbf{X}, B$ and for $B|\mathbf{X}$ for the internal data. For both the outcome and the predictors we consider both continuous and binary variables to illustrate the computational implementation in a range of situations. We also consider the situation of multiple Bs to evaluate the applicability of the synthetic method in the multi-dimensional cases. In some cases a mis-specified imputation model is used within the synthetic data approach, thus allowing us to evaluate the robustness of the method. Only in special cases (see Section 5) can we provide a theoretical justification for the synthetic data approach, thus the simulations are intended to provide numerical properties of the synthetic data approach in situations where the relevant theoretical properties are not yet available.

In real situations, we expect a moderate number of X variables, and their joint distribution could be quite complex with skew distributions and correlations between different Xs. To achieve this we adopted a procedure of generating Xs as described in Xu, Daniels & Winterstein (2016). We generate nine correlated Xs in each of the four simulation scenarios as described below:

$$u_j \sim N(0,1), \quad j = 1, \dots, 5, \quad X_1 \sim N(0,1),$$

$$X_j|u_1, \dots, u_5 = \begin{cases} u_1 u_j + \epsilon_j, \epsilon_j \sim \frac{2}{3}N_+(0, 0.2) + \frac{1}{3}N_-(0, 0.1) & j = 2, \dots, 5 \\ u_2 u_{j-3} + \epsilon_j, \epsilon_j \sim \frac{1}{4}N_+(0, 0.1) + \frac{3}{4}N_-(0, 0.3) & j = 6, \dots, 8 \\ u_3 u_{j-5} + \epsilon_j, \epsilon_j \sim \frac{4}{5}N_+(0, 0.4) + \frac{1}{5}N_-(0, 0.1) & j = 9, \end{cases},$$

where N_+ and N_- represent half normal distributions, and either one or the other is selected with the shown probability. We then generate B from the $B|\mathbf{X}$ distribution, and finally generate Y from the $Y|\mathbf{X}, B$ distribution. For scenario 1 (where Y is continuous, as described below) the form of the external model for $Y|\mathbf{X}$ is readily available. For cases (scenarios 2, 3 and 4 where Y is binary, as described below) where the closed-form of model $Y|\mathbf{X}$ is not available, we numerically derive the external model $Y|\mathbf{X}$. Specifically, we generate an independent dataset of $(Y, \mathbf{X}, B)$ of size 10,000 and fit a linear or logistic regression model $g(E(Y|\mathbf{X}))$ depending on the type of Y. The estimated coefficients of this model serve as the external information we obtained from the established model $Y|\mathbf{X}$.

For each simulation scenario, we first simulate 500 datasets of size n. Then we create the synthetic data following the steps introduced in Section 2, and combine them with the original data to get 500 datasets of size $n + m$ with m missing B values. For each simulated dataset, we create 50 complete datasets by imputing the missing B values given Y and $\mathbf{X}$. In all four scenarios, we use linear additive models for imputing from the $B|\mathbf{X}, Y$ distribution, without including any interaction terms. We compare the results of the synthetic data method to the direct MLE, which uses the complete dataset of size n, in terms of estimation accuracy and prediction ability. We report the average estimated coefficients, standard deviation and 95% coverage rate for $\hat{\boldsymbol{\gamma}}$. To measure the predictive performance, we generate a new dataset of size 1,500 for each scenario, and evaluate the prediction $\widehat{Y}_i$ in this new dataset. In the new dataset, let $\hat{p}$ or $\overline{Y}$ denote the average of the generated Y values. For the continuous Y, we use the mean squared error (MSE) defined as $\sum_{i=1}^{1,500}(\widehat{Y}_i - \overline{Y})^2 / \sum_{i=1}^{1,500}(Y_i - \overline{Y})^2$. For binary Y, we use AUC and scaled Brier score (defined as $\sum_{i=1}^{1,500}(Y_i - \widehat{Y}_i)^2 / \sum_{i=1}^{1,500}(Y_i - \widehat{p})^2$) as measures of predictive performance.

TABLE 1: Simulation results for scenario 1 with Gaussian Y, one Gaussian B and nine correlated Xs: for each method, we report mean (Monte Carlo standard deviation) [95% coverage rate] and MSE across 500 simulated datasets.

	Not including B	True value	Direct MLE	Synthetic data method
γ_0	0	0	−0.04 (0.14) [97%]	0.00 (0.18) [96%]
γ_{X_1}	0.7	0.5	0.50 (0.13) [94%]	0.51 (0.08) [95%]
γ_{X_2}	0.7	0.5	0.48 (0.13) [94%]	0.50 (0.07) [95%]
γ_{X_3}	0.7	0.5	0.49 (0.12) [95%]	0.50 (0.07) [95%]
γ_{X_4}	0.7	0.5	0.50 (0.11) [98%]	0.50 (0.07) [95%]
γ_{X_5}	0.7	0.5	0.50 (0.13) [93%]	0.50 (0.08) [93%]
γ_{X_6}	0.7	0.5	0.50 (0.12) [94%]	0.50 (0.07) [95%]
γ_{X_7}	0.7	0.5	0.50 (0.12) [94%]	0.50 (0.07) [95%]
γ_{X_8}	0.7	0.5	0.50 (0.11) [97%]	0.50 (0.07) [95%]
γ_{X_9}	0.7	0.5	0.50 (0.13) [94%]	0.51 (0.07) [95%]
γ_B	–	0.5	1.00 (0.13) [95%]	1.00 (0.11) [95%]
MSE	0.464	0.334	0.355	0.345

The four simulation scenarios and results are described as follows:

- Scenario 1: Y and B are Gaussian distributed. The true model of $Y|\mathbf{X}, B$ is $Y_i|\mathbf{X_i}, B_i = 0.5\sum_{j=1}^{9} X_{ji} + B_i + e_i$, $e_i \sim N(0, 3)$, and B_i is simulated as $B_i = 0.2(\sum_{j=1}^{9} X_{ji}) + f_i$, $f_i \sim N(0, 1)$. The corresponding $Y|\mathbf{X}$ model is $Y_i = 0.7\sum_{j=1}^{9} X_{ji} + B_i + k_i$, $k_i \sim N(0, 4)$. The current data sample size $n = 200$, replication number $S = 10$, and thus the synthetic data sample size $m = nS = 2{,}000$.
  The results in Table 1 for scenario 1 show that compared to the direct MLE, the synthetic data method leads to an obvious reduction in standard deviation of γ_Xs and good coverage rates of confidence intervals. In addition, it is able to move the MSE closer to the true value by 50%.
- Scenario 2: Y is binary and B is Gaussian distributed. The true model of $Y|\mathbf{X}, B$ is $\text{logit}(Pr(Y_i = 1|\mathbf{X_i}, B_i)) = 1 + 2\sum_{j=1}^{9} X_{ji} - 3B_i$, which gives $Pr(Y = 1) \approx 0.67$. B_i is simulated as $B_i = 0.5\big(\sum_{j=1}^{9} X_{ji}\big) + e_i$, $e_i \sim N(0, 0.1)$. The current data sample size $n = 400$, $S = 10$ and $m = nS = 4{,}000$.
  In Table 2, where Y is binary and B follows Gaussian distribution, including B into the regression model can reduce the scaled Brier score by 15%, and improve the AUC by 9%.
- Scenario 3: Y and B are both binary. The true model of $Y|\mathbf{X}, B$ is $\text{logit}(Pr(Y_i = 1|\mathbf{X_i}, B_i)) = -1 + 0.2\sum_{j=1}^{4} X_{ji} - 0.2\sum_{j=5}^{7} X_{ji} - 0.5\sum_{j=8}^{9} X_{ji} + 1.5B_i$, and B_i is simulated as $\text{logit}(Pr(B_i = 1|\mathbf{X_i})) = -0.5 + 0.5\sum_{j=1}^{5} X_{ji} - \sum_{j=6}^{9} X_{ji}$. The $Pr(Y = 1)$ and $Pr(B = 1)$ are around 0.5 and 0.55, respectively. $n = 400$, $S = 8$ and $m = nS = 3{,}200$.
  The simulation results in Table 3 for scenario 3, in which Y and B are both binary, show that including B in the regression model can reduce the scaled Brier score by 10.8%, and increase the AUC by 5%.
- Scenario 4: Y is binary and two mixed types of B are included, one binary and another Gaussian. The true model of $Y|\mathbf{X}, B_1, B_2$ is $\text{logit}(Pr(Y_i = 1|\mathbf{X_i}, B_{1i}, B_{2i})) = -0.2 - 0.2X_1 + 0.2\sum_{j=2}^{3} X_{ji} + 0.1\sum_{j=4}^{5} X_{ji} - 0.1X_6 - 0.3X_7 + 0.3\sum_{j=8}^{9} X_{ji} + 2B_{1i} - B_{2i}$, from which $\Pr(Y = 1)$

TABLE 2: Results for simulation scenario 2 with binary Y, one Gaussian B and nine correlated Xs: for each method, we report mean (Monte Carlo standard deviation) [95% coverage rate], average scaled Brier score and AUC across 500 simulated datasets.

	Not including B	True value	Direct MLE	Synthetic data method
γ_0	0.849	1	1.00 (0.17) [94%]	0.96 (0.08) [92%]
γ_{X_1}	0.435	2	2.01 (0.29) [96%]	1.92 (0.24) [94%]
γ_{X_2}	0.432	2	2.00 (0.28) [96%]	1.90 (0.23) [94%]
γ_{X_3}	0.437	2	2.01 (0.28) [95%]	1.90 (0.23) [95%]
γ_{X_4}	0.433	2	2.01 (0.30) [96%]	1.91 (0.24) [95%]
γ_{X_5}	0.422	2	2.02 (0.28) [95%]	1.90 (0.24) [93%]
γ_{X_6}	0.421	2	2.01 (0.28) [97%]	1.89 (0.23) [93%]
γ_{X_7}	0.431	2	2.01 (0.29) [96%]	1.91 (0.23) [94%]
γ_{X_8}	0.415	2	2.00 (0.27) [97%]	1.89 (0.23) [94%]
γ_{X_9}	0.445	2	2.01 (0.29) [96%]	1.92 (0.23) [95%]
γ_B	–	−3	−3.02 (0.45) [97%]	−2.85 (0.43) [95%]
Scaled Brier score	0.801	0.680	0.702	0.686
AUC	0.767	0.837	0.828	0.835

TABLE 3: Results for simulation scenario 3 with binary Y, one binary B and nine correlated Xs: for each method, we report mean (Monte Carlo standard deviation) [95% coverage rate], average scaled Brier score and AUC across 500 simulated datasets.

	Not including B	True value	Direct MLE	Synthetic data method
γ_0	−0.328	−1	−1.00 (0.21) [94%]	−1.00 (0.15) [96%]
γ_{X_1}	0.305	0.2	0.20 (0.13) [94%]	0.20 (0.05) [96%]
γ_{X_2}	0.318	0.2	0.21 (0.13) [95%]	0.21 (0.05) [94%]
γ_{X_3}	0.296	0.2	0.20 (0.13) [95%]	0.21 (0.05) [95%]
γ_{X_4}	0.296	0.2	0.19 (0.13) [93%]	0.20 (0.05) [95%]
γ_{X_5}	−0.066	−0.2	−0.21 (0.13) [94%]	−0.19 (0.05) [96%]
γ_{X_6}	−0.405	−0.2	−0.20 (0.13) [96%]	−0.20 (0.06) [96%]
γ_{X_7}	−0.420	−0.2	−0.19 (0.13) [96%]	−0.21 (0.06) [95%]
γ_{X_8}	−0.698	−0.5	−0.50 (0.15) [93%]	−0.51 (0.06) [96%]
γ_{X_9}	−0.713	−0.5	−0.51 (0.14) [95%]	−0.52 (0.06) [94%]
γ_B	–	1.5	1.50 (0.28) [96%]	1.49 (0.28) [95%]
Scaled Brier score	0.750	0.666	0.687	0.669
AUC	0.789	0.833	0.823	0.831

TABLE 4: Results for simulation scenario 4 with binary Y, binary $B1$ and continuous $B2$ and nine correlated Xs: for each method, we report mean (Monte Carlo standard deviation) [95% coverage rate], average scaled Brier score and AUC across 500 simulated datasets.

	Not including B	True value	Direct MLE	Synthetic data method
γ_0	0.250	−0.2	−0.197 (0.21) [96%]	−0.13 (0.15) [93%]
γ_{X_1}	0.441	−0.2	−0.19 (0.19) [96%]	−0.14 (0.13) [94%]
γ_{X_2}	0.779	0.2	0.21 (0.21) [94%]	0.23 (0.13) [94%]
γ_{X_3}	−0.132	0.2	0.19 (0.17) [95%]	0.17 (0.10) [94%]
γ_{X_4}	−0.218	0.1	0.09 (0.17) [95%]	0.08 (0.10) [95%]
γ_{X_5}	1.047	0.1	0.10 (0.26) [96%]	0.15 (0.21) [95%]
γ_{X_6}	0.705	−0.1	−0.10 (0.28) [94%]	−0.06 (0.21) [94%]
γ_{X_7}	0.529	−0.3	−0.29 (0.27) [94%]	−0.25 (0.21) [93%]
γ_{X_8}	−0.750	0.3	0.29 (0.25) [97%]	0.23 (0.21) [95%]
γ_{X_9}	−0.757	0.3	0.30 (0.25) [96%]	0.22 (0.21) [94%]
γ_{B1}	–	1	0.996 (0.31) [96%]	0.88 (0.30) [93%]
γ_{B2}	–	2	1.99 (0.45) [95%]	1.83 (0.43) [93%]
Scaled Brier score	0.637	0.575	0.598	0.582
AUC	0.849	0.876	0.868	0.873

is approximately 0.53. The binary B_{1i} is simulated as $\text{logit}(Pr(B_{1i} = 1|\mathbf{X_i})) = -0.5 + 0.5 \sum_{j=1}^{5} X_{ji} - \sum_{j=6}^{9} X_{ji}$, which gives $Pr(B_1 = 1) \approx 0.56$. The Gaussian B_{2i} is simulated as $B_{2i} = 0.3 \sum_{j=1}^{2} X_{ji} - 0.2 \sum_{j=3}^{4} X_{ji} + 0.5 \sum_{j=5}^{7} X_{ji} - 0.5 \sum_{j=8}^{9} X_{ji} + e_i$, $e_i \sim \text{N}(0, 0.1)$, where $n = 400$, $S = 8$ and $m = nS = 3{,}200$.

The results in Table 4 show that the synthetic data method does improve the scaled Brier score and AUC compared to the MLE and these almost attain the best possible values, and that the coverage rates of the confidence intervals for the $\boldsymbol{\gamma}$s are good.

Overall, the simulation studies show that: (1) the synthetic data method can improve the efficiency of estimating γ_Xs and reduce the MSE of the predictions and increase the AUC for binary Y; (2) in scenario 1 where the $B|\mathbf{X}, Y$ model used for imputation is correctly specified, there is no bias in the estimates of γ_B and the γ_Xs; (3) in scenarios 2, 3 and 4 where the $B|\mathbf{X}, Y$ model used in the imputation is mis-specified, despite the improved predictive performance there is some bias in the estimates of γ_B and the γ_Xs. In future work, we will investigate if even further improvements in performance can be achieved using alternative or more flexible or more non-parametric approaches for imputing B.

4. PCPT DATA EXAMPLE

To assess the performance of the synthetic data method in a real example, we apply it to the PCPT calculator. The high-grade prostate cancer calculator (PCPThg) (Thompson et al., 2016) predicts the probability of high-grade prostate cancer derived from a logistic regression based on standard clinical variables—PSA level, age, DRE findings, prior biopsy result and ethnicity.

The equation for the model is:

$$\text{logit}(p_i) = -6.25 + 0.03\text{age}_i + 0.96\text{race}_i + 1.29\log(\text{PSA}_i) + 1.00\text{DRE}_i - 0.36\text{biopsy}_i, \quad (2)$$

where p_i is the probability of observing high grade prostate cancer for subject i. A detailed description of the calculator and the external and internal and a validation dataset are given in Tomlins et al. (2015) and Cheng et al. (2018). We consider incorporating two biomarkers that have been shown to be predictive of prostate cancer into model (2). One is prostate cancer antigen 3 (PCA3), a continuous variable, and the other is the indicator variable of TMPRSS2:ERG (T2:ERG) gene fusions. We consider three different expanded models, one with the addition of PCA3 only, one with the addition of T2:ERG only and one with the addition of both PCA3 and T2:ERG. To compare the coefficient estimation across methods, we show the estimated coefficients and standard errors in Table 5 from 679 observations in the internal dataset. To compare prediction power, we calculate the scaled Brier score and the AUC based on the validation dataset with 1,218 observations.

For both of the expanded PCPThg models incorporating PCA3 score or binary T2:ERG, if we compare the standard errors across different methods, it is easily seen that the synthetic data method can reduce the standard errors of regression coefficients compared to direct regression by at least 50%.

The expanded PCPThg model incorporating both PCA3 score and binary T2:ERG fitted to the training dataset again shows that the method can reduce the standard errors of regression coefficients compared to direct regression. The results in Table 5 show no improvement in AUC from using the synthetic data approach compared to direct MLE, but noticeable improvement in the Brier score.

We also include in Table 5 the estimates from applying the CSPML. It is a published method that can be applied in this case. We see it gives similar predictive performance as the synthetic data method, but the estimated coefficients differ.

5. ALGEBRAIC JUSTIFICATION IN TWO SPECIAL CASES

5.1. Estimation and Variance of $\boldsymbol{\gamma}$

To establish that the synthetic data approach is asymptotically as efficient as constrained ML approaches, we consider two special cases where closed-form results of MLE for the combined dataset of size $n + m$ in the synthetic data approach are available, so multiple imputation does not need to be used. For these cases, we compare the synthetic data approach to the basic constrained ML method (CML, Cheng et al., 2018) and the constrained semi-parametric ML method (CSPML, Chatterjee et al., 2016). These two maximum likelihood approaches are optimal based on their assumptions. The standard ML approach based on just the observed data without incorporating external information is also provided for reference and comparison. For each approach, we derive the explicit formulas for the asymptotic variance of estimated coefficients, namely, $\hat{\boldsymbol{\gamma}}$ in model (1).

The rationale for studying these two examples in depth is to establish some theoretical underpinning for the synthetic data approach. Given its broad applicability to other more general situations with a mixed set of continuous and categorical multivariable predictors in $\mathbf{X}$ and B, a justification in simpler cases that can be studied analytically makes the approach more plausible in other situations where studying the analytical properties is complicated.

We will be considering three different likelihoods, one based on the distribution $Y|\mathbf{X}, B$, one based on the distribution $(Y, B)|\mathbf{X}$ and one based on the joint distribution of $Y, \mathbf{X}$ and B. When writing distributions, we will include the parameters when necessary, for example, $f(Y|\mathbf{X}, B, \boldsymbol{\gamma})$, but parameters will be excluded when not necessary.

TABLE 5: Expanded PCPThg model: for each method, point estimate (standard error) from the internal dataset, and the scaled Brier score and the AUC from the validation dataset. The sample size of the internal dataset is 679. The sample size of the validation dataset is 1,218. There are $S = 10$ replicates giving $m = 6{,}790$ in the synthetic data method.

Model	PSA	Age	DRE findings	Prior biopsy history	Race	PCA3	T2:ERG	Scaled Brier score	AUC
Original PCPThg	1.29	0.031	1.00	−0.36	0.96	–	–	0.933	0.707
Estimated PCPThg	1.06 (0.18)	0.033 (0.012)	1.15 (0.26)	−1.44 (0.27)	0.44 (0.29)	–	–	0.975	0.716
Expanded model with PCA3 score									
Direct regression[a]	0.97 (0.19)	0.009 (0.013)	1.06 (0.27)	−1.27 (0.27)	0.05 (0.31)	0.56 (0.08)	–	0.953	0.767
Synthetic data method	1.30 (0.08)	0.012 (0.006)	0.91 (0.13)	−0.56 (0.12)	0.50 (0.14)	0.57 (0.08)	–	0.878	0.765
CSPML	1.22 (0.08)	0.007 (0.005)	0.86 (0.10)	−0.20 (0.08)	0.58 (0.11)	0.56 (0.097)	–	0.888	0.759
Expanded model with binary T2:ERG									
Direct regression[a]	0.98 (0.18)	0.032 (0.012)	1.02 (0.26)	−1.41 (0.27)	0.57 (0.29)	–	0.76 (0.20)	0.930	0.744
Synthetic data method	1.21 (0.07)	0.030 (0.005)	0.96 (0.10)	−0.59 (0.09)	0.99 (0.11)	–	0.76 (0.22)	0.897	0.741
CSPML	1.14 (0.07)	0.032 (0.004)	1.06 (0.14)	−0.52 (0.11)	0.80 (0.17)	–	0.72 (0.20)	0.931	0.742
Expanded model with PCA3 score and binary T2:ERG									
Direct regression[a]	0.94 (0.19)	0.010 (0.010)	1.00 (0.28)	−1.27 (0.28)	0.15 (0.31)	0.52 (0.08)	0.47 (0.21)	0.928	0.776
Synthetic data method	1.23 (0.09)	0.008 (0.007)	0.83 (0.13)	−0.53 (0.11)	0.63 (0.15)	0.55 (0.10)	0.45 (0.20)	0.867	0.773
CSPML	1.20 (0.08)	0.008 (0.005)	0.78 (0.11)	−0.21 (0.09)	0.67 (0.12)	0.52 (0.10)	0.48 (0.27)	0.879	0.769

[a]Firth corrected MLE is used.

For this study we either know the full form of the distribution of $Y|\mathbf{X}$, the mean of which may be characterized by a linear combination of $\mathbf{X}$s, as given in Equation (3), with known $\boldsymbol{\beta}$s and a known link function g_1:

$$g_1(E(Y|\mathbf{X})) = \beta_0 + \beta_{X_1}X_1 + \cdots + \beta_{X_p}X_p \tag{3}$$

or we just know the mean structure but not the full distribution of $Y|\mathbf{X}$.

As mentioned earlier, our interest is in building the mean structure of the $Y|\mathbf{X}, B$ distribution as given in model (1). For some approaches, we will also need to consider the relationship between $\mathbf{X}$ and B for which we specify a model, the mean of which is given by

$$g_2(E(B|\mathbf{X})) = \theta_0 + \theta_1 X_1 + \cdots + \theta_p X_p. \tag{4}$$

We note that for all of these models there may be additional parameters necessary to define the full distributions (e.g., the variance σ_β for Gaussian Y). But for ease of notation we will not include these additional parameters unless it is necessary, thus we denote the distributions as $f(Y|\mathbf{X}, \boldsymbol{\beta})$, $f(Y|\mathbf{X}, B, \boldsymbol{\gamma})$ and $f(B|\mathbf{X}, \boldsymbol{\theta})$.

For comparison, we will also present results for standard ML estimation on a complete dataset of size n. In this approach, we estimate the parameters of model (1) using the internal dataset of $Y, \mathbf{X}$ and B without taking the external summary-level information into account. We obtain the estimates by maximizing the likelihood $\prod_{i=1}^{n} f(Y_i|\mathbf{X_i}, B_i, \boldsymbol{\gamma})$ over $\boldsymbol{\gamma}$. Then the asymptotic covariance matrix of $\hat{\boldsymbol{\gamma}}$ is obtained from the inverse of the Fisher information matrix.

- *Approach 1: The synthetic data method.* In special cases in which a direct solution is possible, the likelihood for a dataset of size $n + m$ is $\prod_{i=1}^{n} f(Y_i, B_i|\mathbf{X_i}) \prod_{i=n+1}^{n+m} f(Y_i|\mathbf{X_i})$, and can also be written as $\prod_{i=1}^{n} f(B_i|Y_i, \mathbf{X_i}) \prod_{i=1}^{n+m} f(Y_i|\mathbf{X_i})$. This likelihood is then maximized over $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ to obtain the ML estimates and the asymptotic variance is obtained from the inverse of the Fisher information.
- *Approach 2: Constrained ML on a complete dataset of size n.* For this approach we posit a model $f(B|\mathbf{X}, \boldsymbol{\theta})$ then maximize the likelihood $\prod_{i=1}^{n} f(Y_i, B_i|\mathbf{X_i})$, which can be written as

$$\prod_{i=1}^{n} f(Y_i|\mathbf{X_i}, B_i, \boldsymbol{\gamma}) f(B_i|\mathbf{X_i}, \boldsymbol{\theta})$$

  subject to a constraint on the parameters that is derived from the external information. The equation $f(Y|\mathbf{X}, \boldsymbol{\beta}) = \int f(Y|\mathbf{X}, B, \boldsymbol{\gamma}) f(B|\mathbf{X}, \boldsymbol{\theta}) dB$ gives a relationship between the unknown parameters $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ and the known parameter $\boldsymbol{\beta}$. Assuming $\boldsymbol{\theta}$ can be written as a function of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, that is, as $\boldsymbol{\theta}(\boldsymbol{\gamma}, \boldsymbol{\beta})$, then since β is known the optimization problem becomes an unconstrained optimization problem, specifically maximization of

$$\prod_{i=1}^{n} f(Y_i|\mathbf{X_i}, B_i, \boldsymbol{\gamma}) f(B_i|\mathbf{X_i}, \boldsymbol{\theta}(\boldsymbol{\gamma}, \boldsymbol{\beta} = \boldsymbol{\beta}^*))$$

  with respect to $\boldsymbol{\gamma}$ using the known value $\boldsymbol{\beta}^*$ of $\boldsymbol{\beta}$. We consider two variations of the CML method: Approach 2.1 where only the coefficients $\boldsymbol{\beta}$ are known, and Approach 2.2 where both $\boldsymbol{\beta}$ and σ_β are known.
- *Approach 3: Constrained semi-parametric ML method applied to a dataset of size n.* For this method, the estimates are obtained by maximizing the likelihood $\prod_{i=1}^{n} f(Y_i, \mathbf{X_i}, B_i)$ over $\boldsymbol{\gamma}$ and the empirical distribution of $(\mathbf{X}, B)$, subject to a constraint. In this approach the distribution

of $(\mathbf{X}, B)$ is treated non-parametrically, and the constraint is derived from the integrated score equation of model (3). In this case the constraint is $E_{XB}\left[E_{Y|XB}\left[\frac{\partial}{\partial\boldsymbol{\beta}}\log\{f(Y|\mathbf{X},\boldsymbol{\beta})\}\right]\right] = 0$. The constrained optimization problem is implemented via Lagrange multipliers and gives both an estimate of $\boldsymbol{\gamma}$ and the non-parametric MLE of the distribution of $(\mathbf{X}, B)$. The asymptotic variance of $\hat{\boldsymbol{\gamma}}$ for this approach is given by $(\mathbf{I} + \mathbf{C}\mathbf{L}^{-1}\mathbf{C}^T)^{-1}$, where

$$\mathbf{I} = E_{XB}\left[E_{Y|XB}\left[-\frac{\partial^2}{\partial\boldsymbol{\gamma}^2}\log\{f(Y|\mathbf{X},B,\boldsymbol{\gamma})\}\right]\right]$$

$$\mathbf{C} = E_{XB}\left[E_{Y|XB}\left[\frac{\partial}{\partial\boldsymbol{\gamma}}\log\{f(Y|\mathbf{X},B,\boldsymbol{\gamma}\}\frac{\partial}{\partial\boldsymbol{\beta}}\log\{f(Y|\mathbf{X},\boldsymbol{\beta}\}\right]\right]$$

$$\mathbf{L} = E_{XB}\left[u_\gamma(\mathbf{X},B)u_\gamma^T(\mathbf{X},B)\right]$$

with $u_\gamma(\mathbf{X},B) = E_{Y|XB}\left[\frac{\partial}{\partial\boldsymbol{\beta}}\log\{f(Y|\mathbf{X},\boldsymbol{\beta})\}\right]$.

Intuitively, the asymptotic variance of this constrained ML estimator is the inverse of information matrix $\mathbf{I}$ of $f(Y|\mathbf{X},B,\boldsymbol{\gamma})$ plus the additional information due to knowing $\boldsymbol{\beta}$ from the external study $\mathbf{C}\mathbf{L}^{-1}\mathbf{C}^{\mathrm{T}}$.

5.2. Description of Two Special Cases

In the following two special cases, the goal is to derive the asymptotic efficiency of $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_X, \hat{\gamma}_B)^T$ through a closed-form expression for $\mathrm{Var}(\hat{\boldsymbol{\gamma}})$, and then compare the efficiency gain among all three approaches through the asymptotic relative efficiency (ARE) of $\mathrm{Var}(\hat{\boldsymbol{\gamma}})$s, compared to $\mathrm{Var}(\hat{\boldsymbol{\gamma}})$ from the standard MLE. This will show how much efficiency we can gain by incorporating the external information from the $Y|\mathbf{X}$ model and what determines that gain. We provide all the algebraic details of the derivations in the Appendix.

Special case 1: Y and B are Gaussian distributed

In this section, we assume that Y and B are continuous and have a Gaussian distribution, and assume the identity link for g_1 and g_2 in models (3) and (4).

Without loss of generality, we consider a simplified situation where $p = 1$. We also assume the marginal means of Y_i, X_i and B_i are all equal to zero, thus we use a no-intercept model. Let σ_X^2 denote the variance of X. Then

$$Y|X \sim N(\beta X, \sigma_\beta^2), \tag{5}$$

$$Y|X,B \sim N(\gamma_X X + \gamma_B B, \sigma_\gamma^2), \tag{6}$$

$$B|X \sim N(\theta X, \sigma_\theta^2). \tag{7}$$

Depending on the information available from the external model $Y|X$, we consider two possible situations which correspond to two different constraints. The first situation is when the estimated coefficient $\beta = \beta^*$ is known from model (5). This gives the constraint $\theta = \theta^* = \frac{\beta^* - \gamma_X}{\gamma_B}$. The second situation is when both of the estimated coefficient $\beta = \beta^*$ and the variance $\sigma_\beta^2 = \sigma_\beta^{*2} = \gamma_B^2\sigma_\theta^{*2} + \sigma_\gamma^2$ are known.

For the standard MLE of the complete dataset of size n, it is easy to show that the asymptotic variance of $\hat{\gamma}_X$ and $\hat{\gamma}_B$ is equal to $\frac{\sigma_\gamma^2}{n\sigma_\theta^2}\left(\theta^2 + \frac{\sigma_\theta^2}{\sigma_X^2}\right)$ and $\frac{\sigma_\gamma^2}{n\sigma_\theta^2}$, respectively. The detailed algebraic

TABLE 6: Summary of three approaches when Y and B are Gaussian.

Approach	Method for including external information	Available form of the external information	ARE($\hat{\gamma}$)[a]	
			$\hat{\gamma}_X$	$\hat{\gamma}_B$
Standard MLE (ref)	None	NA	1	1
1: Synthetic data method	m additional synthetic data observations	Ability to draw Y values from $Y\|X$ distribution, regardless of the form	$1 - A^{\mathrm{b}} - \frac{\sigma_{\mathrm{X}}^2\theta^{*2}}{\sigma_{\mathrm{X}}^2\theta^{*2}+\sigma_\theta^{*2}}D^{\mathrm{c}}$	1 − D
2: Constrained MLE (Cheng et al., 2018)	Constraint	2.1: The estimated coefficient β is known.	$1 - A$	1
		2.2: Both of the estimated coefficient β and the standard deviation σ_β are known.	$1 - A - \frac{\sigma_{\mathrm{X}}^2\theta^{*2}}{\sigma_{\mathrm{X}}^2\theta^{*2}+\sigma_\theta^{*2}}D$	1 − D
3: CSPML (Chatterjee et al., 2016)	Constraint	Known expectation of $Y\|X$	$1 - A$	1

[a] $\mathrm{ARE}(\hat{\gamma}) = \mathrm{Var}_M(\hat{\gamma})/\mathrm{Var}_{\mathrm{MLE}}(\hat{\gamma})$, $M \in \{\text{Synthetic Data}, \text{CML}, \text{CSPML}\}$.
[b] $A = \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_{\mathrm{X}}^2\theta^{*2}}\frac{\sigma_\gamma^2}{\sigma_\gamma^2+\gamma_B^2\sigma_\theta^2}$, where $\theta^* = \frac{\beta^*-\gamma_X}{\gamma_B}$.
[c] $D = \frac{2\sigma_\gamma^2\gamma_B^2\sigma_\theta^{*2}}{\sigma_\beta^{*4}}$, where $\sigma_\theta^{*2} = \frac{\sigma_\beta^{*2}-\sigma_\gamma^2}{\gamma_B^2}$.

derivation for each of the three approaches can be found in the Appendix A1–A3. The comparison results for all three approaches are shown in Table 6.

Special case 2: Y, X and B are all binary

Assume we are interested in a saturated model:

$$\mathrm{logit}\{Pr(Y = 1|X, B)\} = \gamma_0 + \gamma_X X + \gamma_B B + \gamma_{XB} XB \tag{8}$$

describing the joint effect of X, B on Y, when Y, X, B are all binary variables. The external information from model (3) can be expressed as:

$$\mathrm{logit}\{Pr(Y = 1|X)\} = \beta_0 + \beta_1 X. \tag{9}$$

The association between B and X is defined through the model:

$$\mathrm{logit}\{Pr(B = 1|X)\} = \theta_0 + \theta_1 X.$$

We denote $P(X = a, Y = b)$ as the probability of $(X = a, Y = b)$ combination and $P(B = 0, X = a, Y = b)$ as the probability of $(B = 0, X = a, Y = b)$ combination, where $a, b \in \{0, 1\}$. The detailed derivation for each of the three approaches can be found in the Appendix B1–B4. The comparison results, showing the AREs, for all three approaches are given in Table 7.

5.3. Summary

Based on the detailed derivation in the Appendix, in Tables 6 and 7, we summarize the methods and the assumed forms of the summary-level external information from the $Y|X$ model for each

approach. The result of ARE of variance $\hat{\boldsymbol{\gamma}}$ compared to the variance from the standard MLE is also given for each of the special cases.

Let the asymptotic relative efficiency under approach M relative to the standard MLE (without external information) be denoted by $\mathrm{ARE_M}(\hat{\gamma}) = \frac{\mathrm{Var}_M(\hat{\gamma})}{\mathrm{Var_{MLE}}(\hat{\gamma})}$, where $M \in \{\text{Synthetic Data}, \text{CML(2.1)}, \text{CML(2.2)}, \text{CSPML}\}$ for Gaussian Y, and $M \in \{\text{Synthetic Data}, \text{CML}, \text{CSPML}\}$ for binary Y.

Result 1. Special case 1: (Y, B) continuous and normally distributed

- $\mathrm{ARE}_{\text{Synthetic Data}}(\hat{\gamma}_X) = \mathrm{ARE}_{\text{CML(2.2)}}(\hat{\gamma}_X) = 1 - A - \frac{\sigma_{\mathrm{X}}^2 \theta^{*2}}{\sigma_{\mathrm{X}}^2 \theta^{*2} + \sigma_\theta^{*2}} D$
  $\mathrm{ARE}_{\text{CSPML}}(\hat{\gamma}_X) = \mathrm{ARE}_{\text{CML(2.1)}}(\hat{\gamma}_X) = 1 - A$
- $\mathrm{ARE}_{\text{Synthetic Data}}(\hat{\gamma}_B) = \mathrm{ARE}_{\text{CML(2.2)}}(\hat{\gamma}_{\mathrm{B}}) = 1 - D$
  $\mathrm{ARE}_{\text{CSPML}}(\hat{\gamma}_B) = \mathrm{ARE}_{\text{CML(2.1)}}(\hat{\gamma}_B) = 1$

where $A = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_{\mathrm{X}}^2 \theta^{*2}} \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \gamma_B^2 \sigma_\theta^2}$, $D = \frac{2\sigma_\gamma^2 \gamma_B^2 \sigma_\theta^{*2}}{\sigma_\beta^{*4}}$, $\theta^* = \frac{\beta^* - \gamma_X}{\gamma_B}$ and $\sigma_\theta^{*2} = \frac{\sigma_\beta^{*2} - \sigma_\gamma^2}{\gamma_B^2}$. Table 6 summarizes the results. In summary, the synthetic data method has the same asymptotic variance as the CML (approach 2.2), and both are more efficient than the CSPML and the CML (approach 2.1). For $\hat{\gamma}_X$ the CSPML and the CML (approach 2.1) are more efficient than the standard MLE. For $\hat{\gamma}_B$ the CSPML and the CML (approach 2.1) have the same efficiency as the standard MLE.

Result 2. Special case 2: (Y, X, B) are all binary

- $\mathrm{ARE}_{\text{Synthetic Data}}(\hat{\gamma}_0) = \mathrm{ARE}_{\text{CML}}(\hat{\gamma}_0) = \mathrm{ARE}_{\text{CSPML}}(\hat{\gamma}_0) = 1 - F$
- $\mathrm{ARE}_{\text{Synthetic Data}}(\hat{\gamma}_X) = \mathrm{ARE}_{\text{CML}}(\hat{\gamma}_X) = \mathrm{ARE}_{\text{CSPML}}(\hat{\gamma}_X) = 1 - G$
- $\mathrm{ARE}_{\text{Synthetic Data}}(\hat{\gamma}_B) = \mathrm{ARE}_{\text{CML}}(\hat{\gamma}_B) = \mathrm{ARE}_{\text{CSPML}}(\hat{\gamma}_B) = 1$
- $\mathrm{ARE}_{\text{Synthetic Data}}(\hat{\gamma}_{XB}) = \mathrm{ARE}_{\text{CML}}(\hat{\gamma}_{XB}) = \mathrm{ARE}_{\text{CSPML}}(\hat{\gamma}_{XB}) = 1$

where $F = \frac{\sum_{(a,b)\in\{(0,1)(0,0)\}} 1/P(X=a,Y=b)}{\sum_{(a,b)\in\{(0,1)(0,0)\}} 1/P(B=0,X=a,Y=b)}$ and $G = \frac{\sum_{a,b\in\{0,1\}} 1/P(X=a,Y=b)}{\sum_{a,b\in\{0,1\}} 1/P(B=0,X=a,Y=b)}$. Table 7 summarizes the results. In conclusion, the synthetic data method, CML and CSPML all converge to the same asymptotic variance. For $\hat{\gamma}_0$ and $\hat{\gamma}_X$, they are more efficient than the standard MLE. For $\hat{\gamma}_B$ and $\hat{\gamma}_{XB}$, they have the same efficiency as the standard MLE.

5.4. Justification from Another Perspective

In the two special cases, we show that using the synthetic data approach with very large m gives identical asymptotic variance for the parameters of model (1) as the constrained ML approach. Below we provide a different intuitive justification for the synthetic data approach, for a more general situation, if certain conditions apply. Assume that Y and B are scalar random variables and that $\mathbf{X}$ is a vector of covariates. We will assume parametric models for all the conditional distributions, and that these can be written as $f(Y, B|\mathbf{X}, \boldsymbol{\phi})$, $f(Y|\mathbf{X}, B, \boldsymbol{\gamma})$, $f(Y|\mathbf{X}, \boldsymbol{\beta})$, $f(B|\mathbf{X}, \boldsymbol{\theta})$ and $f(B|\mathbf{X}, Y, \boldsymbol{\kappa})$. Assume that $f(Y|\mathbf{X}, B, \boldsymbol{\gamma})$ is the model of interest, and that $f(Y|\mathbf{X}, \boldsymbol{\beta})$ is the form of the model that was fit to the external data, and that the estimate of $\boldsymbol{\beta}$ from the external data approximates the true value of $\boldsymbol{\beta}$. We assume that all these models represent the true distributions and are compatible with each other in the sense that $f(Y, B|\mathbf{X}, \boldsymbol{\phi}) = f(Y|\mathbf{X}, B, \boldsymbol{\gamma}) \times f(B|\mathbf{X}, \boldsymbol{\theta}) = f(B|\mathbf{X}, Y, \boldsymbol{\kappa}) \times f(Y|\mathbf{X}, \boldsymbol{\beta})$. We assume there is a 1-to-1 mapping between $\boldsymbol{\phi}$ and $(\boldsymbol{\gamma}, \boldsymbol{\theta})$ and between $\boldsymbol{\phi}$ and $(\boldsymbol{\kappa}, \boldsymbol{\beta})$, and that $\boldsymbol{\kappa}$ and $\boldsymbol{\beta}$ are distinct and that $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ are distinct. With these conditions, we can write $f(Y, B|\mathbf{X}, \boldsymbol{\phi})$ as $f(Y, B|\mathbf{X}, \boldsymbol{\kappa}, \boldsymbol{\beta})$.

TABLE 7: Summary of three approaches when Y, X and B are binary.

Approach	Method for including external information	Available form of the external information	ARE($\hat{\gamma}$)[a]		
			$\hat{\gamma}_0$	$\hat{\gamma}_X$	$\hat{\gamma}_B$ ($\hat{\gamma}_{XB}$)
Standard MLE (ref)	None	NA	1	1	1
1: Synthetic data method	m additional synthetic data observations	Ability to draw Y values from $Y\|X$ distribution, regardless of the form	$1-F$[b]	$1-G$[c]	1
2: Constrained MLE (Cheng et al., 2018)	Constraint	Known estimated coefficient β	$1-F$	$1-G$	1
3: CSPML (Chatterjee et al., 2016)	Constraint	Known expectation of $Y\|X$	$1-F$	$1-G$	1

[a] $\text{ARE} = \text{Var}_M(\hat{\gamma})/\text{Var}_{\text{MLE}}(\hat{\gamma})$, $M \in \{\text{Synthetic Data, CML, CSPML}\}$.
[b] $F = \frac{\sum_{(a,b)\in\{(0,1)(0,0)\}} 1/P(X=a,Y=b)}{\sum_{(a,b)\in\{(0,1)(0,0)\}} 1/P(B=0,X=a,Y=b).}$
[c] $G = \frac{\sum_{a,b\in\{0,1\}} 1/P(X=a,Y=b)}{\sum_{a,b\in\{0,1\}} 1/P(B=0,X=a,Y=b)}$.

With this set-up, the constrained ML estimate is obtained by maximizing the likelihood $\prod_{i=1}^{n} f(Y_i, B_i|\mathbf{X_i}, \boldsymbol{\phi})$ over $\boldsymbol{\phi}$, subject to the known $\boldsymbol{\beta}$. This can be rewritten as maximizing the likelihood $\prod_{i=1}^{n} f(Y_i, B_i|\mathbf{X_i}, \boldsymbol{\kappa}, \boldsymbol{\beta})$ over $\boldsymbol{\kappa}$, subject to the known $\boldsymbol{\beta}$. Then from the combination of the estimate of $\boldsymbol{\kappa}$ and the known $\boldsymbol{\beta}$ we can obtain the estimate of $\boldsymbol{\gamma}$.

The synthetic data method consists of maximizing the likelihood

$$\prod_{i=1}^{n} f(Y_i, B_i|\mathbf{X_i}, \boldsymbol{\phi}) \prod_{i=n+1}^{n+m} f(Y_i|\mathbf{X_i}, \boldsymbol{\beta})$$

which is equivalent to maximizing

$$\prod_{i=1}^{n} f(Y_i, B_i|\mathbf{X_i}, \boldsymbol{\kappa}, \boldsymbol{\beta}) \prod_{i=n+1}^{n+m} f(Y_i|\mathbf{X_i}, \boldsymbol{\beta})$$

over $\boldsymbol{\kappa}$ and $\boldsymbol{\beta}$. When optimizing over $\boldsymbol{\beta}$ for fixed $\boldsymbol{\kappa}$, the second term $\prod_{i=n+1}^{n+m} f(Y_i|\mathbf{X_i}, \boldsymbol{\beta})$ will dominate the optimization procedure when m is very large. Thus the estimate of $\boldsymbol{\beta}$ will essentially reproduce the known value from the external data (since this was the value used to generate the synthetic data). Thus the synthetic data method will reduce to the maximization of the remaining part of the likelihood $\prod_{i=1}^{n} f(Y_i, B_i|\mathbf{X_i}, \boldsymbol{\kappa}, \boldsymbol{\beta})$ with $\boldsymbol{\beta}$ fixed, which is identical to the constrained ML method.

The requirement that all the conditional distributions are compatible with each other will not usually be true, but it may be a reasonable approximation if flexible enough models are being used. The conditions do hold for the normal and the tri-binary examples in Section 5.2 and 5.2 respectively. Another case where they hold is when Y and B follow a bivariate normal distribution given $\mathbf{X}$, that is, $Y, B|\mathbf{X} \sim N\left(\begin{pmatrix}\beta X\\ \theta X\end{pmatrix}, \begin{pmatrix}\sigma_\beta^2 & \rho\sigma_\beta\sigma_\theta\\ \rho\sigma_\beta\sigma_\theta & \sigma_\theta^2\end{pmatrix}\right)$. Then the constrained

ML is to maximize the likelihood $\prod_{i=1}^{n} f(Y_i, B_i|\mathbf{X_i}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_\beta, \sigma_\theta, \rho)$ over $\boldsymbol{\theta}$, σ_θ and ρ subject to known $\boldsymbol{\beta}$ and σ_β.

6. DISCUSSION

In this article, we have introduced the synthetic data method for incorporating summary-level information from well-established external models into the regression model estimation based on internal data. We demonstrated that in some special cases that with a large number of synthetic data observations, the synthetic data approach is asymptotically as efficient as the constrained ML approach. This provides some justification for what at first sight might seem to be an ad-hoc approach. In a simulation study, we demonstrated the ability of the method to improve the predictive ability of the model

A key advantage of the synthetic data method is that it naturally incorporates the prior knowledge into the internal data by creating a large amount of "fake" data that is compatible with the $Y|\mathbf{X}$ established model. By creating pseudo-data from $Y|\mathbf{X}$ instead of using constrained optimization, the synthetic data method not only simplified the task from solving complex constrained optimization, but also provides a potentially more flexible and general framework to handle this problem. The only requirement for the synthetic data approach is the ability to generate Y values given $\mathbf{X}$ from the information of the external models, without the need to know the exact form of model. It is broadly applicable for general data types for $Y, \mathbf{X}$ and B, and when B is more than one new biomarker. It can be extended to the situation where more than one external model is available, that is, $Y|\mathbf{X_1}$, $Y|\mathbf{X_2}, \ldots, Y|\mathbf{X_k}$. In this setting, a combination of external studies that measured overlapping but necessarily identical covariates can provide joint information to develop a model for $Y|\mathbf{X}$ model, where $\mathbf{X}$ is the union of $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_k}$.

The CSPML approach is also broadly applicable, and can handle multiple Bs, and it has some optimality properties. But it does require knowledge of the form of the $Y|\mathbf{X}$ model and requires that the distribution of the Xs are identical in the external and the internal populations, which seems unlikely to be satisfied in practice.

When analyzing the synthetic dataset, the value of B can be considered to be missing, which converts the problem of incorporating external information into a problem of analyzing data with missing values. If multiple imputation procedures are to be used to impute the value of B, then further research would be needed to suggest efficient and robust ways in which this should be implemented. There is the potential to improve even further on the method by using different ways of imputing B, beyond the approach we illustrated in the simulation study.

Another interesting issue that will need to be investigated is the size of m. The theoretical result in this article suggests that m should be very large, but this is under the assumption that the $Y|\mathbf{X}$ and $Y|\mathbf{X}, B$ models are compatible with each other. In practice, they are unlikely to be exactly compatible, which would suggest limiting the size of m. A pragmatic suggestion is to make m equal to the size of the external data, if that is known. By doing this the amount of information in the synthetic data about the relationship between Y and $\mathbf{X}$ is similar to the amount of information in the external data about the relationship between Y and $\mathbf{X}$.

APPENDIX

Derivation of asymptotic variances for the special case 1

A.1. Approach 1: Synthetic Data Method

If the synthetic data approach is applied, and under the assumption that the true value of β and σ_β are used to generate the synthetic data, then the combined data will have the same distribution as a dataset of size $n + m$ in which m values of B have been removed. For this particular data structure, it is possible to obtain formulas for the asymptotic variance of the

maximum likelihood estimates of $\boldsymbol{\gamma}$. In particular, Gourieroux and Monfort (1981) give the exact expression of the ML estimators and the corresponding asymptotic covariance. The likelihood for the combined data is $\prod_{i=1}^{n} f(Y_i, B_i|X_i) \times \prod_{i=n+1}^{n+m} f(Y_i|X_i)$, which can be rewritten as $\prod_{i=1}^{n+m} f(Y_i|X_i) \times \prod_{i=1}^{n} f(B_i|X_i, Y_i)$. Based on this they introduce a set of transformed parameters, and re-parameterized the distributions (5)–(7). They then identify the 1-to-1 relationship among the original parameters and the new set of parameters.

We obtain the estimators of the original parameters by the re-parameterization method, and then apply the delta method to get the asymptotic variance of $\hat{\gamma}_B$ and $\hat{\gamma}_X$.

According to Gourieroux and Monfort (1981), we introduce a set of transformed parameters a, b, c, d and e, and re-parameterized the distributions (5)–(7) as $Y|X \sim N(bX, a^2)$, and $B|Y, X \sim N(dY + eX, c^2)$. We then identify the 1-to-1 relationship among the original parameters and the new set of parameters:

$$\begin{aligned} a^2 &= \sigma^2 + \gamma_B^2 \eta^2, \\ b &= \gamma_X + \theta \gamma_B, \\ c^2 &= \frac{\sigma^2 \eta^2}{a^2}, \\ d &= \frac{\gamma_B \eta^2}{a^2}, \\ e &= \theta - db. \end{aligned} \tag{A.1}$$

The ML estimators $\hat{a}$, $\hat{b}$ and their asymptotic variances are easy to obtain from the linear model

$$Y_i = bX_i + u_i, \mathrm{Var}(u_i) = a^2,$$

where $i = 1, \ldots, n + m$. Similarly, the ML estimators $\hat{c}$, $\hat{d}$ and $\hat{e}$ and their asymptotic variances are easy to obtain from the linear model

$$B_i = dY_i + eX_i + v_i, \mathrm{Var}(v_i) = c^2$$

where $i = 1, \ldots, n$. The estimators of the original parameters are obtained through the relationship derived from equations (A.1), where

$$\begin{aligned} \theta &= bd + e, \\ \eta^2 &= a^2 d^2 + c^2, \\ \gamma_B &= \frac{a^2 d}{\eta^2}, \\ \gamma_X &= b - \gamma_B \theta, \\ \sigma^2 &= \frac{a^2 c^2}{\eta^2}, \end{aligned}$$

and the asymptotic variance of $\hat{\gamma}_B$ and $\hat{\gamma}_X$ can be derived using the delta method:

$$\begin{cases} \mathrm{Var}(\hat{\gamma}_B) = \frac{1}{n}\left[\frac{\sigma^2}{\eta^2} + 2(\lambda - 1)\frac{\gamma_B^2 \sigma^4}{\omega^4}\right] \\ \mathrm{Var}(\hat{\gamma}_X) = \theta^2 \mathrm{Var}(\hat{\gamma}_B) + \frac{1}{n}\left(\sigma^2 \Omega^{-1} \frac{\lambda \sigma^2 + \gamma_B^2 \eta^2}{\omega^2}\right). \end{cases}$$

From this we find the relative efficiency gain of $\mathrm{Var}(\gamma_X, \gamma_B)^T$ by adding m synthetic data observations compared to the original dataset of size n is

$$\mathrm{ARE}(\mathrm{Var}(\hat{\boldsymbol{\gamma}})) = \mathbf{1} - (1-\lambda)\begin{pmatrix} \frac{2\sigma_\gamma^2\gamma_B^2\sigma_\theta^2}{\sigma_\beta^4} + \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_X^2\theta^2}\frac{\sigma_\gamma^2(2\sigma_\gamma^2-\sigma_\beta^2)}{\sigma_\beta^4} \\ \frac{2\gamma_B^2\sigma_\theta^2\sigma_\gamma^2}{\sigma_\beta^4} \end{pmatrix},$$

where $\theta = \frac{\beta-\gamma_X}{\gamma_B}$, and $\sigma_\theta^2 = \frac{\sigma_\beta^2-\sigma_\gamma^2}{\gamma_B^2}$. When m gets very large such that $\lambda \approx 0$, $\mathrm{ARE}(\mathrm{Var}(\hat{\gamma_X})) = 1 - \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_X^2\theta^2}\frac{\sigma_\gamma^2}{\sigma_\gamma^2+\gamma_B^2\sigma_\theta^2}\frac{2\sigma_\gamma^2-\sigma_\beta^2}{\sigma_\beta^2} - \frac{2\sigma_\gamma^2\gamma_B^2\sigma_\theta^2}{\sigma_\beta^4}$, and $\mathrm{ARE}(\mathrm{Var}(\hat{\gamma_B})) = 1 - \frac{2\sigma_\gamma^2\gamma_B^2\sigma_\theta^2}{\sigma_\beta^4}$. This demonstrates some gain in efficiency for both γ_X and γ_B.

A.2. Approach 2: Constrained MLE

Depending on the information available from the external model $Y|X$, there are two possible situations which correspond to two different constraints:

- *Approach 2.1: Only the estimated coefficient β is known from model* (5)
  For models (5)–(7), it is easy to see that the constraint describing the relationship between the unknown variable θ, the known variable β and the target variable $\boldsymbol{\gamma}$ is given by $\theta = \frac{\beta-\gamma_X}{\gamma_B}$. The log-likelihood is given by

$$\begin{aligned} l &= l(\boldsymbol{\gamma}, \theta, \sigma_\gamma^2, \sigma_\theta^2) \\ &= -\frac{n}{2}\log(\sigma_\gamma^2) - \frac{1}{2\sigma_\gamma^2}\sum_{i=1}^n \left(Y_i - \gamma_X X_i - \gamma_B B_i\right)^2 - \frac{n}{2}\log(\sigma_\theta^2) - \frac{1}{2\sigma_\theta^2}\sum_{i=1}^n \left(B_i - \theta Xi\right)^2 \end{aligned} \tag{A.2}$$

  The goal is to maximize the log-likelihood (A.2) over $\boldsymbol{\gamma}$, σ_γ and σ_θ subject to the constraint $\theta = \theta^*$, where $\theta^* = \frac{\beta^*-\gamma_X}{\gamma_B}$. By replacing θ with θ^*, taking the derivative over $\boldsymbol{\gamma}$, and taking the inverse of the matrix, we obtain the asymptotic variance of $\hat{\boldsymbol{\gamma}}$ equals to $\frac{1}{n}\frac{\sigma_\gamma^2}{\sigma_\theta^2}(\theta^{*2} + \frac{\sigma_\theta^4\gamma_B^2}{\sigma_\gamma^2+\sigma_\theta^2\gamma_B^2}\frac{1}{\sigma_X^2}, 1)^T$. The corresponding AREs can be found in Table 6, where we notice that there is some gain in efficiency for γ_X but no gain in efficiency for γ_B. We can see that the largest gain in efficiency is when γ_B, θ and σ_X are small.
- *Approach 2.2: Both of the estimated coefficient β and the standard deviation σ_β are known from model* (5)
  In this situation, knowing the true σ_β gives us more information which is incorporated through an additional constraint. In addition to the constraint $\theta = \theta^* = \frac{\beta-\gamma_X}{\gamma_B^*}$ used in approach 2.1, we add another constraint $\mathrm{Var}(Y|X) = \sigma_\beta^2 = \gamma_B^2\sigma_\theta^2 + \sigma_\gamma^2$, that is, $\sigma_\theta^2 = \sigma_\theta^{*2}$, where $\sigma_\theta^{*2} = \frac{\sigma_\beta^{*2}-\sigma_\gamma^2}{\gamma_B^2}$. Then the log-likelihood (A.2) is maximized with respect to ($\boldsymbol{\gamma}, \sigma_\gamma^2$ at fixed $\sigma_\theta^2 = \sigma_\theta^{*2}, \theta = \theta^*$). Note that unlike in approach 2.1, σ_γ^2 and $\boldsymbol{\gamma}$ are not independent anymore. Thus, we need to consider σ_γ^2 in the information matrix, and take the inverse of a 3×3 matrix to get the correct asymptotic variance.

Let $\boldsymbol{\phi} = (\boldsymbol{\gamma}, \sigma^2)^T$,

$$\mathbf{I} = -E_{XB}\left(\frac{\partial^2 l}{\partial\boldsymbol{\phi}\boldsymbol{\phi}^T}\right) = n\begin{pmatrix} \left(\frac{1}{\sigma^2}+\frac{1}{\gamma_B^2\eta^{*2}}\right)\Omega & \left(\frac{1}{\sigma^2}+\frac{1}{\gamma_B^2\eta^{*2}}\right)\Omega\theta^* & 0 \\ \left(\frac{1}{\sigma^2}+\frac{1}{\gamma_B^2\eta^{*2}}\right)\Omega\theta^* & \left(\frac{1}{\sigma^2}+\frac{1}{\gamma_B^2\eta^{*2}}\right)(\eta^{*2}+\Omega\theta^{*2})+\frac{1}{\gamma_B^2} & \frac{1}{\eta^{*2}\gamma_B^3} \\ 0 & \frac{1}{\eta^{*2}\gamma_B^3} & \frac{1}{2}\left(\frac{1}{\sigma^4}+\frac{1}{\gamma_B^4\eta^{*4}}\right) \end{pmatrix},$$

By taking the inverse of $\mathbf{I}$, we can get the asymptotic variance of $\hat{\boldsymbol{\gamma}}$:

$$\begin{cases} \mathrm{Var}(\hat{\gamma}_B) &= \frac{1}{n}\frac{\sigma^2}{\eta^{*2}}\frac{\sigma^4+\gamma_B^4\eta^{*4}}{(\sigma^2+\gamma_B^2\eta^{*2})^2} = \frac{1}{n}\frac{\gamma_B^2\sigma^2}{\omega^{*2}-\sigma^2}\frac{\sigma^4+(\omega^{*2}-\sigma^2)^2}{\omega^{*4}} \\ \mathrm{Var}(\hat{\gamma}_X) &= \frac{1}{n}\frac{\sigma^2}{\eta^{*2}}\frac{1}{(\sigma^2+\gamma_B^2\eta^{*2})^2}\left[(\eta^{*2}\Omega^{-1}+\theta^{*2})(\sigma^4+\gamma_B^4\eta^{*4})-(\sigma^2-\gamma_B^2\eta^{*2})\sigma^2\eta^{*2}\Omega^{-1}\right] \\ &= (\eta^{*2}\Omega^{-1}+\theta^{*2})\mathrm{Var}(\hat{\gamma}_B)-\frac{1}{n}\sigma^4\Omega^{-1}\frac{\sigma^2-\gamma_B^2\eta^{*2}}{\omega^{*4}}. \end{cases}$$

Thus, we find that the ARE of $\mathrm{Var}(\hat{\boldsymbol{\gamma}})$ from the constrained MLE compared to the standard MLE is identical to the synthetic data method (approach 1). This demonstrates the asymptotic equivalence of the synthetic data approach with large m, to the constrained ML approach that uses knowledge of all the parameters in the $Y|X$ distribution.

A.3. Appendix A3

Approach 3: Constrained Semi-parametric MLE

This approach assumes that β is known, but does not assume that σ_β is known. For this method calculation of the asymptotic variance of $\hat{\boldsymbol{\gamma}}$ requires calculation of the three matrices $\mathbf{I}$, $\mathbf{C}$ and $\mathbf{L}$.

After some algebra for the situation that $Y|X$, B and $B|X$ are both normal it can be shown that $\mathbf{C} = \frac{\Omega}{\omega^2}(1,\theta^*)^T, L = \frac{n\gamma_B^2\eta^*\Omega}{\omega^{*4}}$. Thus,

$$\mathrm{Cov}(\hat{\boldsymbol{\gamma}}) = (\mathbf{I}+\mathbf{C}\mathbf{L}^{-1}\mathbf{C}^T)^{-1} = \frac{1}{n}\frac{\sigma^2}{\eta^2}\begin{pmatrix} \theta^{*2}+\frac{\eta^4\gamma_B^2}{\sigma^2+\eta^2\gamma_B^2}\Omega^{-1} & -\theta^* \\ -\theta^* & 1 \end{pmatrix},$$

which is identical to the covariance matrix in approach 2.1.

Derivation of asymptotic variances for the special case 2, Y, X and B binary

B.1. Standard MLE

We will use the following notation: $S \equiv S_\gamma(X,B) = \gamma_0+\gamma_X X+\gamma_B B+\gamma_{XB}XB$, $M \equiv M_\beta(X) = \beta_0+\beta_1 X$ and $K \equiv K_\theta(X) = \theta_0+\theta_1 X$.

The ML estimators are the solution of maximizing $\prod_{i=1}^n f(Y_i|X_i,B_i,\boldsymbol{\gamma})$, that is, $\max_{\boldsymbol{\gamma}}\left\{\sum_{i=1}^n[Y_iS_i-\log\{1+\exp(S_i)\}]\right\}$.

In the tri-binary case, since $X = X^2$, and $B = B^2$, the Fisher information is

$$\mathbf{I} = E_{XB}I(X,B) = E_{XB}\left[\mathrm{expit}(S)(1-\mathrm{expit}(S))\begin{pmatrix} 1 & X & B & XB \\ X & X & XB & XB \\ B & XB & B & XB \\ XB & XB & XB & XB \end{pmatrix}\right].$$

TABLE A.1: Formulas for $\boldsymbol{\gamma}$ in terms of $P(B|XY)$ and P(XY).

(X, B) combination	Transformation of model (8)
(0, 0)	$\gamma_0 = log\frac{P(B=0\|XY=01)P(XY=01)}{P(B=0\|XY=00)P(XY=00)}$
(1, 0)	$\gamma_0 + \gamma_X = log\frac{P(B=0\|XY=11)P(XY=11)}{P(B=0\|XY=10)P(XY=10)}$
(0, 1)	$\gamma_0 + \gamma_B = log\frac{P(B=1\|XY=01)P(XY=01)}{P(B=1\|XY=00)P(XY=00)}$
(1, 1)	$\gamma_0 + \gamma_X + \gamma_B + \gamma_{XB} = log\frac{P(B=1\|XY=11)P(XY=11)}{P(B=1\|XY=10)P(XY=10)}$

There are a total of four possible combinations of binary (X, B). Thus, the expectation terms of the matrix $I(X, B)$ can be obtained through $\frac{1}{n}\sum_{a,b\in\{0,1\}} I(a,b)P(X=a, B=b)$. The asymptotic variance of $\hat{\boldsymbol{\gamma}}$ is given by:

$$\begin{cases} \mathrm{Var}(\hat{\gamma}_0) = \frac{1}{n}\left(\frac{1}{P(BXY=001)} + \frac{1}{P(BXY=000)}\right), \\ \mathrm{Var}(\hat{\gamma}_X) = \frac{1}{n}\sum_{a,b\in\{0,1\}} \frac{1}{P(BXY=0ab)}, \\ \mathrm{Var}(\hat{\gamma}_B) = \frac{1}{n}\left(\frac{1}{P(B=1|XY=01)P(BXY=001)} + \frac{1}{P(B=1|XY=00)P(BXY=000)}\right), \\ \mathrm{Var}(\hat{\gamma}_{XB}) = \frac{1}{n}\sum_{a,b\in\{0,1\}} \frac{1}{P(B=1|XY=ab)P(BXY=0ab)}, \end{cases}$$

where $P(BXY = 0ab)$ is the probability of the $(B = 0,\ X = a,\ Y = b)$ combination, and $P(B = 1|XY = ab)$ is the probability of $B = 1$ given $X = a$ and $Y = b$, $a, b \in \{0, 1\}$.

B.2. Approach 1: Synthetic Data Method

Motivated by the ML estimation in the missing data problem (Little, 1992), we re-formulate our target likelihood as follows:

$$\begin{aligned} \prod_{i=1}^{n} f(Y_i, X_i, B_i) \prod_{i=n+1}^{m+n} f(X_i, Y_i) &= \prod_{i=1}^{n} f(X_i, Y_i) f(B_i|X_i Y_i) \prod_{i=n+1}^{m+n} f(X_i, Y_i) \\ &= \prod_{i=1}^{n} f(B_i|X_i, Y_i) \prod_{i=1}^{m+n} f(X_i, Y_i), \end{aligned} \tag{B.1}$$

where $f(B_i|X_i, Y_i)$ and $f(X_i, Y_i)$ are independent from each other. The goal is to maximize likelihood (B.1) over $\boldsymbol{\gamma}$.

Let $P(XY = ab) \equiv Pr(X_i = a, Y_i = b),\ a, b \in \{0, 1\},\ i = 1, \ldots, m + n$. With the constraint $\sum_{a,b\in\{0,1\}} P(XY = ab) = 1$, there are a total of three unknown variables in $\prod_{i=1}^{m+n} f(X_i, Y_i)$, that is, P(XY=ab), $a, b \in \{0, 1\}$. Similarly, denote $P(B=1|XY=ab) \equiv Pr(B_i = 1|X_i = a, Y_i = b)$, $i = 1, \ldots, n$. Since there are four different combinations of a and b, there are four unknown parameters (i.e., $P(B = 1|XY = ab)$) in $\prod_{i=1}^{n} f(B_i|X_i, Y_i)$, which are independent from each other.

By plugging the four possible combinations of X and B into model (8) in the main text, we can easily derive the expressions for $\boldsymbol{\gamma}$ as presented in Table A.1.

Let $mn(a,b)$ denote the number of observations with $(X=a, Y=b)$ in the sample size of $m+n$. Since $mn(a,b) \sim \text{Multinomial}(m+n, P(XY=ab))$, we can easily obtain the ML estimation of $P(XY)$, and corresponding estimated covariance as follows:

$$\begin{cases} \hat{P}(XY=ab) = \frac{mn(a,b)}{m+n} \\ \text{Var}(\hat{P}(XY=ab)) = \hat{P}(XY=ab)(1-\hat{P}(XY=ab)) \\ \text{Cov}(\hat{P}(XY=ab), \hat{P}(XY=a'b')) = -\hat{P}(XY=ab)\hat{P}(XY=a'b') \end{cases}$$

Denote $n(a,b)$ as the number of observations with $(X=a, Y=b)$ in the sample of size n, and $n(B=1|XY=ab)$ as the count of $B=1$ given (X = a, Y = b). Since $n(B=1|XY=ab) \sim Binomial(n(a,b), P(B=1|XY=ab))$, the ML estimation of $P(B|X,Y)$ and its estimated covariance can be expressed as:

$$\begin{cases} \hat{P}(B=1|XY=ab) = \frac{n(B=1|XY=ab)}{n(a,b)} \\ \text{Var}(\hat{P}(B=1|XY=ab)) = \frac{\hat{P}(B=1|XY=ab)\hat{P}(B=0|XY=ab)}{n(a,b)} \\ \text{Cov}(\hat{P}(B=1|XY=ab), \hat{P}(B=1|XY=a'b')) = 0 \end{cases}.$$

Therefore, the ML estimation of $\boldsymbol{\gamma}$ can be expressed as:

$$\begin{cases} \hat{\gamma}_0 = log\left(\frac{\hat{P}(B=0|XY=01)}{\hat{P}(B=0|XY=00)} \frac{\hat{P}(XY=01)}{\hat{P}(XY=00)}\right) \\ \hat{\gamma}_X = log\left(\frac{\hat{P}(B=0|XY=11)\hat{P}(B=0|XY=00)}{\hat{P}(B=0|XY=10)\hat{P}(B=0|XY=01)} \frac{\hat{P}(XY=11)\hat{P}(XY=00)}{\hat{P}(XY=10)\hat{P}(XY=01)}\right) \\ \hat{\gamma}_B = log\left(\frac{\hat{P}(B=1|XY=01)}{\hat{P}(B=0|XY=01} \frac{\hat{P}(B=0|XY=00)}{\hat{P}(B=1|XY=00)}\right) \\ \hat{\gamma}_{XB} = log\left(\frac{\hat{P}(B=1|XY=11)\hat{P}(B=1|XY=00)}{\hat{P}(B=1|XY=10)P(B=1|XY=01)} \frac{\hat{P}(B=0|XY=10)\hat{P}(B=0|XY=01)}{\hat{P}(B=0|XY=00)\hat{P}(B=0|XY=11)}\right) \end{cases}.$$

By the delta method, and replacing estimated proportions by the corresponding probabilities we obtain the asymptotic variances

$$\begin{cases} \text{Var}(\hat{\gamma}_0) = \frac{1}{n}\sum_{(a,b)\in\{(0,1),(0,0)\}} \frac{P(B=1|XY=ab)}{P(BXY=0ab)} + \frac{1}{m+n}\sum_{(a,b)\in\{(0,1),(0,0)\}} \frac{1}{P(XY=ab)} \\ \text{Var}(\hat{\gamma}_X) = \frac{1}{n}\sum_{a,b\in\{0,1\}} \frac{P(B=1|XY=ab)}{P(BXY=0ab)} + \frac{1}{m+n}\sum_{a,b\in\{0,1\}} \frac{1}{P(XY=ab)} \\ \text{Var}(\hat{\gamma}_B) = \frac{1}{n}\sum_{(a,b)\in\{(0,1),(0,0)\}} \frac{1}{P(B=1|XY=ab)P(BXY=0ab)} \\ \text{Var}(\hat{\gamma}_{XB}) = \frac{1}{n}\sum_{a,b\in\{0,1\}} \frac{1}{P(B=1|XY=ab)P(BXY=0ab)} \end{cases},$$

where $P(BXY=0ab) = P(B=0|XY=ab)P(XY=ab)$.

Therefore, we find that the ARE of $\text{Var}(\hat{\boldsymbol{\gamma}})$ by adding m synthetic data observations compared to the original dataset of size n is

$$\text{ARE}(\text{Var}(\hat{\boldsymbol{\gamma}})) = \mathbf{1} - (1-\lambda)\begin{pmatrix} \frac{\sum_{(a,b)\in\{(0,1)(0,0)\}} 1/P(XY=ab)}{\sum_{(a,b)\in\{(0,1)(0,0)\}} 1/P(BXY=0ab)} \\ \frac{\sum_{a,b\in\{0,1\}} 1/P(XY=ab)}{\sum_{a,b\in\{0,1\}} 1/P(BXY=0ab)} \\ 0 \\ 0 \end{pmatrix}. \tag{B.2}$$

B.3. Approach 2: Constrained MLE

The summary-level information from model (9) is available in the form of coefficient estimates $\boldsymbol{\beta}$. The constrained ML estimator is the solution of maximizing $\prod_{i=1}^{n} f(Y_i|X_i, B_i, \boldsymbol{\gamma})f(B_i|X_i, \boldsymbol{\theta})$ subject to the constraint that $Pr(Y = 1|X = x, \boldsymbol{\beta}) = \sum_{b=0}^{1} Pr(Y = 1|X = x, B = b, \boldsymbol{\gamma})Pr(B = b|X = x, \boldsymbol{\theta})$.

The log-likelihood can be rewritten as

$$\max_{\gamma}\Big\{ \sum_{i=1}^{n}[Y_iS_i - \log\{1 + \exp(S_i)\} + B_iK_i - \log\{1 + \exp(K_i)\}]\Big\}.$$

From the constraints we can write $\boldsymbol{\theta}$ as a function of $\boldsymbol{\gamma}$ in the following way

$$\begin{cases} \theta_0(\boldsymbol{\gamma}) & = \text{logit}\{\frac{\text{expit}(\beta_0)-\text{expit}(\gamma_0)}{\text{expit}(\gamma_0+\gamma_B)-\text{expit}(\gamma_0)}\} \\ \theta_1(\boldsymbol{\gamma}) & = \text{logit}\{\frac{\text{expit}(\beta_0+\beta_1)-\text{expit}(\gamma_0+\gamma_X)}{\text{expit}(\gamma_0+\gamma_X+\gamma_B+\gamma_{XB})-\text{expit}(\gamma_0+\gamma_X)}\} - \theta_0(\boldsymbol{\gamma}) \end{cases}.$$

Then K becomes $K_\gamma(X) = \theta_0(\boldsymbol{\gamma}) + \theta_1(\boldsymbol{\gamma})X$. Denote

$$\begin{cases} \sigma_{\beta_{0j}} & \equiv \frac{\partial}{\partial\gamma_j}\theta_0(\boldsymbol{\gamma}) \\ \sigma_{\beta_{1j}} & \equiv \frac{\partial}{\partial\gamma_j}\theta_1(\boldsymbol{\gamma}) \end{cases},$$

where $j = 0, 1, 2, 3$. The asymptotic variance of $\hat{\boldsymbol{\gamma}}$ can be derived through the 4×4 matrix $\frac{1}{n}\{E_{XB}[E_{Y|XB}(u_{\boldsymbol{\gamma}}u_{\boldsymbol{\gamma}}{}^T)]\}^{-1}$, where

$$\begin{aligned} u_{\boldsymbol{\gamma}} &= \frac{\partial}{\partial\boldsymbol{\gamma}}\log\{f(Y, B|X, \boldsymbol{\gamma}, \boldsymbol{\theta}(\boldsymbol{\gamma}))\} \\ &= \begin{pmatrix} (\sigma_{\beta_{00}} + \sigma_{\beta_{10}}X)(B - \text{expit}(K)) + Y - \text{expit}(S) \\ (\sigma_{\beta_{01}} + \sigma_{\beta_{11}}X)(B - \text{expit}(K))) + (Y - \text{expit}(S))X \\ (\sigma_{\beta_{02}} + \sigma_{\beta_{12}}X)(B - \text{expit}(K))) + (Y - \text{expit}(S))B \\ (\sigma_{\beta_{03}} + \sigma_{\beta_{13}}X)(B - \text{expit}(K))) + (Y - \text{expit}(S))XB \end{pmatrix}. \end{aligned}$$

Since all Y, X, and B are binary variables, there are a total of eight possible combinations of (Y, X, B). Thus, the expectation term in the matrix $E(\boldsymbol{u}_{\boldsymbol{\gamma}}\boldsymbol{u}_{\boldsymbol{\gamma}}^T)$ can be obtained through

$$\frac{1}{n}\sum_{a,b,c\in\{0,1\}} u_\gamma u_\gamma^T P(Y = a, X = b, B = c).$$

A variation on the above approach is when the external summary information comes in the form of the predicted probability for any X, that is, we are simply provided with $\overline{P}(X_i) = \hat{P}r(Y_i = 1|X_i)$. In these cases, it is easy to construct an estimation method that uses this as a constraint. Also in the special case being considered here where Y and X are binary, it is easy to see that knowing $\overline{P}(0)$ and $\overline{P}(1)$ is equivalent to knowing $\beta_0 = \text{logit}(\overline{P}(0))$ and $\beta_1 = \text{logit}(\overline{P}(1)) - \text{logit}(\overline{P}(0))$, so this also fits into the above framework to obtain the asymptotic variance of $\hat{\boldsymbol{\gamma}}$.

B.4. Approach 3: Constrained Semi-parametric MLE

By implementing the specific distribution into the given formulas for **I**, **C** and **L**, we find that **I** is the same as the information matrix in approach 1, that the 4×2 matrix **C** is the first two columns of matrix **I**, and that

$$\mathbf{L} = E_{XB}\left[(1 - \text{expit}(M))\text{expit}(S) - \text{expit}(M)(1 - \text{expit}(S))\begin{pmatrix} 1 & X \\ X & X^2 \end{pmatrix}\right],$$

where $S \equiv S_\gamma(X, B) = \gamma_0 + \gamma_X X + \gamma_B B + \gamma_{XB} XB$ and $M \equiv M_\beta(X) = \beta_0 + \beta_1 X$. The calculation of **L** is simple under the situation where X and B are both binary. Then **I**, **C** and **L** can be combined to give the variance of $\hat{\boldsymbol{\gamma}}$.

Although we have not written out the formulas for the ARE of $\text{Var}(\hat{\boldsymbol{\gamma}})$ for approaches 2 and 3, we find that their values are numerically identical to those in equation B.2 with $\lambda = 0$.

ACKNOWLEDGEMENTS

This research was partially supported by the National Science Foundation (NSF) and National Institutes of Health (NIH).

BIBLIOGRAPHY

Chatterjee, N., Chen, Y. -H., Maas, P., & 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.

Cheng, W., Taylor, J. M. G., Gu, T., Tomlins, S. A., & 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.

Cheng, W., Taylor, J. M. G., Vokonas, P. S., Park, S. K., & Mukherjee, B. (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., Mukherjee, B., & 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, https://doi.org/10.1007/s12561-018-9217-4.

Freedman, A. N., Slattery, M. L., Ballard-Barbash, R., Willis, G., Cann, B. J., Pee, D., Gail, M. H., & Pfeiffer, R. M. (2009). *Journal of Clinical Oncology*, 27, 686–693.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., & Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81, 1879–1886.

Gourieroux, C. & Monfort, A. (1981). On the problem of missing data in linear models. *Review of Economic Studies*, 48, 579–586.

Grill, S., Fallah, M., Leach, R. J., Thompson, I. M., Hemminki, K., & Ankerst, D. P. (2015). A simple-to-use method incorporating genomic markers into prostate cancer risk prediction tools facilitated future validation. *Journal of Clinical Epidemiology*, 68, 563–573.

Grill, S., Ankerst, D. P., Gail, M. H., Chatterjee, N., & Pfeiffer, R. M. (2017). Comparison of approaches for incorporating new information into existing risk prediction models. *Statistics in Medicine*, 36, 1134–1156.

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

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, 87, 1227–1237.

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1.

Reiter, J. P. & Kinney, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics*, 28, 583.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18, 531.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons Inc., New York.

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

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., & Chinnaiyan, A. M. (2015). Urine TMPRSS2:ERG plus PCA3 for individualized prostate cancer risk assessment. *European Urology*, 70, 45–53. 2015.

Van Buuren, S. & Oudshoom, C. G. M. (2000). *MICE: Multivariate imputation by chained equations (S software for missing-data imputation)*. Available at web.inter.nl.net/users/S.van.Buuren/mi/.

Xu, D., Daniels, M. J., & Winterstein, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics*, 17, 589–602.

Received 19 February 2019
Accepted 06 May 2019