Received: 15 December 2020 | Revised: 22 July 2021 | Accepted: 21 October 2021

DOI: 10.1002/sim.9252

RESEARCH ARTICLE

Statistics in Medicine WILEY

Improving main analysis by borrowing information from auxiliary data

Chixiang Chen[1] | **Peisong Han**[2] | **Fan He**[3]

[1]Division of Biostatistics and Bioinformatics, Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, Maryland

[2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan

[3]Division of Biostatistics and Bioinformatics, Department of Public Health Sciences, Penn State College of Medicine, Hershey, Pennsylvania

Correspondence
Chixiang Chen, Division of Biostatistics and Bioinformatics, Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD 21201, USA.
Email: chixiang.chen@som.umaryland.edu

In many clinical and observational studies, auxiliary data from the same subjects, such as repeated measurements or surrogate variables, will be collected in addition to the data of main interest. Not directly related to the main study, these auxiliary data in practice are rarely incorporated into the main analysis, though they may carry extra information that can help improve the estimation in the main analysis. Under the setting where part of or all subjects have auxiliary data available, we propose an effective weighting approach to borrow the auxiliary information by building a working model for the auxiliary data, where improvement of estimation precision over the main analysis is guaranteed regardless of the specification of the working model. An information index is also constructed to assess how well the selected working model works to improve the main analysis. Both theoretical and numerical studies show the excellent and robust performance of the proposed method in comparison to estimation without using the auxiliary data. Finally, we utilize the Atherosclerosis Risk in Communities study for illustration.

KEYWORDS

auxiliary data, empirical likelihood, estimation efficiency improvement, information borrowing, information index

1 | INTRODUCTION

To improve the estimation accuracy in the main study, auxiliary measurements may play an important role. In the literature, there are multiple types of auxiliary measurements, of which the most popular one is from an external independent study. Such information could be either at the summary level or from other validation data through shared covariate effects between the main and external studies. In presence of this type of auxiliary data, methods based on the generalized method of moment, generalized regression, weight calibration, constrained maximum likelihood, empirical likelihood, etc., have been proposed to borrow auxiliary information to power up the main study.[1-13] In this article, we consider a different type of auxiliary data that is also widely seen in applications, that is, an auxiliary measurement collected in the same study but served as the outcome in a secondary analysis. Usually, such kind of auxiliary measurement is highly associated with the primary outcome, and how to incorporate this secondary information to enhance estimation precision for the main analysis is of high interest.

There are many examples in epidemiological and clinical studies where auxiliary outcomes that are associated with primary outcome are available but the information contained in these variables is rarely incorporated into the main analysis. For instance, in the prospective Cardiovascular Health Study,[14,15] both cardiovascular and cancer outcomes were measured. Investigators may borrow information from cancer outcomes to improve studies in which cardiovascular

disease is the main outcome, or vice versa. Another example is the long-term Health Professional Follow-up Study[16] on the association between coffee drinking status at the baseline and prostate cancer risk. Since data on coffee consumption during the follow-up were also repeatedly measured, these data not used in the main analysis become auxiliary but contain extra information about the main study. A third example is a study on the relationship between risk of influenza and vaccination status in the current flu-season, in which case individuals' vaccination and disease status in previous seasons are auxiliary measurements and may help improve the main analysis. In this article, we use the Atherosclerosis Risk in Communities study[17,18] for illustration. We are interested in detecting baseline risk factors for the development of essential hypertension during the follow-up. In addition to the primary interest, there are certain auxiliary measurements available in the study that are associated with the hypertension development, for example, longitudinal measurements of systolic blood pressure. Since these auxiliary measurements are not typically treated as risk factors for hypertension, in the existing literature they are rarely incorporated into the main analysis for risk factor detection. However, the systolic blood pressure is highly associated with the occurrence of hypertension, it may contain useful information for the estimation in the main analysis.

In the remaining article, let us focus on the auxiliary measurements collected from all or part of the subjects in the same study. To make use of these auxiliary data, we propose an estimation procedure by reweighting the study subjects in their contribution towards the main analysis. The weights are calculated based on the empirical likelihood method[19] through specifying a working model for the auxiliary data. Note that weighting based on the empirical likelihood have already been investigated in certain areas, including missing data problems, casual inference, and longitudinal quantile regression.[6,20-25] Distinct from existing literature, we allow the auxiliary data and the main data to contain completely different variables, which implies a broader applicability for information borrowing. The auxiliary data are also allowed to be available only from a subset of study subjects. One of desirable features of the proposed method is that under mild conditions, mis-specification of the working model for the auxiliary data will not affect estimation consistency of the main analysis but still improves estimation precision in comparison to estimation without the auxiliary data. Here, misspecification also refers to informative missing data issues when building the working model, in addition to misspecification of the working model itself. We further propose an index of information borrowing (*IIB*) to assess the overall performance of the estimation procedure, serving as a criterion in practice for selecting a proper auxiliary data set or a desirable working model.

The rest of paper is organized as follows. Section 2 describes the new estimation procedure with its theoretical properties in a general setting. Two examples are then illustrated. Section 3 conducts simulation studies to evaluate the numerical performance of the proposed method. An application to the Atherosclerosis Risk in Communities study data is presented in Section 4. Some discussions are given in Section 5. The technical proofs, extra discussions, and other simulation results are presented in Appendix S1.

2 | PROPOSED ESTIMATION PROCEDURE

We start with a general framework. For each subject i from a random sample, $i = 1, \dots, n$, let $\boldsymbol{D}_i^u$ be the data for the main analysis, which include both the outcome Y_i (eg, occurrence of hypertension in ARIC study) and covariates $\boldsymbol{X}_i$ (eg, risk factors such as age, bmi, etc.) of primary interest. The model of interest is a regression of the main outcome on potential risk factors, with regression parameters $\boldsymbol{\beta}$ estimated by solving the estimating equations

$$\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{D}_i^u, \boldsymbol{\beta}) = \boldsymbol{0}. \tag{1}$$

Here, the estimating functions $\boldsymbol{g}(\boldsymbol{D}_i^u, \boldsymbol{\beta})$ can be the derivative of least squares, the score function from a likelihood, the generalized estimating equations (GEE), etc., based on different specifications of the regression model of interest. For illustration, we assume that all primary data $\boldsymbol{D}_i^u$ are observed. Some more complicated setups, such as in missing data framework, are discussed in Section 5 and section 3.2 in Appendix S1. Let $\boldsymbol{\beta}_0$ be the true parameter values such that $E\{\boldsymbol{g}(\boldsymbol{D}_i^u, \boldsymbol{\beta}_0)\} = \boldsymbol{0}$. Then, the solution $\hat{\boldsymbol{\beta}}$ to the equation (1) is consistent for $\boldsymbol{\beta}_0$ and has an asymptotic normal distribution under some regularity conditions.[26]

In addition to the data $\boldsymbol{D}_i^u$ for the main analysis, many studies have secondary (auxiliary) variables collected, such as longitudinal measurements of systolic blood pressure in ARIC study. Let $\boldsymbol{D}_i^a$ be the auxiliary data collected on subject i. It includes an auxiliary outcome with certain covariates, which is believed to contain information to improve over the main estimation. To avoid confusion, hereafter, we call $\boldsymbol{D}_i^a$ the auxiliary data/variables and use terms auxiliary

outcome/covariates to denote the outcome/covariates in the auxiliary data $\boldsymbol{D}_i^a$. In practice, not every subject has auxiliary data. Without loss of generality, we assume that the first m_1 subjects in the study have auxiliary data, that is, $\boldsymbol{D}_i^a$ for $i = 1, \dots, m_1$. A more concrete specification for the main data $\boldsymbol{D}_i^u$ and auxiliary data $\boldsymbol{D}_i^a$ will be provided in later examples and Section 3.

To improve the estimation precision from (1), auxiliary data $\boldsymbol{D}_i^a$ can play an important role. Due to potential association between the main data and the auxiliary data, the estimation of main parameters of interest $\boldsymbol{\beta}$ will be affected if the information in the auxiliary data is incorporated. One way to incorporate the auxiliary information is to jointly model the two sets of data. However, the joint modeling approach may be theoretically complicated (sometimes even intractable) and computationally intensive (more discussions in Section 5). Instead, to effectively use the auxiliary data to improve estimation for the main parameters $\boldsymbol{\beta}$, we consider the following combined weighted estimating equations

$$\sum_{i=1}^{m_1} m_1 \hat{p}_i \boldsymbol{g}(\boldsymbol{D}_i^u; \boldsymbol{\beta}) + \sum_{i=m_1+1}^{n} \boldsymbol{g}(\boldsymbol{D}_i^u; \boldsymbol{\beta}) = \boldsymbol{0}, \tag{2}$$

where the weights $\hat{p}_i$ on subjects $i = 1, \dots, m_1$ are non-negative and maximize $\prod_{i=1}^{m_1} p_i$ under the constraints

$$\sum_{i=1}^{m_1} p_i = 1, \sum_{i=1}^{m_1} p_i \boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta}) = \boldsymbol{0}, \tag{3}$$

where the estimating functions $\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta})$ are based on some working model for the auxiliary data $\boldsymbol{D}_i^a$ that satisfies $E\{\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta}^*)\} = \boldsymbol{0}$. The working model has parameters $\boldsymbol{\theta}$ with dimension r and true values $\boldsymbol{\theta}^*$ if the working model for auxiliary data is correctly specified. From (2) and (3), intuitively, we can see that the information from auxiliary data $\boldsymbol{D}_i^a$ is incorporated into the main model by the constructed weights $\hat{p}_i$, making data integration and information borrowing possible. Denote the dimension of $\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta})$ as q, and we require $q > r$, thereby resulting in over-identified estimating functions. Two working examples for auxiliary data and corresponding specification of over-identified functions $\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta})$ are provided below.

Example 1. The typical auxiliary measurement in longitudinal format is oftentimes available in practice and can carry the information that is useful for the main analysis. For example, the repeated measurements on systolic blood pressure are informative for the analysis where the risk of hypertension occurrence is the outcome with primary interest. In general, the auxiliary data $\boldsymbol{D}_i^a$ from the ith subject contains repeatedly measured auxiliary outcomes $\tilde{\boldsymbol{Y}}_i$ with dimension T and some covariates $\tilde{\boldsymbol{X}}_i$. Then, the over-identified estimating functions accounting for the association within repeated measurements can take the form of

$$\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{Z}_i^T \boldsymbol{R}_i^{-1/2} \boldsymbol{V}_1 \boldsymbol{R}_i^{-1/2} \left\{ \tilde{\boldsymbol{Y}}_i - \boldsymbol{\mu}(\tilde{\boldsymbol{X}}_i; \boldsymbol{\theta}) \right\} \\ \dots \\ \boldsymbol{Z}_i^T \boldsymbol{R}_i^{-1/2} \boldsymbol{V}_\tau \boldsymbol{R}_i^{-1/2} \left\{ \tilde{\boldsymbol{Y}}_i - \boldsymbol{\mu}(\tilde{\boldsymbol{X}}_i; \boldsymbol{\theta}) \right\} \end{pmatrix}, \tag{4}$$

where $\boldsymbol{Z}_i = \partial \boldsymbol{\mu}(\tilde{\boldsymbol{X}}_i; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T$; $\boldsymbol{R}_i$ is a diagonal matrix containing variances of each element in $\tilde{\boldsymbol{Y}}_i$; $\boldsymbol{\mu}(\tilde{\boldsymbol{X}}_i; \boldsymbol{\theta})$ are the conditional mean of $\tilde{\boldsymbol{Y}}_i$ indexed by parameters $\boldsymbol{\theta}$. The τ matrices $\boldsymbol{V}_j$ lead to the over-identified estimating function with length $q = r \times \tau > r$, given $\tau \geq 2$. One possible candidate for $\boldsymbol{V}_j$ is the basis matrix as suggested in References 6,27. Another option is to take $\boldsymbol{V}_j$ to be the inverse of a working correlation matrix, in which case (4) is equivalent to stacking multiple GEEs[28] with different working correlation structures. In what follows, we will adopt basis matrices for constructing $\boldsymbol{V}_j$. In practice, we recommend using more distinct and small number of basis matrices to avoid potential co-linearity issue. Detailed specifications are referred to Section 3.1.

Example 2. Sometimes, the secondary outcomes that are highly associated with the disease of main interest is available. For example, the measurement of amyloid plaques is often accessible in addition to Alzheimer's disease outcome.[29] Those measurements may not be in a longitudinal format but could be still informative for the analysis of the disease of primary interest. Suppose for the auxiliary data, we specify a working model with auxiliary outcome $\tilde{Y}_i$ and covariates $\tilde{\boldsymbol{X}}_i$ such that the conditional mean of $\tilde{Y}_i$ given $\tilde{\boldsymbol{X}}_i$ is $\mu(\tilde{\boldsymbol{X}}_i; \boldsymbol{\theta})$. Then, we may consider the over-identified estimating function $\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta})$ with the form

$$\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta}) = \boldsymbol{d}(\tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{Z}}_i; \boldsymbol{\theta}) \left\{ \tilde{Y}_i - \mu(\tilde{\boldsymbol{X}}_i; \boldsymbol{\theta}) \right\}, \tag{5}$$

where $\boldsymbol{d}(\tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{Z}}_i; \boldsymbol{\theta})$ is a user-specified vector function with dimension $q > r$ and $\tilde{\boldsymbol{Z}}_i$ are some additionally available variables. One special case is to set $\boldsymbol{d}(\tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{Z}}_i; \boldsymbol{\theta}) = (\tilde{\boldsymbol{X}}_i^T, \tilde{\boldsymbol{Z}}_i^T)^T$. The variables $\tilde{\boldsymbol{Z}}_i$ may be part of or all the covariates $\boldsymbol{X}_i$ used for the main analysis. The selected redundant variables $\tilde{\boldsymbol{Z}}_i$ should satisfy $E\{\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta}_*)\} = \boldsymbol{0}$ for some parameter values $\boldsymbol{\theta}_*$ (not necessarily to be the true ones $\boldsymbol{\theta}^*$).

Given a specific type of auxiliary data with corresponding estimating functions $\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta})$, the constrained optimization in (3) results in estimated weights $\hat{p}_i$ (and estimated $\hat{\boldsymbol{\theta}}$ as by product) that are used to reweight subjects $i = 1, \ldots, m_1$ in (2). The solution to (2) is our proposed estimator $\hat{\boldsymbol{\beta}}_{EN}$. The following theorem summarizes the asymptotic property of $\hat{\boldsymbol{\beta}}_{EN}$.

Theorem 1. *Under certain regularity conditions Appendix S1, suppose there exist parameter values $\boldsymbol{\theta}_*$ such that $\tilde{E}\{\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta}_*)\} = \boldsymbol{0}$, then we have*

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{EN} - \boldsymbol{\beta}_0) \to N(\boldsymbol{0}, \boldsymbol{V}_{EN}),$$

in distribution, where asymptotic covariance matrix $\boldsymbol{V}_{EN}$ equals $\boldsymbol{\Gamma}^{-1}(\boldsymbol{\Sigma} - \rho\boldsymbol{\Lambda}\boldsymbol{S}\boldsymbol{\Lambda}^T)(\boldsymbol{\Gamma}^T)^{-1}$ with $\rho = \lim_{n\to\infty} m_1/n$, $\boldsymbol{\Gamma} = E\{\partial \boldsymbol{g}(\boldsymbol{D}_i^u; \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}^T\}$, $\boldsymbol{\Sigma} = E\{\boldsymbol{g}^{\otimes 2}(\boldsymbol{D}_i^u; \boldsymbol{\beta}_0)\}$, $\boldsymbol{\Lambda} = \tilde{E}\{\boldsymbol{g}(\boldsymbol{D}_i^u; \boldsymbol{\beta}_0)\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta}_*)^T\}$, and $\boldsymbol{S} = \boldsymbol{S}_{11}^{-1}(\boldsymbol{\theta}_*) - \boldsymbol{S}_{11}^{-1}(\boldsymbol{\theta}_*)\boldsymbol{S}_{12}(\boldsymbol{\theta}_*)\boldsymbol{\Omega}(\boldsymbol{\theta}_*)\boldsymbol{S}_{21}(\boldsymbol{\theta}_*)\boldsymbol{S}_{11}^{-1}(\boldsymbol{\theta}_*)$. Here, $\boldsymbol{\Omega}(\boldsymbol{\theta}_*) = \{\boldsymbol{S}_{21}(\boldsymbol{\theta}_*)\boldsymbol{S}_{11}^{-1}(\boldsymbol{\theta}_*)\boldsymbol{S}_{12}(\boldsymbol{\theta}_*)\}^{-1}$, $\boldsymbol{S}_{11}(\boldsymbol{\theta}_*) = \tilde{E}\{\boldsymbol{h}^{\otimes 2}(\boldsymbol{D}_i^a; \boldsymbol{\theta}_*)\}$, $\boldsymbol{S}_{12}(\boldsymbol{\theta}_*) = \tilde{E}\{\partial \boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta}_*)/\partial \boldsymbol{\theta}^T\}$, and $\boldsymbol{S}_{21}(\boldsymbol{\theta}_*) = \boldsymbol{S}_{12}^T(\boldsymbol{\theta}_*)$. The notation $\boldsymbol{f}^{\otimes 2}$ is $\boldsymbol{f}\boldsymbol{f}^T$, and $\tilde{E}(\boldsymbol{f})$ denotes the limit value of $(1/m_1)\sum_{i=1}^{m_1}\boldsymbol{f}_i$, for any measurable function vector $\boldsymbol{f}$.

The proof is presented in section 1.1 in Appendix S1. The regularity conditions are similar to those for the empirical likelihood[19] and the method of moments.[26] In $\boldsymbol{V}_{EN}$, the term $\tilde{\boldsymbol{V}} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}(\boldsymbol{\Gamma}^{-1})^T$ is the asymptotic covariance matrix for the estimator $\hat{\boldsymbol{\beta}}$ solving the equations in (1).[26] Since the matrix $\boldsymbol{S}$ is nonnegative definite, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_{EN}$ is no larger than that of $\hat{\boldsymbol{\beta}}$. In addition, the term ρ quantifies the degree of information borrowing from the auxiliary data, with the efficiency of $\hat{\boldsymbol{\beta}}_{EN}$ reaches the maximum when auxiliary data are available for all subjects, that is, $m_1 = n$.

Efficiency improvement of $\hat{\boldsymbol{\beta}}_{EN}$ over $\hat{\boldsymbol{\beta}}$ requires $q > r$; that is, the length of the estimating function $\boldsymbol{h}(\cdot)$ should be larger than the number of paramemters $\boldsymbol{\theta}$. If $q = r$ and the matrix $\boldsymbol{S}_{12}(\boldsymbol{\theta}_*)$ has full rank, then

$$\boldsymbol{\Omega}(\boldsymbol{\theta}_*) = \left\{\boldsymbol{S}_{21}(\boldsymbol{\theta}_*)\boldsymbol{S}_{11}^{-1}(\boldsymbol{\theta}_*)\boldsymbol{S}_{21}(\boldsymbol{\theta}_*)\right\}^{-1} = \boldsymbol{S}_{21}^{-1}(\boldsymbol{\theta}_*)\boldsymbol{S}_{11}(\boldsymbol{\theta}_*)\boldsymbol{S}_{21}^{-1}(\boldsymbol{\theta}_*),$$

which implies $\boldsymbol{S} = \boldsymbol{S}_{11}^{-1}(\boldsymbol{\theta}_*) - \boldsymbol{S}_{11}^{-1}(\boldsymbol{\theta}_*) = \boldsymbol{0}$ and thus $\boldsymbol{\Lambda}\boldsymbol{S}\boldsymbol{\Lambda}^T = \boldsymbol{0}$. In this case, all information in the auxiliary data has been used to estimate nuisance parameters $\boldsymbol{\theta}$, and thereby the auxiliary data no longer improves over the main analysis. Indeed, such a situation would cause the weights $\hat{p}_i$ equal to $1/n$ for all i, being totally noninformative to the main analysis. Moreover, the quantity $\boldsymbol{\Lambda} = E[\boldsymbol{g}(\boldsymbol{D}_i^u; \boldsymbol{\beta}_0)\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta}^*)^T]$ typically describes the strength of association between the main outcome in $\boldsymbol{D}_i^u$ and auxiliary outcome in $\boldsymbol{D}_i^a$. Higher association implies larger potential of efficiency gain, thus highlighting the motivation of this article to incorporate auxiliary outcome that are highly associated with the trait of main interest into the estimation procedure.

In theory, for a particular specified working model, we only require the existence of $\boldsymbol{\theta}_*$ such that $\tilde{E}\{\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta}_*)\} = \boldsymbol{0}$. Thus, the working model for the auxiliary data does not have to be correctly specified for efficiency improvement for the main parameter estimation. Note that the misspecification of the working model will occur if the mean structure fails to incorporate appropriate covariates and/or the working model fails to address the issue of informative missingness in auxiliary data. In any case, however, our proposed weighting scheme always enables unbiased estimation of main parameters $\boldsymbol{\beta}$, as long as $\tilde{E}\{\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta}_*)\} = \boldsymbol{0}$ hold with some values $\boldsymbol{\theta}_*$. This benefit can be seen from our numerical evaluations (Section 3). It is also worthwhile pointing out that, since we do not intend to make any inference on the auxiliary data, the working model does not need to be practically interpretable. It only serves as a bridge to deliver the information from the auxiliary data to the main analysis.

To illustrate the efficiency gain of our method in using the auxiliary data, we consider the following set of estimating functions

$$\mathcal{S} = \left\{ \boldsymbol{G}_n^*(\boldsymbol{\beta}) \middle| \boldsymbol{G}_n^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{D}_i^u; \boldsymbol{\beta}) - \frac{1}{n}\sum_{i=1}^{m_1} \boldsymbol{C}\boldsymbol{h}(\boldsymbol{D}_i^a; \hat{\boldsymbol{\theta}}) \right\}, \tag{6}$$

where $\boldsymbol{C}$ is any $p \times r$ constant matrix, and $\hat{\boldsymbol{\theta}}$ is the empirical likelihood estimator under the constraints in (3). When the weight matrix $\boldsymbol{C}$ is set to be zero, the class reduces to the estimating function in (1). Thus, the constructed

class contains the unweighted estimation function as a special case. Furthermore, (6) augments the unweighted estimation function based on the main data by using a weighted estimating function based on the auxiliary data, with an arbitrary weight. The following property summarizes the optimality of our proposed estimator within this class.

Property 1. The asymptotic covariance matrix of the proposed estimator $\hat{\boldsymbol{\beta}}_{EN}$ is the minimum that can be achieved by any estimator solving $\boldsymbol{G}_n^*(\boldsymbol{\beta}) = \mathbf{0}$ for $\boldsymbol{G}_n^*(\boldsymbol{\beta}) \in \mathcal{S}$.

The proof is given in section 1.2 in Appendix S1. Note that Property 1 is for a given working function $\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta})$. In general, it is rather challenging to determine the optimal form of $\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta})$ that leads to the most efficiency improvement for the estimation of $\boldsymbol{\beta}$. Thus, a criterion assessing several candidate working models is highly desired. Noting that ρ and $\boldsymbol{\Lambda S \Lambda}^T$ play essential roles in the efficiency improvement based on Theorem 1, the quantity trace$\{\rho \boldsymbol{\Gamma}^{-1} \boldsymbol{\Lambda S \Lambda}^T (\boldsymbol{\Gamma}^T)^{-1}\}$ may serve as such a criterion. It is equivalent to the sum of eigenvalues for the matrix $\rho \boldsymbol{\Gamma}^{-1} \boldsymbol{\Lambda S \Lambda}^T (\boldsymbol{\Gamma}^T)^{-1}$, providing an overall evaluations for the efficiency gain. Therefore, we propose the following index for information borrowing (IIB), which is a scaled version of this quantity,

$$\text{IIB} = \text{trace}\{\rho \text{Diag}(\hat{\tilde{\boldsymbol{V}}})^{-\frac{1}{2}} \hat{\boldsymbol{\Gamma}}^{-1} \hat{\boldsymbol{\Lambda}} \hat{\boldsymbol{S}} \hat{\boldsymbol{\Lambda}}^T (\hat{\boldsymbol{\Gamma}}^T)^{-1} \text{Diag}(\hat{\tilde{\boldsymbol{V}}})^{-\frac{1}{2}}\}, \tag{7}$$

where the terms $\hat{\tilde{\boldsymbol{V}}}$, $\hat{\boldsymbol{\Gamma}}$, $\hat{\boldsymbol{\Lambda}}$, and $\hat{\boldsymbol{S}}$ are some consistent estimators for $\tilde{\boldsymbol{V}}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{S}$, respectively. The multiplier $\text{Diag}(\hat{\tilde{\boldsymbol{V}}})^{-1/2}$ is used to adjust the asymptotic standard error, thereby making it comparable to the one from unweighted estimator based on the equations in (1). The IIB is nonnegative, and a larger value may indicate a better performance in borrowing information by incorporating the auxiliary data.

3 | SIMULATION

3.1 | Example 1

In this section, we will examine the performance of our proposed estimator in the case specified in Example 1, where auxiliary outcomes are repeated measurements. We consider the setting where the outcome in the main analysis is binary but the longitudinal outcomes in the auxiliary data are continuous. This setup considers different types of outcomes in the main and auxiliary data and mimics the situation in ARIC study, where occurrence of hypertension (binary-scale) is our main interest with systolic blood pressure (continuous-scale) as an auxiliary outcome. Specifically, for $i = 1, \ldots, m_1$, the auxiliary data $\boldsymbol{D}_i^a$ contain continuous repeated outcomes $\tilde{\boldsymbol{Y}}_i$ with dimension $T = 4$ generated by the model $\tilde{\boldsymbol{Y}}_i = \tilde{\boldsymbol{X}}_i \boldsymbol{\theta} + \tilde{\boldsymbol{\epsilon}}_i$ with $\boldsymbol{\theta} = (-1, 1, 2, 1)^T$ and covariates $\tilde{\boldsymbol{X}}_i = (\mathbf{1}, \tilde{\boldsymbol{X}}_{i1}, \tilde{\boldsymbol{X}}_{i2}, \tilde{\boldsymbol{X}}_{i3})$ with $\tilde{\boldsymbol{X}}_{ij} = (\tilde{X}_{ij1}, \ldots, \tilde{X}_{ijT})^T$. Here, time-dependent covariate vector $\tilde{\boldsymbol{X}}_{i1}$ follows multivariate normal distribution with mean zero, variance one, and exchangeable correlation matrix with correlation coefficient 0.3; time-dependent covariate vector $\tilde{\boldsymbol{X}}_{i2}$ follows multivariate Bernoulli distribution with success probability 0.5 and exchangeable correlation matrix with correlation coefficient 0.3; $\tilde{\boldsymbol{X}}_{i3}$ is a time-independent covariate vector where all components are equal and are from Bernoulli distribution with success probability 0.5. The residual vector $\tilde{\boldsymbol{\epsilon}}_i = (\tilde{\epsilon}_{i1}, \ldots, \tilde{\epsilon}_{iT})^T$ follows multivariate normal distribution with mean zero, variance one, and the exchangeable correlation structure with correlation coefficient $\tilde{\rho} = 0.4, 0.8$ in this simulation.

In the main analysis, the outcome of interest is binary with success probability $p_i^\star = \{1 + \exp(-\boldsymbol{X}_i^T \boldsymbol{\beta})\}^{-1}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T = (1, -1, -1, 1)^T$ and $\boldsymbol{X}_i = (1, \tilde{X}_{i11}, \tilde{X}_{i21}, \tilde{X}_{i31})^T$. In other words, the covariates used in the main analysis are the baseline covariates in the auxiliary data. The association between the binary outcome of primary interest and the continuous auxiliary outcome in the auxiliary data is generated as follows. For each i, generate a standard normal random variable $\overline{Z}_i$ and then a variable $\overline{x}_i = \tilde{\epsilon}_{i1} r_0 + \overline{Z}_i (1 - r_0^2)^{0.5}$ with $0 \le r_0 \le 1$. It is clear that $\overline{x}_i$ follows the standard normal distribution. Then the binary outcome of main interest Y_i is equal to 1 if $\overline{x}_i \ge x_{0i}$ and equal to 0 otherwise, where x_{0i} is the $(1 - p_i^\star)^{\text{th}}$ percentile of the standard Normal distribution. The constant r_0 controls the strength of association between the binary Y_i and the continuous $\tilde{\boldsymbol{Y}}_i$, with larger values indicating a stronger association. In our simulations, we consider two possible values $r_0 = 0.5$ and $r_0 = 0.9$, leading to correlation coefficient between Y_i and $\tilde{\boldsymbol{Y}}_{i1}$ equal to 0.34 and 0.63, respectively.

We consider three different scenarios and specify $\boldsymbol{h}(\boldsymbol{D}_i^a; \boldsymbol{\theta})$ in (4) for the auxiliary data by employing basis matrices for different correlation structures. In scenario 1, we select the basis matrices $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ corresponding to an exchangeable correlation structure, where $\boldsymbol{V}_1$ is the identity matrix, and $\boldsymbol{V}_2$ is the matrix with 0 on the diagonal and 1 off the

diagonal. In scenario 2, in addition to $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ as above, we also use $\boldsymbol{V}_3$ and $\boldsymbol{V}_4$ corresponding to the basis matrices for an AR-1 correlation structure, where $\boldsymbol{V}_3$ is a matrix with 1 on the two main off-diagonal and 0 elsewhere, and $\boldsymbol{V}_4$ is a matrix with 1 at the left-top and right-bottom corners and 0 elsewhere. This is to study if adding more basis matrices will further improve estimation efficiency for main parameters $\boldsymbol{\beta}$. In scenario 3, we evaluate the performance under a misspecified working model for $\boldsymbol{\mu}_i(\tilde{\boldsymbol{X}}_i;\boldsymbol{\theta})$ with covariates $\tilde{\boldsymbol{X}}_i$ replaced by $\check{\boldsymbol{X}}_i = (\mathbf{1}, \tilde{\boldsymbol{X}}_{i1}, \tilde{\boldsymbol{X}}_{i2})$. The basis matrices are the same as in scenario 2. We consider settings where 50% and 100% of the subjects have auxiliary data for sample sizes $n = 300, 600$. We are not aware of an existing method that is directly applicable to our simulation setting, so we compare our proposed estimator to the main-data-only estimator. More discussions about the potential comparison to existing methods are provided in Section 5. All simulation results are summarized based on 1000 Monte Carlo runs in Table 1 ($\rho = 100\%$) with the results of partially observed auxiliary data ($\rho = 50\%$) provided in Table S2 in Appendix S1.

Overall, the proposed estimator has a good performance evidenced by the small bias and greater-than-one ERE, which is the ratio between the empirical variance of the maximum likelihood estimator using the main study data alone and the empirical variance of the proposed estimator. The empirical coverage probabilities of the 95% confidence intervals are all close to the nominal level. There is more efficiency improvement as more auxiliary data become available, that is, as ρ increases. For all scenarios, as expected, the efficiency improvement increases as the correlation coefficient between the repeated measurements or between the outcome of interest and the repeated measurements increases, reflected by the increased ERE. Also from the ERE, for any given $\tilde{\rho}$ and r_0, efficiency improvement in scenario 2 is always higher than that in scenario 1, showing that adding additional valid estimating equations in $\boldsymbol{h}(\boldsymbol{D}_i^a;\boldsymbol{\theta})$ further improves the efficiency. This is to be expected, since with a fixed parameter $\boldsymbol{\theta}$, more estimating equations lead to higher efficiency from the estimating equation theory.[26] In scenario 3 where the working model is misspecified, the proposed estimator has little bias and still has a better efficiency compared to the maximum likelihood estimator based on the main study data alone, showing the flexibility and robustness of our proposed method. We also observe that, for all the settings considered, the estimates corresponding to $\tilde{X}_{i11}$ and $\tilde{X}_{i21}$, the baseline value of the time-dependent covariates $\tilde{\boldsymbol{X}}_{i1}$ and $\tilde{\boldsymbol{X}}_{i2}$, have more efficiency gain compared to the estimates corresponding to $\tilde{X}_{i31}$, the baseline value of the time-independent covariates $\tilde{\boldsymbol{X}}_{i3}$. More discussion regarding this observation is given in section 2 in the Appendix S1. The IIB seems to be a good measure to compare the overall efficiency gain across different scenarios. Note that, under our data generating scheme the correlation coefficient between main outcome and auxiliary outcome can only take values up to around 0.6, therefore the efficiency gains observed in Table 1 are mild. Substantial efficiency gain could be discovered in real studies or from other simulation setups in Section 3.2 and Appendix S1 (Tables S1 and S3).

3.2 | Example 2

In this section, we assess the performance of our proposed estimator under the setting in Example 2, where a surrogate auxiliary outcome (not in a longitudinal format) is available to the main trait. There are five covariates in total generated as following: $\tilde{X}_{i1}$ follows the Uniform distribution on $[0, 1]$; $\tilde{Z}_{i1}$ follows Bernoulli distribution with success probability 0.5; $(\tilde{X}_{i2}, \tilde{Z}_{i2})^T$ follows the bivariate Normal distribution with mean zero, variance one, and correlation coefficient 0.3; $\tilde{X}_{i3}$ is the sum of $\tilde{Z}_{i2}$ and a standard normal random variable with a standardization such that $\tilde{X}_{i3}$ has mean zero and variance one. With these covariates, the outcome of primary interest is Y_i that follows Bernoulli distribution with success probability $p_i^\star = \{1 + \exp(-\boldsymbol{X}_i^T\boldsymbol{\beta})\}^{-1}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T = (1, -1, -1, 1)^T$ and $\boldsymbol{X}_i = (1, \tilde{X}_{i2}, \tilde{Z}_{i1}, \tilde{X}_{i3})^T$. The auxiliary outcome is modeled as $\tilde{Y}_i = \tilde{\boldsymbol{X}}_i^T\boldsymbol{\theta} + \tilde{\epsilon}_i$, with $\tilde{\boldsymbol{X}}_i = (1, \tilde{X}_{i1}, \tilde{X}_{i2})^T$ and $\boldsymbol{\theta} = (-1, 1, 1)^T$. Here, to generate a correlation between Y_i and $\tilde{Y}_i$, we set $\tilde{\epsilon}_i = r_0 Z_{0i} + (1 - r_0^2)^{0.5}\epsilon_i$, where ϵ_i follows standard Normal distribution, and Z_{0i} is Y_i with a standardization to have mean zero and unit variance. The residual $\tilde{\epsilon}_i$ then has mean zero and unit variance as well and is associated with Y_i, where the degree of association is controlled by $0 \le r_0 \le 1$. In this setup, we consider three values for r_0, $0.5, 0.7, 0.9$, and adopt (5) as our working estimating functions with redundant variables $\tilde{\boldsymbol{Z}}_i = (\tilde{Z}_{i1}, \tilde{Z}_{i2})^T$.

We consider two scenarios, with correctly specified and misspecified working models for the auxiliary outcome, respectively. In the latter scenario, $\check{\boldsymbol{X}}_i = (1, \tilde{X}_{i2})^T$ instead of $\tilde{\boldsymbol{X}}_i$ is used to construct the working model $\mu_i(\check{\boldsymbol{X}}_i;\boldsymbol{\theta})$ for the auxiliary outcome $\tilde{Y}_i$. The simulation results based on 1000 Monte Carlo runs are summarized in Table 2. Similar to Section 3.1, our proposed estimator for the main parameters $\boldsymbol{\beta}$ shows a better performance compared to logistic regression without integrating the auxiliary data. It has smaller standard errors and is robust against misspecification of the working model. More discussions about the efficiency gain are referred to section 3.1 in Appendix S1.

TABLE 1 Simulation results for Example 1. All subjects have auxiliary data ($\rho = 100\%$)

			$n = 300$						$n = 600$					
			Bias	**ESE**	**ASE**	**ERE**	**95%CP**	**IIB**	**Bias**	**ESE**	**ASE**	**ERE**	**95%CP**	**IIB**
S2	$\tilde{\rho} = 0.4$	β_0	2	24	24	1.01	94.8	0.222	1	17	17	1.02	94.5	0.167
	$r_0 = 0.5$	β_1	−2	28	28	1.04	95.8		−2	19	20	1.03	96.5	
		β_2	−3	17	16	1.01	94.2		−1	12	11	1.02	95.0	
		β_3	1	29	29	1.01	94.2		0	21	20	1.03	94.0	
	$\tilde{\rho} = 0.4$	β_0	2	24	24	1.04	95.3	0.479	1	17	17	1.08	95.4	0.420
	$r_0 = 0.9$	β_1	−2	27	27	1.14	95.1		−1	19	19	1.12	95.6	
		β_2	−3	17	16	1.07	93.5		−1	11	11	1.10	93.8	
		β_3	1	28	28	1.06	95.0		0	20	20	1.05	94.9	
	$\tilde{\rho} = 0.8$	β_0	2	24	24	1.00	95.2	0.181	1	17	17	1.04	95.8	0.198
	$r_0 = 0.5$	β_1	−2	27	28	1.06	95.7		−2	19	20	1.06	95.9	
		β_2	−2	16	16	1.05	94.7		−1	12	11	1.07	94.6	
		β_3	1	29	29	1.00	94.3		0	21	20	1.01	93.0	
	$\tilde{\rho} = 0.8$	β_0	2	25	24	1.03	94.5	0.582	2	17	17	1.08	94.4	0.523
	$r_0 = 0.9$	β_1	−3	26	26	1.24	94.8		−1	18	18	1.23	95.3	
		β_2	−3	16	15	1.18	93.9		−1	11	11	1.18	95.6	
		β_3	2	29	29	1.00	94.8		0	20	20	1.01	94.3	
S1	$\tilde{\rho} = 0.4$	β_0	2	25	25	1.00	95	0.042	1	18	17	1.00	94.8	0.042
	$r_0 = 0.5$	β_1	−2	29	29	1.00	95.5		−2	20	20	1.01	96.5	
		β_2	−2	17	17	1.01	95.1		−1	12	12	1.01	95.4	
		β_3	1	30	29	0.99	94.8		0	21	20	1.00	94.1	
	$\tilde{\rho} = 0.4$	β_0	2	25	25	1.00	95.5	0.124	2	17	18	1.02	95.5	0.105
	$r_0 = 0.9$	β_1	−2	28	29	1.03	95.6		−1	20	21	1.05	95.5	
		β_2	−3	17	16	1.03	94.6		−1	12	12	1.05	94.3	
		β_3	1	29	29	0.99	95.5		0	20	21	1.00	95.9	
	$\tilde{\rho} = 0.8$	β_0	2	24	25	1.00	95.2	0.114	1	17	17	1.02	96.1	0.099
	$r_0 = 0.5$	β_1	−2	28	29	1.03	95.5		−2	20	20	1.04	96.4	
		β_2	−2	16	16	1.04	94.7		−1	12	12	1.06	95	
		β_3	1	29	29	0.99	94.4		0	21	20	1.00	93.9	
	$\tilde{\rho} = 0.8$	β_0	2	25	24	1.03	95.9	0.312	2	17	17	1.03	94.8	0.294
	$r_0 = 0.9$	β_1	−2	28	27	1.14	95.2		−1	19	19	1.13	95.3	
		β_2	−3	16	16	1.13	94.1		−1	11	11	1.13	95.4	
		β_3	1	29	29	0.99	95.1		0	20	20	1.00	94.5	
S3	$\tilde{\rho} = 0.4$	β_0	2	24	24	1.00	94.7	0.16	1	17	17	1.01	95.1	0.120
	$r_0 = 0.5$	β_1	−2	28	28	1.03	95.9		−2	20	20	1.02	96.1	
		β_2	−3	17	16	1.00	94.8		−1	12	12	1.02	95	
		β_3	1	30	29	0.99	94.2		0	21	20	1.00	93.8	
	$\tilde{\rho} = 0.4$	β_0	2	24	24	1.03	95.8	0.332	1	17	17	1.04	95.5	0.285
	$r_0 = 0.9$	β_1	−2	27	27	1.13	95.2		−1	19	19	1.10	95.9	
		β_2	−3	17	16	1.07	94.3		−1	11	11	1.08	94.1	
		β_3	1	30	29	0.98	94.9		0	20	20	0.99	95.6	
	$\tilde{\rho} = 0.8$	β_0	2	24	24	1.00	95	0.200	1	17	17	1.02	95.3	0.154
	$r_0 = 0.5$	β_1	−2	27	28	1.05	95.9		−2	20	20	1.04	96.3	
		β_2	−2	16	16	1.03	94.4		−1	12	11	1.05	95	
		β_3	1	29	29	0.99	94.6		0	21	20	1.01	93.3	
	$\tilde{\rho} = 0.8$	β_0	2	25	24	1.03	95	0.444	2	17	17	1.06	94.6	0.399
	$r_0 = 0.9$	β_1	−3	27	27	1.20	95.4		−1	18	19	1.19	95.2	
		β_2	−3	16	16	1.13	94.3		−1	11	11	1.14	94.5	
		β_3	2	29	29	0.97	94.9		0	20	20	0.99	94.5	

Abbreviations: ASE, estimated asymptotic standard error; CP, coverage probability; ERE, empirical relative efficiency, the empirical variance of the maximum likelihood estimator using the main study data alone divided by the empirical variance of the proposed estimator; ESE, empirical standard error; S1, S2, and S3, scenarios 1, 2, 3.

Notes: All values except ERE and IIB are multiplied by 100.

TABLE 2 Simulation results for Example 2

			$n = 300$						$n = 600$					
			Bias	**ESE**	**ASE**	**ERE**	**95%CP**	**IIB**	**Bias**	**ESE**	**ASE**	**ERE**	**95%CP**	**IIB**
S1	prop = 100%	β_1	1	20	20	1.10	94.6	0.514	1	14	14	1.18	94.2	0.487
	$r_0 = 0.5$	β_2	−3	18	17	1.01	93.7		−1	12	12	1.00	95.0	
		β_3	−1	26	25	1.30	93.9		−1	18	17	1.38	95.0	
		β_4	2	17	16	1.11	93.6		1	11	11	1.16	94.7	
	prop = 100%	β_1	1	19	18	1.24	94.0	0.975	1	13	13	1.37	94.7	0.947
	$r_0 = 0.7$	β_2	−3	17	16	1.03	93.8		−1	12	12	1.02	95.0	
		β_3	−1	21	21	1.85	93.4		−1	15	15	2.03	96.0	
		β_4	2	16	15	1.25	92.9		1	10	10	1.34	95.2	
	prop = 100%	β_1	1	17	16	1.51	94.4	1.591	1	12	12	1.70	94.7	1.564
	$r_0 = 0.9$	β_2	−3	17	16	1.05	93.5		−1	12	11	1.04	95.1	
		β_3	−1	14	14	4.21	94.2		−1	9	10	4.95	95.8	
		β_4	2	14	13	1.53	93.6		1	9	9	1.63	96.0	
	prop = 50%	β_1	1	21	20	1.04	94.1	0.275	1	15	14	1.08	94.7	0.248
	$r_0 = 0.5$	β_2	−3	18	17	0.99	93.8		−1	12	12	0.99	95.0	
		β_3	−1	27	27	1.13	93.9		−1	19	19	1.14	94.4	
		β_4	2	17	16	1.04	92.7		1	11	11	1.07	94.4	
	prop = 50%	β_1	1	20	20	1.09	94.2	0.506	1	14	14	1.15	94.8	0.476
	$r_0 = 0.7$	β_2	−3	18	17	1.00	93.8		−1	12	12	1.00	94.8	
		β_3	−1	26	25	1.29	94.3		−1	18	17	1.33	94.4	
		β_4	2	17	16	1.09	92.9		1	11	11	1.14	94.6	
	prop = 50%	β_1	1	20	19	1.19	94.2	0.814	1	13	13	1.27	94.9	0.783
	$r_0 = 0.9$	β_2	−3	18	16	1.00	94.0		−1	12	12	1.01	94.7	
		β_3	−1	23	22	1.59	94.7		0	16	16	1.70	94.4	
		β_4	2	16	15	1.20	93.6		1	11	11	1.24	95.0	
S2	prop = 100%	β_1	1	20	20	1.10	94.2	0.477	1	14	14	1.16	94.2	0.451
	$r_0 = 0.5$	β_2	−3	18	17	1.01	93.8		−1	12	12	1.00	95.0	
		β_3	−1	26	25	1.27	93.5		−1	18	18	1.33	94.6	
		β_4	2	17	16	1.08	93.5		1	11	11	1.15	94.8	
	prop = 100%	β_1	1	19	18	1.23	93.8	0.906	−1	13	13	1.33	94.6	0.877
	$r_0 = 0.7$	β_2	−3	18	16	1.02	93.7		−1	12	12	1.02	95.1	
		β_3	−1	22	21	1.75	94.0		1	15	15	1.87	95.2	
		β_4	2	16	15	1.21	93.2		1	10	10	1.31	95.1	
	prop = 100%	β_1	1	18	17	1.47	94.3	1.477	−1	12	12	1.61	94.7	1.448
	$r_0 = 0.9$	β_2	−3	17	16	1.04	93.5		−1	12	11	1.04	95.2	
		β_3	−1	15	15	3.52	94.7		1	11	11	3.80	95.7	
		β_4	2	15	14	1.44	92.6		1	9	10	1.57	95.7	
	prop = 50%	β_1	1	21	20	1.03	94.1	0.260	1	15	14	1.07	94.8	0.232
	$r_0 = 0.5$	β_2	−3	18	17	0.99	93.5		−1	12	12	1.00	95.0	
		β_3	−1	28	27	1.11	93.3		−1	19	19	1.14	94.2	
		β_4	2	18	16	1.03	92.9		1	11	11	1.07	94.7	
	prop = 50%	β_1	1	20	20	1.09	94.1	0.476	1	14	14	1.14	94.8	0.444
	$r_0 = 0.7$	β_2	−3	18	17	1.00	93.6		−1	12	12	1.01	94.9	
		β_3	−1	26	25	1.26	93.7		−1	18	18	1.31	94.4	
		β_4	2	17	16	1.08	93.0		1	11	11	1.13	94.9	
	prop = 50%	β_1	1	20	19	1.18	94.3	0.763	1	14	13	1.23	94.6	0.728
	$r_0 = 0.9$	β_2	−3	18	16	1.00	94.1		−1	12	12	1.01	94.7	
		β_3	−1	24	23	1.53	93.5		−1	16	16	1.62	94.6	
		β_4	2	16	15	1.17	93.5		1	11	11	1.23	95.0	

Abbreviations: ASE, estimated asymptotic standard error; CP, coverage probability; ERE, empirical relative efficiency, the empirical variance of the maximum likelihood estimator using the main study data alone divided by the empirical variance of the proposed estimator; ESE, empirical standard error; S1 and S2, scenarios 1 and 2, with correctly specified and misspecified working model, respectively.

Notes: All values except ERE and IIB are multiplied by 100.

TABLE 3 Simulation results for auxiliary data with informative missingness

		$n = 300$						$n = 600$					
		Bias	**ESE**	**ASE**	**ERE**	**95%CP**	**IIB**	**Bias**	**ESE**	**ASE**	**ERE**	**95%CP**	**IIB**
$r_0 = 0.5$	β_1	3	20	20	0.99	96	0.202	0	15	14	1.02	95	0.185
	β_2	−2	17	17	0.98	95		−1	12	12	1.01	94	
	β_3	−2	27	27	1.03	94		0	20	19	1.06	94	
	β_4	2	17	16	1.05	95		1	12	12	1.05	95	
$r_0 = 0.7$	β_1	3	20	20	1.01	96	0.367	0	15	14	1.06	95	0.349
	β_2	−2	17	17	0.99	95		−1	12	12	1.01	94	
	β_3	−2	26	26	1.11	94		0	19	18	1.15	94	
	β_4	2	16	16	1.10	96		1	11	11	1.10	95	
$r_0 = 0.9$	β_1	3	20	19	1.06	95	0.587	0	14	14	1.11	94	0.568
	β_2	−2	17	17	0.99	95		−1	12	12	1.02	94	
	β_3	−2	25	24	1.25	95		−1	18	17	1.31	94	
	β_4	2	16	16	1.17	95		1	11	11	1.17	95	

Abbreviations: ASE, estimated asymptotic standard error; CP, coverage probability; ERE, empirical relative efficiency, the empirical variance of the maximum likelihood estimator using the main study data alone divided by the empirical variance of the proposed estimator; ESE, empirical standard error.
Note: All values except ERE and IIB are multiplied by 100.

3.3 | Informative missing in auxiliary data

Section 3.1 and 3.2 evaluate the performance when the mean structure of auxiliary outcomes is misspecified. Not limited to this, our weighting scheme is also robust to other misspecification of the working model, such as informative missingness in auxiliary data (Section 2). Note that, though previous simulation setups have already considered partially observed auxiliary data, we implicitly assume that the unobserved auxiliary data are missing completely at random. Thus, this section provides more evaluation to the case in presence of nonignorable missing auxiliary data.

For illustration, we take the setup from Example 2 and consider the situation where the mean structure is misspecified, and also the auxiliary outcome is informatively missing. To be specific, we first generate main data with sample size 300 and 600 by following the lines in Section 3.2. Then we simulate complete auxiliary data, where the auxiliary outcome is modeled as $\tilde{Y}_i = \tilde{\boldsymbol{X}}_i^T \boldsymbol{\theta} + \tilde{\alpha}_i + \tilde{\epsilon}_i$. Here, $\tilde{\alpha}_i$ follows standard normal distribution, and $\tilde{\boldsymbol{X}}_i$, $\tilde{\epsilon}_i$, and $\boldsymbol{\theta}$ are defined the same in Section 3.2. We then simulate observing indicator $\tilde{R}_i$, so that $\tilde{R}_i = 1$ to keep subject i's auxiliary data and $\tilde{R}_i = 0$, otherwise. The success probability for $\tilde{R}_i$ is given as $\{1 + \exp(-\boldsymbol{H}_i^T \tilde{\boldsymbol{\theta}})\}^{-1}$, where $\boldsymbol{H}_i = (\check{\boldsymbol{X}}_i^T, \tilde{\alpha}_i)^T$, $\tilde{\boldsymbol{\theta}} = (1, 1, 0.5)^T$, and $\check{\boldsymbol{X}}_i$ are already defined in Section 3.2. By incorporating $\tilde{\alpha}_i$ into both models, $\tilde{R}_i$ and $\tilde{Y}_i$ are not independent given covariates $\tilde{\boldsymbol{X}}_i$, leading to non-ignorable missingness.[30,31] To evaluate our proposed estimation, we only take auxiliary data with $R_i = 1$ to construct the working model, of which the mean structure is misspecified with covariates $\check{\boldsymbol{X}}_i$. Thus, both informative missing data issue and misspecification of the mean structure are present for the auxiliary outcome. Table 3 summarizes the results under 1000 Monte Carlo runs under two sample sizes and three r_0 values (defined in Section 3.2). It is seen that, even under the worst situation with both missing data problem and mean structure misspecification, our weighting scheme leads to little bias from the estimation for main parameters $\boldsymbol{\beta}$ and still improve estimation precision in comparison to the estimation without the auxiliary data.

4 | DATA APPLICATION

Hypertension impacts over one-third of the U.S. adults and is a major risk factor for cardiovascular disease and stroke, one of the top causes of death in the United States and worldwide. The Atherosclerosis Risks in Communities study, beginning in 1987, is a prospective epidemiological study conducted among approximately 16 000 middle-aged adults from four U.S. communities. The objective of the study is to investigate the distribution and causes of atherosclerosis and its clinical outcomes, as well as other cardiovascular risk factors. In this application, we are particularly interested

TABLE 4 Data analysis results for the Atherosclerosis Risks in Communities study based on different auxiliary datasets

	Data F (IIB: 1.09)		Date S (IIB: 1.26)	
	Estimate	**Relative variance**	**Estimate**	**Relative variance**
Intercept	−6.2934	1.08	−7.6346	1.17
BMI	0.0054	1.11	0.0352	1.24
DRNKR	0.1633	1.27	0.2350	1.19
CIGT	0.0361	1.18	0.0019	1.13
AGE	0.1036	1.08	0.1031	1.12
Hemoglobin	−0.0847	1.14	−0.0504	1.19

Abbreviations: BMI, body mass index; CIGT, current cigarette smoking status; DRNKR, current alcohol drinking status; Data F, auxiliary data including all subjects; Data S, auxiliary data including the subjects with four observations; Relative variance, variance of the maximum likelihood estimate divided by that of the proposed estimate.

in detecting baseline risk factors for the development of essential hypertension. Our analysis focuses on the subjects who did not have hypertension at baseline and who were not taking antihypertensive medications within the 2 weeks prior to the baseline. The outcome of primary interest is the occurrence of hypertension (binary-scale) during the follow-up (ie, systolic blood pressure $\geq$ 140 mm Hg or diastolic blood pressure $\geq$ 90 mm Hg). To further narrow down the focus, we select white males sampled from center *B*. The potential risk factors of interest include baseline measurements of body mass index (kg/m^2), current alcohol drinking status (1 = Yes, 0 = No), current cigarette smoking status (1 = Yes, 0 = No), age (years), and hemoglobin (g/dL). By removing a small portion of subjects with missing values, the main dataset we use has 1143 subjects.

There are quite a few variables with repeated measurements that could be considered as a candidate auxiliary outcome, including the systolic blood pressure, the diastolic blood pressure, the glucose, etc. Among them, the systolic blood pressure and the diastolic blood pressure are the most informative because that is how hypertension is defined. Noting that only 2.96% subjects in the data have diastolic blood pressure $\geq$ 90 while 17.6% subjects have systolic blood pressure$\geq$ 140, we take the systolic blood pressure as the auxiliary outcome in this application. We have also computed the IIB for different variables, and using the systolic blood pressure measurements as the auxiliary outcome leads to substantially higher IIB than the others.

The systolic blood pressure is repeatedly measured for four visits in this study. To borrow the auxiliary information, we take the estimating functions in (4) with the basis matrices $\boldsymbol{V}_1$, $\boldsymbol{V}_2$, $\boldsymbol{V}_3$, and $\boldsymbol{V}_4$ used in Section 3. Such a choice of basis matrices is based on the highest IIB compared to other three cases that use $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$, $\boldsymbol{V}_1$ and $\boldsymbol{V}_3$ and $\boldsymbol{V}_4$, and $\boldsymbol{V}_2$ and $\boldsymbol{V}_3$ and $\boldsymbol{V}_4$, respectively. The covariates used in the working model include body mass index, alcohol drinking status, cigarette smoking status, hemoglobin, usage of antihypertensive medication (1 = Yes, 0 = No), and age at each visit. Since some study subjects did not complete four visits, we consider two ways of using the auxiliary data. The first way uses all available repeated measurements from all subjects, and the second uses data only from those subjects with all four repeated measurements. For the second way about ρ =76% of the study subjects contribute auxiliary data. Both ways are valid since we just need to specify a working model for the auxiliary data, and misspecification of the constructed working model will lead to unbiased estimation for the main analysis (referred to Sections 2 and 3).

The results are summarized in Table 4. When using all available repeated measurements from all subjects, a better efficiency is observed for estimates for the drinking and cigarette smoking effects, based on a larger relative variance, while the estimates corresponding to the intercept, BMI, age, and hemoglobin are less efficient, compared to those when using only the subjects with complete four repeated measurements. In addition, using partial the subjects with all four repeated measurements leads to a higher IIB. Thus, when the auxiliary data have missing values, using more subjects with incomplete observations may not necessarily lead to a better estimation. Table 5 summarizes the results when using only the subjects with all four repeated measurements. Compared to the maximum likelihood estimator using the main data alone, the proposed method leads to estimates with smaller variances after incorporating the information from the auxiliary data. The maximum likelihood method only detects the significance of baseline age, whereas our method also detects the marginal significance of baseline body mass index on the development of hypertension. Note that in this application, most subjects have records of the occurrence of hypertension during the study as the main outcome. Thus, taking complete data for main analysis will not be improper. However, our proposed estimation can easily address missing

TABLE 5 Data analysis results for the Atherosclerosis Risks in Communities study using subjects with all four repeated measurements as the auxiliary dataset

	Estimate	SE	Relative variance	Odds ratio	Lower limit	Upper limit	*P* value	*P* value MLE
Intercept	−7.6346	1.4169	1.17	0.000	0.000	0.008	0.000	0.001
BMI	0.0352	0.0200	1.24	1.036	0.996	1.077	0.078	0.387
DRNKR	0.2350	0.1505	1.19	1.265	0.942	1.699	0.118	0.357
CIGT	0.0019	0.1725	1.13	1.002	0.715	1.405	0.991	0.724
AGE	0.1031	0.0134	1.12	1.109	1.080	1.138	0.000	0.000
Hemoglobin	−0.0504	0.0664	1.19	0.951	0.835	1.083	0.448	0.672

Abbreviations: BMI, body mass index; CIGT, current cigarette smoking status; DRNKR, current alcohol drinking status; Lower limit and upper limit, lower bound and upper bound of the 95% confidence interval for the odds ratio; Relative variance, variance of the maximum likelihood estimate divided by that of the proposed estimate; SE, estimated standard error.

data issue occurred in the main analysis. Discussions and extensions are referred to Section 5 and section 3.2 in the Appendix S1.

5 | DISCUSSION

In the big data era, it is of tremendous interest to have methods that can incorporate auxiliary information to enhance statistical analysis. This paper considers the case where auxiliary data are collected from the same study subjects. We proposed an effective estimation procedure to borrow information from the auxiliary data, which can be completely different from the trait of primary interest. The auxiliary information may substantially improve the estimation efficiency in the main analysis. Note that in theory, we can reformulate (2) and (3) by introducing an indicator of the availability of auxiliary data, which leads to an alternative equivalent formulation of the proposed estimation procedure.Refer to section 1.3 in Appendix S1 for more details. In addition, we provided an IIB to assess the performance when comparing different working models for the auxiliary data. The magnitude of efficiency gain by borrowing information from auxiliary data depends on the number of subjects that have auxiliary data, the strength of association between the main outcome and the auxiliary outcome, and the specification of the working model for the auxiliary data, among other things.

Some existing methods may be applied to the setting considered in this article after some nontrivial modifications. One such method is the joint modeling approach, which was originally proposed to incorporate the longitudinal information to survival analysis via shared random effects.[32] In the setting we considered, a shared random effect to link the main outcome and the auxiliary outcome may be used. However, most joint modeling methods are limited to certain parametric forms, which may not be applicable to model the data we generated in Section 3. Moreover, when the main and auxiliary data are jointly modeled, the misspecification of a working model for the auxiliary data can lead to inconsistent estimation in the main model. In future research, we will study how to modify and apply some existing methods to our setting and how they compare to our method.

Many extensions could be easily done based on this paper. One direct extension is to address missing data problem in the main analysis. In presence of high missingness, our method can be easily adjusted by adopting some well-known scheme, such as inverse probability weight,[31,33,34] into the estimating function $\boldsymbol{g}(\boldsymbol{D}_i^u; \boldsymbol{\beta})$. More details about this extension is provided in section 3.2 in Appendix S1. Moreover, a generalization of the proposed method to other data types, such as survival data, is of great interest. Borrowing information from possible multiple auxiliary outcomes is another important research problem. We will investigate such generalizations in our future work.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

AUTHOR CONTRIBUTIONS

All authors have made important contributions and have approved this work.

DATA AVAILABILITY STATEMENT

The ARIC study data is managed by the ARIC study Data Coordinating Center and available on the website. The publicly available data used in this manuscript are available from the authors upon request.

ORCID

Chixiang Chen https://orcid.org/0000-0001-8208-281X

REFERENCES

1. Imbens GW, Lancaster T. Combining micro and macro data in microeconometric models. *Rev Econ Stud*. 1994;61(4):655-680.
2. Chen YH, Chen H. A unified approach to regression analysis under double-sampling designs. *J R Stat Soc Ser B Stat Methodol*. 2000;62(3):449-460.
3. Qin J. Miscellanea. combining parametric and empirical likelihoods. *Biometrika*. 2000;87(2):484-490.
4. Glynn AN, Quinn KM. An introduction to the augmented inverse propensity weighted estimator. *Polit Anal*. 2010;18(1):36-56.
5. Lumley T, Shaw PA, Dai JY. Connections between survey calibration estimators and semiparametric models for incomplete data. *Int Stat Rev*. 2011;79(2):200-220.
6. Tang CY, Leng C. Empirical likelihood and quantile regression in longitudinal data analysis. *Biometrika*. 2011;98(4):1001-1006.
7. Qin J, Zhang H, Li P, Albanes D, Yu K. Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika*. 2015;102(1):169-180.
8. Chatterjee N, Chen YH, Maas P, Carroll RJ. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J Am Stat Assoc*. 2016;111(513):107-117.
9. Cheng W, Taylor JM, Gu T, Tomlins SA, Mukherjee B. Informing a risk prediction model for binary outcomes with external coefficient information. *J R Stat Soc Ser C Appl Stat*. 2019;68(1):121-139.
10. Han P, Lawless JF. Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Stat Sin*. 2019;29:1321-1342.
11. Yang S, Ding P. Combining multiple observational data sources to estimate causal effects. *J Am Stat Assoc*. 2019;115:1-46.
12. Huang CY, Qin J, Tsai HT. Efficient estimation of the Cox model with auxiliary subgroup survival information. *J Am Stat Assoc*. 2016;111(514):787-799.
13. Gu T, Taylor JM, Cheng W, Mukherjee B. Synthetic data method to incorporate external information into a current study. *Can J Stat*. 2019;47(4):580-603.
14. Fried LP, Borhani NO, Enright P, et al. The cardiovascular health study: design and rationale. *Ann Epidemiol*. 1991;1(3):263-276.
15. Newman AB, Sachs MC, Arnold AM, et al. Total and cause-specific mortality in the cardiovascular health study. *J Gerontol A Biol Sci Med Sci*. 2009;64(12):1251-1261.
16. Wilson KM, Kasperzyk JL, Rider JR, et al. Coffee consumption and prostate cancer risk and progression in the health professionals follow-up study. *JNCI*. 2011;103(11):876-884.
17. Investigators A. The atherosclerosis risk in communit (ARIC) study: design and objectwes. *Am J Epidemiol*. 1989;129(4):687-702.
18. González HM, Tarraf W, Harrison K, et al. Midlife cardiovascular health and 20-year cognitive decline: atherosclerosis risk in communities study results. *Alzheimers Dement*. 2018;14(5):579-589.
19. Qin J, Lawless J. Empirical likelihood and general estimating equations. *Ann Stat*. 1994;22(1):300-325.
20. Qin J, Shao J, Zhang B. Efficient and doubly robust imputation for covariate-dependent missing responses. *J Am Stat Assoc*. 2008;103(482):797-810.
21. Zhang B. Empirical likelihood in causal inference. *Econom Rev*. 2016;35(2):201-231.
22. Han P, Wang L. Estimation with missing data: beyond double robustness. *Biometrika*. 2013;100(2):417-430.
23. Han P. Multiply robust estimation in regression analysis with missing data. *J Am Stat Assoc*. 2014;109(507):1159-1173.
24. Han P. Combining inverse probability weighting and multiple imputation to improve robustness of estimation. *Scand Stat Theory Appl*. 2016;43(1):246-260.
25. Chan KCG, Yam SCP, Zhang Z. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *J R Stat Soc Ser B Stat Methodol*. 2016;78(3):673.
26. Newey WK, McFadden D. Large sample estimation and hypothesis testing. Engle RF, McFadden DL, eds. *Handbook of Econometrics*. Vol 4; Elsevier; Amsterdam, Netherlands: 1994:2111-2245.
27. Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika*. 2000;87(4):823-836.
28. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22.
29. Gouras GK, Olsson TT, Hansson O. β-amyloid peptides and amyloid plaques in Alzheimera's disease. *Neurotherapeutics*. 2015;12(1):3-11.
30. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc*. 1995;90(429):106-121.

31. Enders CK. *Applied Missing Data Analysis*. New York City, NY: Guilford Press; 2010.
32. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Stat Sin*. 2004;14(3):809-834.
33. Chen C, Shen B, Zhang L, Xue Y, Wang M. Empirical-likelihood-based criteria for model selection on marginal analysis of longitudinal data with dropout missingness. *Biometrics*. 2019;75(3):950-965.
34. Chen C, Shen B, Liu A, Wu R, Wang M. A multiple robust propensity score method for longitudinal analysis with intermittent missing data. *Biometrics*. 2020;77:519-532.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Chen C, Han P, He F. Improving main analysis by borrowing information from auxiliary data. *Statistics in Medicine*. 2022;41(3):567-579. doi: 10.1002/sim.9252