Received: 20 December 2020 | Revised: 6 October 2021 | Accepted: 8 October 2021

DOI: 10.1111/biom.13585

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Noniterative adjustment to regression estimators with population-based auxiliary information for semiparametric models

Fei Gao[1] | **K. C. G. Chan**[2]

[1] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

[2] Department of Biostatistics, University of Washington, Seattle, Washington, USA (Email: kcgchan@u.washington.edu)

Correspondence
Fei Gao, Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WC, USA.
Email: fgao@fredhutch.org

Funding information
National Science Foundation, Grant/Award Number: DMS1711952; National Institutes of Health, Grant/Award Numbers: R01HL122212, U24AG072122, P30MH123248, S10OD028685

Abstract

Disease registries, surveillance data, and other datasets with extremely large sample sizes become increasingly available in providing population-based information on disease incidence, survival probability, or other important public health characteristics. Such information can be leveraged in studies that collect detailed measurements but with smaller sample sizes. In contrast to recent proposals that formulate additional information as constraints in optimization problems, we develop a general framework to construct simple estimators that update the usual regression estimators with some functionals of data that incorporate the additional information. We consider general settings that incorporate nuisance parameters in the auxiliary information, non-*i.i.d.* data such as those from case-control studies, and semiparametric models with infinite-dimensional parameters common in survival analysis. Details of several important data and sampling settings are provided with numerical examples.

KEYWORDS

case-control studies, empirical likelihood, meta-analysis, proportional hazards model, proportional odds model, semiparametric model, survival data

1 | INTRODUCTION

With the development of population-based biomedical science, large studies and datasets become increasingly available. Examples include census data, disease registries, healthcare databases, and various consortia of individual studies (Chatterjee et al., 2016). Using these large datasets, certain characteristics of the population can be accurately described, and they may serve as auxiliary information to improve estimation efficiency in analyzing data from small studies. As the large datasets are usually designed for purposes different from the hypotheses of interest, they may not be directly utilized as a complementary sample for the original research question. However, the moment conditions or unbiased estimating equations that correspond to the characteristic of the underlying population can be utilized in smaller studies of interest. One example is the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (National Cancer Institute, 2021), which is an authoritative source of information on cancer incidence and survival in the United States. It provides information on the most recent cancer incidence, mortality, survival, prevalence, and lifetime risk statistics that would serve as auxiliary information in cancer clinical trials.

In the absence of auxiliary information, regression methods are routinely used to study the association between exposures and outcomes and many such methods are based on maximizing a likelihood or other objective function. These estimators can also be formulated as

solving first-order (score) equations that are just-identified, that is, the number of equations are the same as the dimension of the model parameters. Auxiliary information in the form of additional estimating functions would therefore lead to a system of overidentified estimating functions, in which efficient estimation can be attained by empirical likelihood (Qin and Lawless, 1994). The empirical likelihood approach has been applied in the survey sampling setting to incorporate auxiliary information on the finite population in the estimation of the population mean, total, distribution function, or quantiles, while adjusting for different sampling schemes (Chen and Qin, 1993; Chen and Sitter, 1999; Qin, 2000; Wu and Sitter, 2001; Chaudhuri et al., 2008; Qin et al., 2015; Chatterjee et al., 2016). Computation for the empirical likelihood estimation usually involves iteratively maximizing the objective function that includes Lagrange multipliers and solving for the value of the Lagrange multipliers, and may cause computational challenges (Han and Wang, 2013). Generalized method of moment (GMM) provides an alternative framework for handling overidentified estimating functions and is recently studied for incorporating population-based auxiliary information (Kundu et al., 2019; Sheng et al., 2020). However, such methods still requires optimization to obtain the estimators.

When model parameters include infinite dimensional components, for example, in semiparametric models, incorporation of auxiliary information becomes more challenging. Zhou (2006) and Hu and Zhou (2010) studied the proportional hazards model with auxiliary information on functionals of the hazard and formulated the empirical likelihood function in terms of the baseline hazard function. Huang et al. (2016) proposed a double empirical likelihood approach to combine the published subgroup t-year survival probabilities in a proportional hazards model for individual-level data. The above methods made use of certain special structure of the proportional hazards model to obtain a closed-form profile estimator of the baseline hazard function. As the closed-form solution for the infinite-dimensional parameter is not always available, the methods cannot be easily generalized to other semiparametric models.

In this paper, we study a general estimation framework in which a simple noniterative update procedure that incorporates auxiliary information can attain the same statistical efficiency as certain maximum empirical likelihood estimators. Unlike empirical likelihood methods that require constraint optimization through introduction of additional Lagrange multipliers, the proposed method avoids their computation entirely by exploiting intricate mathematical structure of the problem. We consider the general setting of semiparametric models, where the constraints corresponding to the auxiliary information can be summarized as estimating equations, and allow additional unknown parameters present only in the constraints. Our formulation also provides a simple asymptotic variance estimator for inference. We demonstrate use of the proposed methods in various common parametric and semiparametric settings with simulations and some real examples.

2 | METHODS

To simplify the derivations, we consider *i.i.d.* sampling in the exposition, and the generalization to non-*i.i.d.* samples will be discussed in Web Appendix A. Let $\boldsymbol{X}_i$ $(i = 1, \ldots, n)$ be *i.i.d.* observations of a random variable $\boldsymbol{X}$, whose distribution is associated with an unknown p-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^p$ and an infinite-dimensional nuisance parameter η. Suppose that the true value (θ_0, η_0) of (θ, η) maximizes a criterion function $E\{m(\boldsymbol{X}; \theta, \eta)\}$, where $E(\cdot)$ is the expectation function with respect to $\boldsymbol{X}$. Then, an initial estimator $(\widehat{\theta}, \widehat{\eta})$ can be constructed by maximizing the function $\sum_{i=1}^n m(\boldsymbol{X}_i; \theta, \eta)$. In particular, the choice of $m(\boldsymbol{X}; \theta, \eta) = \log f(\boldsymbol{X}; \theta, \eta)$ corresponds to the maximum likelihood estimator, where $f(\boldsymbol{X}; \theta, \eta)$ is the density function of $\boldsymbol{X}$.

Suppose that additional information on the distribution of $\boldsymbol{X}$ can be summarized as another set of q-dimensional functions $\boldsymbol{g}(\boldsymbol{X}; \theta, \pi, \eta)$ with $E\{\boldsymbol{g}(\boldsymbol{X}; \theta, \pi, \eta)\} = \boldsymbol{0}$, where π is a υ-dimensional nuisance parameter that is not of primary interest. Here, even though the same notation $\boldsymbol{X}$ is used in the criterion function $m(\boldsymbol{X}; \theta, \eta)$ and the additional information function $\boldsymbol{g}(\boldsymbol{X}; \theta, \pi, \eta)$, they may involve different subsets of $\boldsymbol{X}$. Based on the observed data $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, we obtain another set of estimating equations

$$\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{X}_i; \theta, \pi, \eta) = \boldsymbol{0}. \tag{1}$$

If $\upsilon < q$, then we cannot directly solve (1) for π even if θ and η are given, as the system of equations is overidentified. To obtain an estimator that makes use of the auxiliary information (1) efficiently, we propose a noniterative update procedure. Particularly, the procedure is based on a comparison of the asymptotic distributions of the estimator that incorporates the auxiliary information and an initial estimator that maximizes $\sum_{i=1}^n m(\boldsymbol{X}_i; \theta, \eta)$.

To incorporate the auxiliary information, we consider a composite empirical likelihood approach such that we maximize $\sum_{i=1}^n \{m(\boldsymbol{X}_i; \theta, \eta) + \log p_i\}$ subject to the constraints $\sum_{i=1}^n p_i = 1$, $p_i \geq 0$ for $i = 1, \ldots, n$, and $\sum_{i=1}^n p_i \boldsymbol{g}(\boldsymbol{X}_i; \theta, \pi, \eta) = \boldsymbol{0}$, where p_i is a point mass corresponding to subject i. By applying the Lagrange multiplier

arguments, it can be seen that the estimator maximizes

$$\sum_{i=1}^{n}\left[m(\boldsymbol{X}_i;\theta,\eta)-\log\left\{1+\boldsymbol{t}^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{X}_i;\theta,\boldsymbol{\pi},\eta)\right\}\right],$$

where $\boldsymbol{t}$ is a q-vector of Lagrange multipliers that satisfies

$$\sum_{i=1}^{n}\frac{\boldsymbol{g}(\boldsymbol{X}_i;\theta,\boldsymbol{\pi},\eta)}{1+\boldsymbol{t}^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{X}_i;\theta,\boldsymbol{\pi},\eta)}=\boldsymbol{0}.$$

Remark 1. The objective function we considered is a composite log-likelihood such that we address a combination of initial objective function $m(\boldsymbol{X}_i;\theta,\eta)$ and empirical mass $\log p_i$ that corresponds to the auxiliary information. Here, we directly sum up $m(\boldsymbol{X}_i;\theta,\eta)$ and $\log p_i$, suggesting equal weights for the internal data and auxiliary information. In settings with certain level of belief on auxiliary information, we may place different weights or include an additional parameter, for example, $m(\boldsymbol{X}_i;\theta,\eta)+\alpha\log p_i$, to form a different composite likelihood. The estimation procedure can be revised to address this change. In the special case of a parametric regression model, where the initial objective function corresponds to a conditional probability of outcome variable and the auxiliary information relates only to the marginal distribution of the independent variable, the proposed composite likelihood function is asymptotic equivalent to the empirical likelihood considered in Qin (2000) and is efficient.

Write $\boldsymbol{s}_\theta(\boldsymbol{X};\theta,\eta)$ and $\dot{\boldsymbol{s}}_{\theta\theta}(\boldsymbol{X};\theta,\eta)$ as the first and second derivatives of $m(\boldsymbol{X};\theta,\eta)$ with respect to θ, respectively. In addition, write $\dot{\boldsymbol{g}}_\theta(\boldsymbol{X};\theta,\boldsymbol{\pi},\eta)$ and $\dot{\boldsymbol{g}}_\pi(\boldsymbol{X};\theta,\boldsymbol{\pi},\eta)$ as the derivatives of $\boldsymbol{g}(\boldsymbol{X};\theta,\boldsymbol{\pi},\eta)$ with respect to θ and $\boldsymbol{\pi}$, respectively. Under some regularity conditions, we show in Web Appendix A that after profiling out the infinite-dimensional parameter η, the composite likelihood estimator $\widetilde{\theta}$ then satisfies

$$\begin{aligned}\sqrt{n}(\widetilde{\theta}-\theta_0)=&\left(\boldsymbol{I}_{p\times p},\boldsymbol{0}_{p\times(v+q)}\right)\left\{E\widetilde{\boldsymbol{A}}(\boldsymbol{X};\theta_0,\boldsymbol{\pi}_0,\eta_0)\right\}^{-1}\\&\times\left\{n^{-1/2}\sum_{i=1}^{n}\widetilde{\boldsymbol{l}}(\boldsymbol{X}_i;\theta_0,\boldsymbol{\pi}_0,\eta_0)\right\}+o_{\mathbb{P}}(1),\end{aligned}\tag{2}$$

where

$$\widetilde{\boldsymbol{A}}(\boldsymbol{X};\theta,\boldsymbol{\pi},\eta)=\begin{pmatrix}-\widetilde{\boldsymbol{s}}_{\theta\theta}(\boldsymbol{X};\theta,\eta) & \boldsymbol{0}_{p\times v} & \widetilde{\boldsymbol{g}}_\theta(\boldsymbol{X};\theta,\boldsymbol{\pi},\eta)^{\mathrm{T}}\\ \boldsymbol{0}_{v\times p} & \boldsymbol{0}_{v\times v} & \dot{\boldsymbol{g}}_\pi(\boldsymbol{X};\theta,\boldsymbol{\pi},\eta)^{\mathrm{T}}\\ -\widetilde{\boldsymbol{g}}_\theta(\boldsymbol{X};\theta,\boldsymbol{\pi},\eta) & -\dot{\boldsymbol{g}}_\pi(\boldsymbol{X};\theta,\boldsymbol{\pi},\eta) & \widetilde{\boldsymbol{G}}(\boldsymbol{X};\theta,\boldsymbol{\pi},\eta)\end{pmatrix},$$

and $\widetilde{\boldsymbol{l}}(\boldsymbol{X};\theta,\boldsymbol{\pi},\eta)=(\widetilde{\boldsymbol{s}}_\theta(\boldsymbol{X};\theta,\eta)^{\mathrm{T}},\boldsymbol{0}_{v\times1}^{\mathrm{T}},\widetilde{\boldsymbol{g}}(\boldsymbol{X};\theta,\boldsymbol{\pi},\eta)^{\mathrm{T}})^{\mathrm{T}}$. The functions $\widetilde{\boldsymbol{s}}_{\theta\theta}$, $\widetilde{\boldsymbol{g}}_\theta$, $\widetilde{\boldsymbol{G}}$, $\widetilde{\boldsymbol{s}}_\theta$, and $\widetilde{\boldsymbol{g}}$ are respective versions of $\dot{\boldsymbol{s}}_{\theta\theta}$, $\dot{\boldsymbol{g}}_\theta$, $\boldsymbol{g}^{\otimes2}$, $\boldsymbol{s}_\theta$, and $\boldsymbol{g}$ with η profiled out, as defined in Web Appendix A. The explicit forms are also given in an example in Section 3.1.

For the initial estimator $(\widehat{\theta},\widehat{\eta})$ that maximizes $\sum_{i=1}^{n}m(\boldsymbol{X}_i;\theta,\eta)$, we show in Web Appendix A that

$$\begin{aligned}\sqrt{n}(\widehat{\theta}-\theta_0)=&\left(\boldsymbol{I}_{p\times p}\boldsymbol{0}_{p\times v}-\left\{E\widetilde{\boldsymbol{s}}_{\theta\theta}(\boldsymbol{X};\theta_0,\eta_0)\right\}^{-1}\right.\\&\left.\times\left\{E\widetilde{\boldsymbol{g}}_\theta(\boldsymbol{X};\theta_0,\boldsymbol{\pi}_0,\eta_0)\right\}^{\mathrm{T}}\right)\\&\times\left\{E\widetilde{\boldsymbol{A}}(\boldsymbol{X};\theta_0,\boldsymbol{\pi}_0,\eta_0)\right\}^{-1}\\&\times\left\{n^{-1/2}\sum_{i=1}^{n}\widetilde{\boldsymbol{l}}(\boldsymbol{X}_i;\theta_0,\boldsymbol{\pi}_0,\eta_0)\right\}\\&+o_{\mathbb{P}}(1).\end{aligned}\tag{3}$$

Comparing (2) and (3), we can obtain an asymptotic equivalent version of $\widetilde{\theta}$ by

$$\begin{aligned}\widehat{\theta}+&\left(\boldsymbol{0}_{p\times p}\boldsymbol{0}_{p\times v}\left\{n^{-1}\sum_{i=1}^{n}\widetilde{\boldsymbol{s}}_{\theta\theta}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\eta})\right\}^{-1}\right.\\&\left.\left\{n^{-1}\sum_{i=1}^{n}\widetilde{\boldsymbol{g}}_\theta(\boldsymbol{X}_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})\right\}^{\mathrm{T}}\right)\left\{n^{-1}\sum_{i=1}^{n}\widetilde{\boldsymbol{A}}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})\right\}^{-1}\\&\times\left\{n^{-1}\sum_{i=1}^{n}\widetilde{\boldsymbol{l}}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})\right\},\end{aligned}\tag{4}$$

where $\widehat{\boldsymbol{\pi}}$ is a consistent estimator of $\boldsymbol{\pi}$ that may be given or estimated (see Remark 3). With a slight abuse of notation, we denote (4) by $\widetilde{\theta}$ because they are asymptotically equivalent. We can further show that $\sqrt{n}(\widetilde{\theta}-\theta_0)$ converges weakly to a p-dimensional zero-mean normal random vector with covariance matrix $\boldsymbol{D}\boldsymbol{V}\boldsymbol{D}^{\mathrm{T}}$, where $\boldsymbol{D}=(\boldsymbol{I}_{p\times p},\boldsymbol{0}_{p\times(v+q)})\{E\widetilde{\boldsymbol{A}}(\boldsymbol{X};\theta_0,\boldsymbol{\pi}_0,\eta_0)\}^{-1}$ and $\boldsymbol{V}=E\{\widetilde{\boldsymbol{l}}(\boldsymbol{X};\theta_0,\boldsymbol{\pi}_0,\eta_0)^{\otimes2}\}$. A natural variance estimator for $\widetilde{\theta}$ is

$$\begin{aligned}&\left[\left(\boldsymbol{I}_{p\times p}\ \boldsymbol{0}_{p\times(v+q)}\right)\left\{n^{-1}\sum_{i=1}^{n}\widetilde{\boldsymbol{A}}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})\right\}^{-1}\right]\\&\times\left\{n^{-1}\sum_{i=1}^{n}\widetilde{\boldsymbol{l}}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})^{\otimes2}\right\}\\&\times\left[\left(\boldsymbol{I}_{p\times p}\ \boldsymbol{0}_{p\times(v+q)}\right)\left\{n^{-1}\sum_{i=1}^{n}\widetilde{\boldsymbol{A}}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})\right\}^{-1}\right]^{\mathrm{T}}.\end{aligned}$$

We show in Web Appendix A that when the original estimator is a semiparametric maximum likelihood estimator of a regression model, the asymptotic variance of $\sqrt{n}(\widetilde{\theta}-\theta_0)$ is given by $\{E\widetilde{\boldsymbol{s}}_\theta(\boldsymbol{X};\theta_0,\eta_0)^{\otimes 2}+\boldsymbol{B}\boldsymbol{Q}^{-1}\boldsymbol{B}^{\mathrm{T}}\}^{-1}$, where $\boldsymbol{B}$ and $\boldsymbol{Q}$ are defined in Web Appendix A. The asymptotic variance is no larger than that of $\sqrt{n}(\widehat{\theta}-\theta_0)$, indicating an improvement of efficiency with the incorporation of the auxiliary information.

Remark 2. Our updating formula is substantially different from the one-step efficient estimation procedure that supplies an initial consistent estimator to the Taylor series expansion of an efficient estimating equation. First, $m(\boldsymbol{X};\theta,\eta)$ is not required to be a log-likelihood function. In addition, the estimator (4) is constructed by comparing the asymptotic distributions of the estimators with and without auxiliary information addressed, and exploits a special structure that the right-hand sides of (2) and (3) can be expressed with a difference of only a matrix factor and an asymptotically negligible term. A similar idea is also used in Cox and Wermuth (1990), where a one-step update is constructed by comparing the information from a model and an extended model with more model parameters.

Remark 3. A consistent estimator for π is often available from previous studies as a form of external information and could be treated as fixed, as discussed in detail in Chatterjee et al. (2016). If such information is not available, we can still obtain an initial estimate $\widehat{\pi}$ by solving $\sum_{i=1}^n \boldsymbol{g}_1(\boldsymbol{X}_i;\widehat{\theta},\pi,\widehat{\eta})=\boldsymbol{0}$ where $\boldsymbol{g}_1$ is defined without loss of generality to be the first υ-dimensional component of $\boldsymbol{g}$. Specific examples are given in Sections 3.2 and 3.3. The asymptotic variance of the updating estimator $\widetilde{\theta}$ does not depend on the version of $\boldsymbol{g}_1$ chosen to compute the initial estimator $\widehat{\pi}$, as long as $\boldsymbol{g}_1$ is a υ-dimensional subcomponent of $\boldsymbol{g}$ and $\widehat{\pi}$ is consistent, see Web Appendix A. This is an advantage for using the composite empirical likelihood to handle nuisance parameters appearing only in the auxiliary information. When the estimation and inference of π is of interest, we can modify the procedure to update both $\widehat{\theta}$ and $\widehat{\pi}$. Details are given in Remark 4.

Remark 4. When the estimation and inference on π is also of interest, we may modify the proposed procedure to update the estimation of π along with θ. Let

$$\begin{aligned}\boldsymbol{H}(\theta,\pi,\eta)=&\left\{n^{-1}\sum_{i=1}^n\widetilde{\boldsymbol{g}}_{1\theta}(\boldsymbol{X}_i;\theta,\pi,\eta)\right\}\\&\times\left\{n^{-1}\sum_{i=1}^n\widetilde{\boldsymbol{s}}_{\theta\theta}(\boldsymbol{X}_i;\theta,\eta)\right\}^{-1}\\&\times\left\{n^{-1}\sum_{i=1}^n\widetilde{\boldsymbol{g}}_{\theta}(\boldsymbol{X}_i;\theta,\pi,\eta)\right\}^{\mathrm{T}}\\&-n^{-1}\sum_{i=1}^n\widetilde{\boldsymbol{g}}_1(\boldsymbol{X}_i;\theta,\pi,\eta)\widetilde{\boldsymbol{g}}(\boldsymbol{X}_i;\theta,\pi,\eta)^{\mathrm{T}},\end{aligned}$$

and $\widetilde{\boldsymbol{g}}_{1\theta}(\boldsymbol{X};\theta,\pi,\eta)$ and $\dot{\boldsymbol{g}}_{1\pi}(\boldsymbol{X};\theta,\pi,\eta)$ be the vectors of the first υ elements of $\widetilde{\boldsymbol{g}}_{\theta}(\boldsymbol{X};\theta,\pi,\eta)$ and $\dot{\boldsymbol{g}}_{\pi}(\boldsymbol{X};\theta,\pi,\eta)$, respectively. We can show that the updating estimator $(\widetilde{\theta},\widetilde{\pi})$ is given by

$$\begin{aligned}\begin{pmatrix}\widehat{\theta}\\ \widehat{\pi}\end{pmatrix}+&\begin{pmatrix}\boldsymbol{0}_{p\times p}&\boldsymbol{0}_{p\times\upsilon}\\ \boldsymbol{0}_{\upsilon\times p}&\boldsymbol{0}_{\upsilon\times\upsilon}\end{pmatrix}\begin{pmatrix}\left\{n^{-1}\sum_{i=1}^n\widetilde{\boldsymbol{s}}_{\theta\theta}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\eta})\right\}^{-1}\left\{n^{-1}\sum_{i=1}^n\widetilde{\boldsymbol{g}}_{\theta}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})\right\}^{\mathrm{T}}\\ -\left\{n^{-1}\sum_{i=1}^n\dot{\boldsymbol{g}}_{1\pi}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})\right\}^{-1}\boldsymbol{H}(\widehat{\theta},\widehat{\pi},\widehat{\eta})\end{pmatrix}\\&\times\left\{n^{-1}\sum_{i=1}^n\widetilde{\boldsymbol{A}}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})\right\}^{-1}\left\{n^{-1}\sum_{i=1}^n\widetilde{\boldsymbol{l}}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})\right\},\end{aligned}\tag{5}$$

and we can estimate the covariance matrix of $(\widetilde{\theta},\widetilde{\pi})$ by

$$\begin{aligned}&\left[\left(\boldsymbol{I}_{(p+\upsilon)\times(p+\upsilon)}\ \boldsymbol{0}_{(p+\upsilon)\times q}\right)\left\{n^{-1}\sum_{i=1}^n\widetilde{\boldsymbol{A}}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})\right\}^{-1}\right]\\&\quad\times\left\{n^{-1}\sum_{i=1}^n\widetilde{\boldsymbol{l}}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})^{\otimes 2}\right\}\\&\quad\times\left[\left(\boldsymbol{I}_{(p+\upsilon)\times(p+\upsilon)}\ \boldsymbol{0}_{(p+\upsilon)\times q}\right)\right.\\&\quad\left.\times\left\{n^{-1}\sum_{i=1}^n\widetilde{\boldsymbol{A}}(\boldsymbol{X}_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})\right\}^{-1}\right]^{\mathrm{T}}.\end{aligned}$$

The justification of joint updating is given in Web Appendix A. Note that even though the asymptotic variance of the updating estimator $\widetilde{\theta}$ does not depend on the version of g_1 chosen to compute the initial estimator $\widehat{\pi}$, a poorly chosen g_1 including little information on π may pose difficulties to updating π because the inversion of $n^{-1}\sum_{i=1}^{n}\dot{g}_{1\pi}(X_i;\widehat{\theta},\widehat{\pi},\widehat{\eta})$ may not be stable.

3 | IMPORTANT SPECIAL CASES

3.1 | Parametric model with known mean

We first consider the simple case of a parametric regression model with auxiliary information on the mean outcome (Qin, 2000). This case arises naturally in microeconometric settings (Imbens and Lancaster, 1994). For example, we may build a linear regression model of food expenditure on income of a sample, with auxiliary information of the national average household expenditure on food in the United States available from the US Census Bureau.

Suppose that we observe an *i.i.d.* sample of $X \equiv (Y,Z)$, and we model the conditional distribution of Y given Z by $f(Y;Z,\theta)$. An initial estimator for θ can be obtained by maximizing the log-likelihood function, that is, $m(X;\theta)=\log f(Y;Z,\theta)$. Suppose that auxiliary information on the mean of Y, μ_Y, is available. The auxiliary information is then identified by $g(Z;\theta)=E_\theta(Y|Z)-\mu_Y$, where $E_\theta(Y|Z)$ is the conditional expectation of Y given Z with the conditional density given by $f(Y;Z,\theta)$.

In the following, we provide numerical simulations to examine three parametric settings presented in Qin (2000). In particular, we consider three settings:

(1) $Y\sim N(\theta_{00}+\theta_{10}Z,1)$, $Z\sim N(1,1)$, and $\theta_0\equiv(\theta_{00},\theta_{10})=(1,0.5)$.
(2) $Y\sim Exp(1/(\theta_{00}+\theta_{10}Z))$, $Z\sim\chi^2(1)$, and $\theta_0=(1,1)$.
(3) $Y\sim Exp(\exp(-\theta_{00}-\theta_{10}Z))$, $Z\sim N(0,1)$, and $\theta_0=(1,1)$.

Here, we illustrate the implementation of the proposed approach in setting (c). Note that $m(X;\theta)=-\theta_0-\theta_1Z-Y\exp(-\theta_0-\theta_1Z)$ and $g(X;\theta)=\exp(\theta_0+\theta_1Z)-\exp(\theta_{00}+\theta_{10}^2/2)$. The proposed algorithm can be implemented as follows:

Step 1. We calculate the maximum likelihood estimator (without incorporating auxiliary information on the mean outcome) by

$$\begin{aligned}\widehat{\theta}=(\widehat{\theta}_0,\widehat{\theta}_1)^{\mathrm{T}}&=\arg\max_{\theta}n^{-1}\sum_{i=1}^{n}m(X_i;\theta)\\&=\arg\max_{\theta}n^{-1}\sum_{i=1}^{n}\{-\theta_0-\theta_1Z_i\\&\quad-Y_i\exp(-\theta_0-\theta_1Z_i)\}.\end{aligned}$$

Step 2. For each $i=1,\ldots,n$, we calculate the key quantities

$$\begin{aligned}s_\theta(X_i;\widehat{\theta})&=-1+Y_i\exp(-\widehat{\theta}_0-\widehat{\theta}_1Z_i)(1,Z_i)^{\mathrm{T}},\\\dot{s}_{\theta\theta}(X_i;\widehat{\theta})&=-Y\exp(-\widehat{\theta}_0-\widehat{\theta}_1Z_i)(1,Z_i)^{\otimes2},\\\dot{g}_\theta(X_i;\widehat{\theta})&=\exp(\widehat{\theta}_0+\widehat{\theta}_1Z_i)(1,Z_i)^{\mathrm{T}},\\A(X_i;\widehat{\theta})&=\begin{pmatrix}-\dot{s}_{\theta\theta}(X_i;\widehat{\theta})&\dot{g}_\theta(X_i;\widehat{\theta})\\-\dot{g}_\theta(X_i;\widehat{\theta})^{\mathrm{T}}&g(X_i;\widehat{\theta})^2\end{pmatrix},\end{aligned}$$

and $l(X_i;\widehat{\theta})=(s_\theta(X_i;\widehat{\theta})^{\mathrm{T}},g(X_i;\widehat{\theta}))^{\mathrm{T}}$.

Step 3. We obtain the proposed estimator $\widetilde{\theta}$ by

$$\begin{aligned}\widetilde{\theta}=\widehat{\theta}+&\begin{pmatrix}0_{2\times2}\left\{n^{-1}\sum_{i=1}^{n}\dot{s}_{\theta\theta}(X_i;\widehat{\theta})\right\}^{-1}\\\times\left\{n^{-1}\sum_{i=1}^{n}\dot{g}_\theta(X_i;\widehat{\theta})\right\}^{\mathrm{T}}\end{pmatrix}\\&\times\left\{n^{-1}\sum_{i=1}^{n}A(X_i;\widehat{\theta})\right\}^{-1}\left\{n^{-1}\sum_{i=1}^{n}l(X_i;\widehat{\theta})\right\}.\end{aligned}$$

Here, a simplified version of (4) is used because the model is parametric and no additional parameter π is introduced by the auxiliary information.

Step 4. We calculate the variance estimator by

$$\begin{aligned}&\left[(I_{2\times2}\ 0_{2\times1})\left\{n^{-1}\sum_{i=1}^{n}A(X_i;\widehat{\theta})\right\}^{-1}\right]\\&\quad\times\left\{n^{-1}\sum_{i=1}^{n}l(X_i;\widehat{\theta})^{\otimes2}\right\}\\&\quad\times\left[(I_{2\times2}\ 0_{2\times1})\left\{n^{-1}\sum_{i=1}^{n}A(X_i;\widehat{\theta})\right\}^{-1}\right]^{\mathrm{T}}.\end{aligned}$$

Remark 5. In parametric model settings, the proposed approach is asymptotic equivalent to the approach in Qin (2000) that directly maximizes the likelihood function subject to moment constraints based on the auxiliary information. Even though Qin (2000) is conceptually easier to

TABLE 1 Simulation results for the parametric model with known mean

		MLE		Proposed				
Setting		**Bias**	**SE**	**Bias**	**SE**	**SEE**	**CP**	**RE**
(a)	θ_0	−0.001	0.100	−0.001	0.078	0.078	0.950	1.65
	θ_1	0.001	0.071	0.001	0.071	0.071	0.949	1.00
(b)	θ_0	<0.001	0.117	0.006	0.118	0.115	0.938	0.98
	θ_1	0.001	0.191	0.007	0.143	0.141	0.947	1.79
(c)	θ_0	−0.004	0.071	0.007	0.061	0.060	0.939	1.34
	θ_1	<0.001	0.071	0.007	0.061	0.061	0.950	1.35

Note: SE, SEE, and CP are the empirical standard error, mean standard error estimator, and empirical coverage probability of the 95% confidence interval, respectively. RE is the relative efficiency defined as the ratio of the variances.

understand, it is more computationally intensive to implement because numerical differentiation is needed if one would like to avoid deriving of the derivatives of the objective function. On the other hand, the proposed noniterative method requires the correct derivation of derivatives but has less computational burden. A reviewer mentioned that the requirement of derivative calculation or not is analogous to the comparison of likelihood ratio and score tests.

Table 1 shows the simulation results with sample size $n = 200$ and 10,000 replicates. The results are similar to those of the composite likelihood estimator in Qin (2000). In particular, the proposed estimators for θ_0 in settings (a) and (c) and θ_1 in settings (b) and (c) have substantial efficiency gain over the initial estimators. The proposed standard error estimator is accurate, and the 95% confidence interval has a proper coverage probability.

We also evaluate the simulation results and computational speed of the proposed algorithm compared to the composite likelihood estimator in Qin (2000) using R package `glmc` and the GMM estimator using R package `gmm`. The results are shown in Web Table 1 of Web Appendix E. The proposed algorithm gives estimators with similar precision, whereas the computation speed is much (∼ 10 times) faster.

3.2 | Covariate-specific disease prevalence in case-control studies

We consider another example of incorporating covariate-specific disease prevalence in case-control studies, which has been considered in Qin et al. (2015) and Chatterjee et al. (2016). Based on the case-control data, the effects of multiple risk factors and their interactions are studied under a logistic regression model, and the disease prevalences at various levels of one of the risk factors are incorporated. Due to the case-control sampling scheme, specialized methodologies were developed. Here, we show that our unified framework covers case-control studies, which can be formulated as independent but not identically distributed observations, such that the generalized estimation procedure in the end of Web Appendix A is applicable.

Let D indicate, by the values of 1 versus 0, whether the disease is present, and $\boldsymbol{Z}$ be a set of risk factors. Suppose that the disease status follows a logistic regression model, with

$$\Pr(D = 1|\boldsymbol{Z}, \alpha^*, \boldsymbol{\beta}^{\mathrm{T}}) = \frac{\exp\{\alpha^* + \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})\}}{1 + \exp\{\alpha^* + \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})\}}, \quad (6)$$

where α^* and $\boldsymbol{\beta}$ are regression parameters. Write $\alpha = \alpha^* + \log\{(1-\pi)/\pi\}$, where $\pi = \Pr(D = 1)$ is the (unknown) disease prevalence in the general population.

For a case-control study of n_1 cases and n_0 controls, let $\boldsymbol{X}_i \equiv (D_i, \boldsymbol{Z}_i)$ $(i = 1, \ldots, n)$ be the observed data, where $n = n_0 + n_1$. It is well known that based on the case-control data, we are only able to identify $\theta \equiv (\alpha, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$, but not α^*. The maximum likelihood estimator $\widehat{\theta}$ solves the estimating equations $\sum_{i=1}^{n} \boldsymbol{s}(\boldsymbol{X}_i; \widehat{\theta}) = \boldsymbol{0}$, where

$$\boldsymbol{s}(\boldsymbol{X}; \theta) = \left\{ D - \frac{\rho \exp(\theta^{\mathrm{T}}\widetilde{\boldsymbol{Z}})}{1 + \rho \exp(\theta^{\mathrm{T}}\widetilde{\boldsymbol{Z}})} \right\} \widetilde{\boldsymbol{Z}},$$

$\rho = n_1/n_0$ and $\widetilde{\boldsymbol{Z}}_i = (1, \boldsymbol{Z}_i^{\mathrm{T}})^{\mathrm{T}}$.

Suppose that the effects of part of $\boldsymbol{Z}$ on D have been well studied, such that the disease prevalence at various levels of $\boldsymbol{Z}$ is available based on published information, that is, $\Pr(D = 1|\boldsymbol{Z} \in \boldsymbol{\Omega}_k) = c_k$ for $k = 1, \ldots, K$, where c_k is the known covariate-specific disease prevalence. We show in Web Appendix B that the auxiliary information can be summarized as $\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{X}_i; \theta, \pi) = \boldsymbol{0}$, where $\boldsymbol{g}(\boldsymbol{X}; \theta, \pi) = (g_1(\boldsymbol{X}; \theta, \pi), \ldots, g_K(\boldsymbol{X}; \theta, \pi))^{\mathrm{T}}$ and

$$g_k(\boldsymbol{X}; \theta, \pi) = I(\boldsymbol{Z} \in \boldsymbol{\Omega}_k) \frac{\pi \exp(\theta^{\mathrm{T}}\widetilde{\boldsymbol{Z}}) - (1-\pi)\dfrac{c_k}{1-c_k}}{1 + \rho \exp(\theta^{\mathrm{T}}\widetilde{\boldsymbol{Z}})}.$$

We then apply the estimator (5), where an initial estimator for π is obtained by solving $\sum_{i=1}^{n} g_1(\boldsymbol{X}_i; \widehat{\theta}, \pi) = 0$. The

TABLE 2 Simulation results for the case-control studies with covariate-specific disease prevalence

		MLE		Proposed				
(n_0, n_1)		Bias	SE	Bias	SE	SEE	CP	RE
(1000,2000)	α	−0.008	0.026	−0.004	0.019	0.019	0.948	1.89
	β_1	0.004	0.056	−0.002	0.025	0.025	0.950	4.94
	β_2	<0.001	0.050	−0.001	0.049	0.049	0.950	1.04
	β_3	0.001	0.048	−0.001	0.019	0.019	0.953	6.35
	π			0.001	0.003	0.003	0.952	
(2000,2000)	α	−0.008	0.023	−0.006	0.016	0.016	0.942	2.09
	β_1	0.002	0.046	−0.002	0.021	0.021	0.953	4.91
	β_2	0.001	0.042	−0.001	0.040	0.040	0.948	1.07
	β_3	<0.001	0.040	<0.001	0.016	0.016	0.952	6.50
	π			0.001	0.003	0.003	0.942	

Note: SE, SEE, and CP are the empirical standard error, mean standard error estimator, and empirical coverage probability of the 95% confidence interval, respectively. RE is the relative efficiency defined as the ratio of the variances.

asymptotic variance of $\widetilde{\theta}$ is given in Web Appendix B, and is the same to that of the empirical likelihood estimator in Qin et al. (2015).

We examined the performance of the proposed procedure in simulation studies. In particular, we generated two covariates Z_1, Z_2 that are standard normal distributed with correlation 0.5 and the disease status D from (6) with $\boldsymbol{Z} = (Z_1, Z_2, Z_1Z_2)^{\mathrm{T}}$, $\alpha^* = -1.5$, and $\boldsymbol{\beta} = (1, 0.08, 0.05)^{\mathrm{T}}$. For each simulated replicate, n_1 cases and n_0 controls were randomly generated. Suppose that the (population) disease prevalences for Z_1 in the intervals $(-\infty, -0.67]$, $(-0.67, 0]$, (0,0.67], and $(0.67, \infty)$ are known.

Table 2 shows the simulation results based on 10,000 replicates. The proposed estimators for α, β_1, and β_3 have substantial efficiency gain over the corresponding initial maximum likelihood estimators. The proposed estimator for π has small bias. The proposed variance estimator is accurate, with reasonable coverage probability of the 95% confidence intervals.

We also evaluated the performance of the proposed approach under different covariate distributions. Specifically, we generate $Z_2^* = I(Z_2 > 0)$, $Z_3 = \Phi(Z_1) - 0.5$, and $Z_4 = \Phi(Z_2) - 0.5$, where Φ is the cumulative distribution function for standard normal. That is, Z_2^* is a binary variable (correlated with Z_1) with success probability 0.5, and Z_3 and Z_4 are correlated Unif(−0.5,0.5) random variables. We considered two additional simulated settings with $\boldsymbol{Z} = (Z_1, Z_2^*, Z_1Z_2^*)^{\mathrm{T}}$ and $\boldsymbol{Z} = (Z_3, Z_4, Z_3Z_4)^{\mathrm{T}}$. For the setting with variables (Z_3, Z_4), we suppose that the (population) disease prevalences for Z_3 in the intervals $[-0.5, -0.25]$, $(-0.25, 0]$, (0,0.25], and (0.25,0.5] are known. The simulation results are shown in Web Table 2 of Web Appendix E. The general conclusion of simulation results is similar, suggesting that the performance of the proposed approach is not sensitive to covariate distributions.

3.3 | Survival regression models with t-year survival constraints

Cancer registries often publish survival probabilities for various cancer sites and subgroups. Here, we consider a semiparametric setting, where the auxiliary information of subgroup t-year survival probabilities is available for the analysis of right-censored data under the proportional hazards or proportional odds model. The setting with the proportional hazards model has been studied in Huang et al. (2016), where a special structure of the proportional hazards model was exploited. Their method does not have a straightforward extension to other semiparametric models. To illustrate our general methodology, we provide specific derivations for the proportional hazards model, whereas some results for the proportional odds model are given in Web Appendix D.

Let T denote the survival time that follows the proportional hazards model, with cumulative hazard function given by

$$\Lambda(t|\boldsymbol{Z}) = \Lambda(t)\exp(\theta^{\mathrm{T}}\boldsymbol{Z}), \tag{7}$$

where $\Lambda(\cdot)$ is an unspecified nondecreasing function and θ is a p-vector of regression parameters. Let C be a censoring time that is conditional independent of T given the covariates $\boldsymbol{Z}$, such that we observe $Y \equiv \min(T, C)$ and $\Delta \equiv I(T \le C)$. For a random sample of n subjects, the observed data include $\boldsymbol{X}_i \equiv \{Y_i, \Delta_i, \boldsymbol{Z}_i\}$ for $i = 1, \ldots, n$. The nonparametric maximum likelihood estimator (NPMLE) $(\widehat{\theta}, \widehat{\Lambda})$ maximizes the objective function $\sum_{i=1}^{n} m(\boldsymbol{X}_i; \theta, \Lambda)$ with $m(\boldsymbol{X}; \theta, \Lambda) = \Delta\{\theta^{\mathrm{T}}\boldsymbol{Z} + \log\Lambda\{Y\}\} - \Lambda(Y)\exp(\theta^{\mathrm{T}}\boldsymbol{Z})$, where $\Lambda\{t\}$ is the jump size of Λ at t. The NPMLE $(\widehat{\theta}, \widehat{\Lambda})$ can be easily obtained using common software, for example, the R package `survival` or SAS procedure `phreg`.

TABLE 3 Simulation results for the proportional hazards model with auxiliary survival probabilities

		MLE		Proposed ($\pi = 1$)					Proposed (π estimated)				
(n, π)		Bias	SE	Bias	SE	SEE	CP	RE	Bias	SE	SEE	CP	RE
(100,1)	θ_1	−0.024	0.215	−0.006	0.069	0.075	0.956	9.7	0.004	0.065	0.066	0.933	11.0
	θ_2	0.030	0.284	−0.026	0.203	0.198	0.942	2.0	0.007	0.275	0.259	0.937	1.1
	θ_3	−0.010	0.290	−0.008	0.212	0.201	0.939	1.9	−0.030	0.217	0.202	0.934	1.8
	π								0.080	0.313	0.293	0.939	
(400,1)	θ_1	−0.004	0.097	−0.002	0.028	0.029	0.954	12.1	0.001	0.027	0.028	0.947	13.0
	θ_2	0.004	0.132	−0.009	0.097	0.096	0.952	1.9	−0.002	0.130	0.127	0.944	1.0
	θ_3	−0.004	0.130	−0.002	0.096	0.096	0.942	1.8	−0.008	0.098	0.097	0.944	1.8
	π								0.017	0.136	0.134	0.948	
(100,1.5)	θ_1	−0.024	0.215	0.013	0.077	0.104	0.983	7.8	0.002	0.064	0.066	0.938	11.2
	θ_2	0.030	0.284	−0.281	0.189	0.205	0.733	2.3	0.007	0.275	0.259	0.937	1.1
	θ_3	−0.010	0.290	0.038	0.206	0.214	0.952	2.0	−0.029	0.217	0.202	0.934	1.8
	π								0.115	0.468	0.440	0.939	
(400,1.5)	θ_1	−0.004	0.097	0.021	0.028	0.046	0.988	11.6	0.001	0.027	0.028	0.948	13.0
	θ_2	0.004	0.132	−0.284	0.089	0.099	0.154	2.2	−0.002	0.130	0.127	0.944	1.0
	θ_3	−0.004	0.130	0.047	0.092	0.102	0.942	2.0	−0.007	0.098	0.097	0.944	1.8
	π								0.025	0.204	0.200	0.948	

Note: SE, SEE, and CP are the empirical standard error, mean standard error estimator, and empirical coverage probability of the 95% confidence interval, respectively. RE is the relative efficiency defined as the ratio of the variances.

Suppose that we obtain the t_k-year survival probability for the kth subgroup of subjects $(k = 1, \ldots, K)$ as auxiliary information. Write $\boldsymbol{\Omega}_k$ as the collection of covariate values for subjects in subgroup k and c_k as the corresponding t_k-year survival probability. Then, the additional estimating equations are given by $\boldsymbol{g}(\boldsymbol{X}; \theta, \Lambda) = (g_1(\boldsymbol{X}; \theta, \Lambda), \ldots, g_K(\boldsymbol{X}; \theta, \Lambda))^{\mathrm{T}}$, with

$$g_k(\boldsymbol{X}; \theta, \Lambda) = I(\boldsymbol{Z} \in \boldsymbol{\Omega}_k)\left[\exp\left\{-\Lambda(t_k)\exp(\theta^{\mathrm{T}}\boldsymbol{Z})\right\} - c_k\right].$$

In some cases, the auxiliary survival information may not be consistent with the original individual-level data due to inclusion or exclusion criteria of the clinical study. Huang et al. (2016) accommodated the inconsistency by the inclusion of an unknown parameter π such that the auxiliary information is summarized as

$$g_k(\boldsymbol{X}; \theta, \pi, \Lambda) = I(\boldsymbol{Z} \in \boldsymbol{\Omega}_k)[\exp\{-\pi\Lambda(t_k)\exp(\theta^{\mathrm{T}}\boldsymbol{Z})\} - c_k] \tag{8}$$

for $k = 1, \ldots, K$. We provide the derivatives of the functions and the least favorable directions, which are essential to calculate the proposed estimator (4) or (5), in Web Appendix C. We also derive the asymptotic variance of $\widetilde{\theta}$, which is the same as that of the double empirical likelihood estimator in Huang et al. (2016).

We illustrate the performance of the proposed estimators in simulated settings. In particular, we generated two independent covariates $Z_1 \sim N(0, 1)$ and $Z_2 \sim$ Bernoulli(0.5). The survival time T was generated from model (7) with $\boldsymbol{Z} = (Z_1, Z_2, Z_1Z_2)^{\mathrm{T}}$, $\theta = (-0.5, 1, -0.5)^{\mathrm{T}}$, and $\Lambda(t) = t^2$. We generated the censoring time $C \sim$ Uniform(0, 2.52) to have a 30% censoring rate. We considered two forms of auxiliary information with $\pi = 1$ and $\pi = 1.5$. The auxiliary information concerns the survival probabilities at $t_1 = t_2 = 0.5$ with $\boldsymbol{\Omega}_1 = \{Z_1 \le 0, Z_2 = 0\}$ and $\boldsymbol{\Omega}_2 = \{Z_1 > 0, Z_2 = 0\}$.

We considered sample sizes $n = 100$ or 400 with 10,000 replicates. Table 3 summarizes the results for the maximum likelihood estimator and the proposed estimators with known $\pi = 1$ or estimated π. When the auxiliary information is consistent with the individual-level data $(\pi = 1)$, the proposed estimator with known $\pi = 1$ has substantial efficiency gain over the initial maximum likelihood estimators, especially for θ_1. The proposed estimator with estimated π has similar efficiency gain for θ_1 and θ_3, but has less efficiency gain for θ_2. The proposed variance estimators are accurate, with reasonable coverage probability of the 95% confidence intervals. When the auxiliary information is inconsistent with the data, the proposed estimator with $\pi = 1$ is biased, especially for θ_2. The proposed estimator with estimated π is virtually unbiased when $n = 400$, and has demonstrated substantial efficiency gain over the initial maximum likelihood estimator.

A similar simulation setting has been considered by Huang et al. (2016). In Web Table 3 of Web Appendix E, we show the simulation results and computation time

based on the approach in Huang et al. (2016). The performance of the double empirical likelihood estimator in Huang et al. (2016) is similar to the proposed estimator, whereas the computation time is much longer, especially for large sample sizes, because an iterative algorithm was applied.

4 | APPLICATION

We applied the proposed methods to a chemotherapy study for Stage III colon cancer that was originally described in Laurie et al. (1989). We considered the open-source dataset that is included in the R package `survival` (Therneau et al., 2021) and is closest to that of the final report in Moertel et al. (1995). In the study, patients diagnosed with stage III colon cancer were enrolled between March 1984 and October 1987. The subjects were randomized such that 315, 310, and 304 patients received observation (Obs), levamisole alone (Lev), and levamisole combined with fluorouracil (Lev + 5FU) treatments, respectively. The patients were followed for up to 9 years for the outcomes of cancer recurrence and death. The analysis with both outcomes was considered in Lin (1994) using a different version of the data with a shorter follow-up period. For the purpose of illustration, we modeled death using the proportional hazards model and associate the survival rate with the treatments, gender, and diagnosis age.

As introduced in the introduction, The SEER Program collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 34.6 % of the U.S. population. Starting 1973, the SEER Program registries routinely collect data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status. It publishes annual report, the SEER Cancer Statistics Review, on the most recent cancer incidence, mortality, survival, prevalence, and lifetime risk statistics. The SEER Cancer Statistics have been used by thousands of researchers, clinicians, public health officials, policy makers, community groups, and the public for cancer incidence and survival statistics in the United States (Huang et al., 2016).

Here, we analyze the data from the chemotherapy study combining with the 5-year gender-specific survival information reported in SEER. In particular, the 5-year survival rates among regional colon cancer patients are 66.7% for males and 66.6% for females among those diagnosed from 1986 to 1992, based on the SEER Cancer Statistics Review, 1973–1993 (National Cancer Institute, 1997). The populations in the chemotherapy study and the SEER Program may be different; however, the conditional effect of the covariates may be more generalizable. Therefore, we apply the proposed method with estimated π as in (8) to accommodate potential inconsistency.

Table 4 shows the results from the colon cancer study and those combined with the SEER statistics. Using the proposed approach, the effect of gender is estimated with a substantial improvement in accuracy. The Wald test for $\pi = 1$ gives a p-value of < 0.0001, indicating that there is significant difference among the population of the chemotherapy study and the SEER population. Based on the proposed approach, the effects of the diagnosis age and the treatments were estimated with slightly larger standard errors.

For comparison, we analyzed the data using the double empirical likelihood method in Huang et al. (2016). As they estimated the standard errors by a bootstrapping approach, we also compared a bootstrapped standard error of the proposed estimator. The results in Table 4 are based on 1000 bootstrapping samples. The main conclusion based on the double empirical likelihood method in Huang et al. (2016) is similar; however, the proposed estimation procedure is much faster (over 300 times faster in this example) than the double empirical likelihood approach.

If the parameter π is not sufficient to capture the heterogeneity of the chemotherapy study and SEER program populations, the difference of the original estimator $\widehat{\theta}$ and the proposed estimator $\widetilde{\theta}$ would diverge. Otherwise, $\sqrt{n}(\widehat{\theta} - \widetilde{\theta})$ converges in distribution to a zero-mean multivariate normal distribution. Therefore, we test the adequacy of applying the auxiliary information based on the SEER program by a Wald test on the difference $\widehat{\theta} - \widetilde{\theta}$. In particular, the test statistic for the difference of the effects of gender takes value 0.003 with a p-value of 0.96, such that the application of the auxiliary information based on the SEER program may not lead to bias in estimation.

5 | DISCUSSION

The conventional empirical likelihood approach assumes the same population for the original study and the auxiliary information. This assumption may be relaxed in multiple ways. In Section 3.3, an unknown parameter π is included to accommodate the potential inconsistency in the t-year survival probabilities in the analysis of survival data. Another approach is to directly model the density ratio between auxiliary data and the sample to reweight the auxiliary information. This idea is similar to the synthetic likelihood in Chatterjee et al. (2016). The proposed method can also be applied to handle additional nuisance parameters in the density ratio model. In addition, the difference of $\widetilde{\theta}$ and $\widehat{\theta}$ would diverge if the original study and the auxiliary information are not compatibility. That is, a test on population compatibility can be formulated based

TABLE 4 Parameter estimates for the regression analysis in colon cancer study

	PLE		Proposed			Huang et al. (2016)	
Covariate	**Est**	**SEE**	**Est**	**SEE**	**BSE**	**Est**	**BSE**
Gender: male	−0.0004	0.094	−0.006	0.005	0.008	−0.006	0.006
Diagnosis age	0.002	0.004	0.002	0.005	0.004	0.002	0.003
Lev	−0.027	0.110	−0.027	0.118	0.109	−0.027	0.113
Lev + 5FU	−0.374	0.119	−0.374	0.129	0.115	−0.374	0.112
π			0.699	0.059	0.079	0.627	0.047

Note: PLE, SEE, and BSE are the partial likelihood estimator, standard error estimator, and bootstrapped standard error, respectively.

on the estimator difference. We have illustrated such test in the application of chemotherapy study in Section 4.

In some cases, the auxiliary information is either precise or comes from a large separate study, where the variation can be ignored. However, we must take the variation into account when the auxiliary information comes from an independent source with a limited sample size. When the auxiliary information comes from the cohort from which the data are drawn, the correlation between the auxiliary information and the data should also be considered. For example, bootstrap procedures were proposed to address the variation and correlation that arise from the overlapping samples in Qin et al. (2015). Recently, Zhang et al. (2020) proposed an analytical modification of the empirical likelihood objective function to jointly model the uncertainty distribution of the parameter estimates. Extending our current methodology to accommodate such cases is not straightforward, and it is especially difficult for semiparametric models with infinite-dimensional parameters. Exploring such extensions would be important future research.

In some cases, the auxiliary information may not be in a form of equality but rather be presented as inequality constraints. For example, when the auxiliary information is on baseline hazard function from a proportional hazards model, people are often reluctant to assume a precise value for the baseline feature, but interval constraints may be more reasonable (Zhou, 2006). However, this type of information may not be useful when sample size goes to infinity and the constraints are inactive, that is, the true expectation lies within the interior of the interval constraints. When the interval constraints are active in large samples, the statistical properties will be equivalent to the case with equality constraints.

ACKNOWLEDGMENTS

The authors thank an Associate Editor and two reviewers for their constructive comments that improved the paper. This work was supported by the U.S. National Institutes of Health grants R01 HL122212, U24 AG072122, P30 MH123248 and National Science Foundation grant DMS 1711952. Scientific Computing Infrastructure at Fred Hutch funded by ORIP grant S10OD028685.

DATA AVAILABILITY STATEMENT

The chemotherapy study data that support the findings in this paper are openly available in R package "survival" at https://cran.r-project.org/web/packages/survival/index.html.

ORCID

Fei Gao https://orcid.org/0000-0001-6797-5468

REFERENCES

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.

Chaudhuri, S., Handcock, M.S. & Rendall, M.S. (2008) Generalized linear models incorporating population level information: an empirical-likelihood-based approach. *Journal of the Royal Statistical Society, Series B*, 70, 311–328.

Chen, J. & Qin, J. (1993) Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107–116.

Chen, J. & Sitter, R. (1999) A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385–406.

Cox, D.R. & Wermuth, N. (1990) An approximation to maximum likelihood estimates in reduced models. *Biometrika*, 77, 747–761.

Han, P. & Wang, L. (2013) Estimation with missing data: beyond double robustness. *Biometrika*, 100, 417–430.

Hu, Y. & Zhou, M. (2010) Censored empirical likelihood with over-determined hazard-type constraints. http://www.ms.uky.edu/~mai/sta709/paperv1_3.pdf

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

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

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

Laurie, J.A., Moertel, C., Fleming, T., Wieand, H., Leigh, J., Rubin, J., McCormack, G., Gerstner, J., Krook, J. & Malliard, J. (1989) Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. the north central cancer treatment group and the mayo clinic. *Journal of Clinical Oncology*, 7, 1447–1456.

Lin, D. (1994) Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine*, 13, 2233–2247.

Moertel, C.G., Fleming, T.R., Macdonald, J.S., Haller, D.G., Laurie, J.A., Tangen, C.M., Ungerleider, J.S., Emerson, W.A., Tormey, D.C., Glick, J.H., Veeder, M.H. & Mailliard, J.A. (1995) Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage III colon carcinoma: a final report. *Annals of Internal Medicine*, 122, 321–326.

National Cancer Institute (1997) SEER Cancer Statistics Review 1973–1993.

National Cancer Institute (2021) Surveillance, Epidemiology, and End Results program, https://seer.cancer.gov (accessed September 30, 2021).

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

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

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

Sheng, Y., Sun, Y., Deng, D. & Huang, C.-Y. (2020) Censored linear regression in the presence or absence of auxiliary survival information. *Biometrics*, 76, 734–745.

Therneau, T.M., Lumley, T., Elizabeth, A. & Cynthia, C. (2021) *survival: Survival Analysis*. R package version 3.2-13.

Wu, C. & Sitter, R.R. (2001) A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185–193.

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

Zhou, M. (2006) The cox proportional hazards model with a partially known baseline. In *Random Walk, Sequential Analysis And Related Topics: A Festschrift in Honor of Yuan-Shih Chow*. eds. A. C. Hsiung, Z. Ying, and C.-H. Zhang, Singapore: World Scientific Publishing Co. 215–232.

SUPPORTING INFORMATION

Web Appendices A, B, C, D, and E, referenced in Sections 2 and 3 and computation codes, are available with this paper at the Biometrics website on Wiley Online Library. In addition, the R codes for simulation and for analyzing data examples are available at the Biometrics website on Wiley Online Library.

How to cite this article: Gao, F., Chan, K.C.G. Noniterative adjustment to regression estimators with population-based auxiliary information for semiparametric models. *Biometrics*. 2021;1–11. https://doi.org/10.1111/biom.13585