Received: 4 January 2022 | Revised: 6 October 2022 | Accepted: 9 November 2022

DOI: 10.1002/sim.9649

RESEARCH ARTICLE

Statistics in Medicine WILEY

Fitting additive risk models using auxiliary information

Jie Ding[1] | **Jialiang Li**[2,3] | **Yang Han**[4] | **Ian W. McKeague**[5] | **Xiaoguang Wang**[1]

[1]School of Mathematical Sciences, Dalian University of Technology, Liaoning, China

[2]Department of Statistics and Data Science, National University of Singapore, Singapore, Singapore

[3]Duke University-NUS Graduate Medical School, National University of Singapore, Singapore, Singapore

[4]Department of Mathematics, University of Manchester, Manchester, United Kingdom

[5]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, USA

Correspondence
Xiaoguang Wang, School of Mathematical Sciences, Dalian University of Technology, Liaoning, China.
Email: wangxg@dlut.edu.cn

Funding information
Academic Research Funds, Grant/Award Number: A-8000016-0-0; NSF, Grant/Award Number: DMS-2112938; NIH, Grant/Award Number: R01AG062401

There has been a growing interest in incorporating auxiliary summary information from external studies into the analysis of internal individual-level data. In this paper, we propose an adaptive estimation procedure for an additive risk model to integrate auxiliary subgroup survival information via a penalized method of moments technique. Our approach can accommodate information from heterogeneous data. Parameters to quantify the magnitude of potential incomparability between internal data and external auxiliary information are introduced in our framework while nonzero components of these parameters suggest a violation of the homogeneity assumption. We further develop an efficient computational algorithm to solve the numerical optimization problem by profiling out the nuisance parameters. In an asymptotic sense, our method can be as efficient as if all the incomparable auxiliary information is accurately acknowledged and has been automatically excluded from consideration. The asymptotic normality of the proposed estimator of the regression coefficients is established, with an explicit formula for the asymptotic variance-covariance matrix that can be consistently estimated from the data. Simulation studies show that the proposed method yields a substantial gain in statistical efficiency over the conventional method using the internal data only, and reduces estimation biases when the given auxiliary survival information is incomparable. We illustrate the proposed method with a lung cancer survival study.

KEYWORDS

adaptive lasso, additive risk model, generalized method of moments, heterogeneity, information synthesis, penalty function, sparse estimation

1 | INTRODUCTION

Efficient data use can be challenging for many medical studies with limited resources. Practical limitations in terms of funding and adequate sample size can lead to unsatisfactory inference. It is thus important to ameliorate such problems by incorporating auxiliary information from relevant external studies (when available). Typically, individual level patient data are only available "internally", but study investigators may be able to exploit information extracted from an earlier study (the "external" study) that addresses the same question.

Normally, due to privacy concerns and/or administrative problems, individual level data from external studies are seldom directly available to the public. Instead, only summary statistics can be obtained from existing publications, public databases or other related projects. It is thus of great interest to combine auxiliary summary information from external studies with individual level data to improve the estimation efficiency. To achieve this goal, various procedures have so far been proposed.[1-5] The aggregated auxiliary summary data accessible from the external studies can be quite helpful

and offer abundant information to us.[6,7] For time-to-event data,[8-10] some useful auxiliary information could be available in multiple subgroups defined by patient characteristics. Perhaps the most common quantity is the t^*-year survival probabilities reported by the literature or the public domain databases, such as the Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute (NCI).[11] To incorporate such auxiliary subgroup survival information from external sources, Huang et al[11] proposed a double empirical likelihood procedure to summarize the subgroup t^*-year survival probability as a population moment constraint under Cox proportional hazards model.[12] Extensions of this approach to the additive risk model[13] and the non-mixture cure model[14] have also been developed by He et al[15] and Han et al,[16] using the empirical likelihood method.[17] As suggested by Imbens and Lancaster,[1] the auxiliary information from the external studies and individual level data can also be integrated by applying the generalized method of moments (GMM),[18] which is asymptotically equivalent to the empirical likelihood estimators and has appealing practical performance in many applications.[19,20] GMM has been applied to address the additive-multiplicative hazard model and the accelerated failure time model.[21,22]

A key assumption in the above references is that the data distribution for individuals in the internal study is identical to that in the population from which the auxiliary data are derived. If the auxiliary information was obtained from a different reference population, borrowing such auxiliary information directly into the internal study may produce biased estimation.[16,23,24] Violation of this type of comparability assumption is a serious issue in practice.[24,25] If we simply adopt existing methods to combine the external and internal data without checking the comparability, we might have incorrectly incorporated potentially incomparable information, and consequently any seemingly efficiency gain could be spurious. We acknowledge that in addition to incomparability, another complicated problem is the non-negligible uncertainty,[26,27] usually caused by the sampling variability of summary auxiliary information. Such a variability can often be ignored when the sample size of external study is sufficiently large.[11,15] We intend to address the incomparability in this paper and leave out the sample uncertainty issue.

To tackle the problem of potential incomparability, methods based on reference-data have been discussed in earlier works.[5,16,28] However, the reference datasets may not always be available in practice. As an alternative solution, Chen et al[24] proposed the penalized maximum empirical likelihood estimator under a Cox proportional hazards model. Similar ideas have been applied to the parametric regression models by Zhai and Han.[25] By including nuisance parameters to characterize the potential discrepancy between the true underlying subgroup survival probabilities of internal data and the auxiliary quantities from external studies, they are able to adaptively determine the degree of information borrowed from the external studies using a penalization approach.[29-32] Despite its novelty, the computational cost of their method is quite expensive due to the nonlinear optimization of regularized empirical likelihood and only relatively small number of discrepancy parameters are allowed in numerical studies.

In this article, we focus on analyzing right censored survival data under the semiparametric additive risk model,[13,33,34] and propose a new method to incorporate external auxiliary survival information which may be incomparable with the internal data. As an important alternative to Cox proportional hazards model, the additive risk model assumes that the covariate effects act in an additive manner on the hazard rate. Since its invention, the model has gained its popularity in plenty of applications where the multiplicative structure fails to approximate the lifetime distribution.[33-35] Focusing on the risk difference instead of the hazard ratio may also induce a more direct interpretation of the regression parameters, drawing a connection to the familiar linear regression.[36,37] In a semiparametric additive hazards model with an unspecified baseline hazard function, Lin and Ying[33] obtained an analytic form estimator for the regression parameters. McKeague and Sasieni[34] presented a weighted least squares estimation method for partly parametric additive risk model and established its asymptotic efficiency. These earlier works encouraged the implementation of additive risk model since its solution does not require any complicated nonlinear optimization. Extensions to multivariate survival data or interval-censored data with a cure fraction have been studied in recent years.[38-41] Furthermore, additive risk models have been favored over Cox models in causal inference since the two-stage method for linear regression can be directly adopted for the former but not the latter.[42-44] Rava and Xu[45] recently studied the explained variation under the additive risk model.

As there is no simple likelihood for the regression parameters under the additive risk model, it is difficult for us to adapt Chen et al's[24] procedure which admits an analytic form for the partial likelihood. We consider the GMM[46-49] framework in this article. Specifically, an adaptive lasso penalty function[31] is adopted so that the magnitude of heterogeneity can be accurately assessed. Owing to the additive structure of our model and the proposed optimization procedure, all nuisance parameters can be profiled out to yield an explicit expression. Therefore increasing the number of nuisance parameters does not increase the computing complexity for estimating the parameters of primary interest. We prove that the proposed estimator is not only more efficient than the one without using the auxiliary information, but also as efficient as the existing estimators when all the undesirable auxiliary information has been excluded from consideration. To facilitate

statistical inference of the regression parameters, a consistent estimator for the asymptotic covariance matrix is further constructed accordingly.

The remainder of this article is organized as follows. In Section 2, we review the additive risk model and introduce the auxiliary subgroup survival information. In Section 3, after formulating the auxiliary information in a set of estimating equations, we propose a penalized GMM approach to control the degree of information borrowed from the external studies and further develop a computational algorithm to estimate all unknown parameters. The asymptotic properties of the proposed estimators are then carefully established. In Section 4, numerical comparisons are conducted based on simulated data under a variety of settings. In Section 5, we apply the proposed method to analyze a medical data set. Section 6 concludes with a brief discussion, and the Appendix contains all the technical details.

2 | MODEL AND AUXILIARY SURVIAL INFORMATION

Let a non-negative random variable T be the failure time which is right censored by a random variable C. In practice, only $\tilde{T} = \min(T, C)$ and $\Delta = I(T \leq C)$ are observed in an empirical sample. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^T$ be a p-dimensional vector of covariates with bounded support. We further assume that given $\boldsymbol{Z}$, T and C are independent. Consider the additive risk model for the conditional hazard function $\lambda(t|\mathbf{Z})$ of T given $\mathbf{Z}$:

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t) + \boldsymbol{\beta}^T\boldsymbol{Z}, \tag{1}$$

where $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression parameters and $\lambda_0(t)$ is an unspecified baseline hazard function.

Let $\{(\tilde{T}_i, \boldsymbol{Z}_i, \Delta_i), i = 1, \ldots, n\}$ be n independent and identically distributed copies of $(\tilde{T}, \boldsymbol{Z}, \Delta)$. Lin and Ying[33] proposed to estimate $\boldsymbol{\beta}$ in model (1) from the following estimating function

$$\boldsymbol{\Phi}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\infty}\{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\beta}^T\boldsymbol{Z}_i dt\}, \tag{2}$$

where $Y_i(t) = I(\tilde{T}_i \geq t)$, $N_i(t) = I(\tilde{T}_i \leq t, \Delta_i = 1)$ and $\overline{\boldsymbol{Z}}(t) = \{\sum_{j=1}^n Y_j(t)\boldsymbol{Z}_j\}/\{\sum_{j=1}^n Y_j(t)\}$ with the convention $0/0 = 0$. If we denote $A_1 = \sum_{i=1}^n \int_0^{\infty} Y_i(u)\{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t)\}^{\otimes 2}du$ and $A_2 = \sum_{i=1}^n \int_0^{\infty}\{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t)\}dN_i(u)$, the resulting estimator of $\boldsymbol{\beta}$ has a "least squares estimator" form $\hat{\boldsymbol{\beta}}_{LY} = A_1^{-1}A_2$, where $a^{\otimes 2} = aa^T$ for a vector or matrix a. The cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ can be subsequently estimated by

$$\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}_{LY}) = \int_0^t \frac{\sum_{i=1}^n\{dN_i(u) - Y_i(u)\hat{\boldsymbol{\beta}}_{\mathrm{LY}}^T\boldsymbol{Z}_i du\}}{\sum_{j=1}^n Y_j(u)}.$$

The survival probabilities for different patient subgroups are sometimes available from an external study. For $k = 1, \ldots, K$, let Ω_k denote the kth subgroup and ϕ_k be the corresponding survival probability at time t^* provided in the external study. These probability values $\mathcal{A} = \{\phi_k, k = 1, \cdots, K\}$ should be treated with caution if we cannot verify their comparability with the internal data $\{(\tilde{T}_i, \boldsymbol{Z}_i, \Delta_i), i = 1, \ldots, n\}$. More precisely, for the internal data we are interested in, define the true underlying survival probability of the subgroup Ω_k as ϕ_k^*, that is,

$$P(T > t^*|\boldsymbol{Z} \in \Omega_k) = \phi_k^*, \tag{3}$$

for $k = 1, \ldots, K$. Most existing methods implicitly demand that $\phi_k^* = \phi_k$ for all $k = 1, \ldots, K$, and such a comparability assumption may not always hold.[11,24,25]

It is possible that some of these available auxiliary subgroup survival probabilities are incorrect due to the potential discrepancy between the internal and external study populations. In particular, we consider that there exists at least one $k \in \{1, \ldots, K\}$, such that $\phi_k^* \neq \phi_k$, and define the following undesirable subset of all biased auxiliary information $\widetilde{\mathcal{A}} = \{\phi_k : \phi_k \neq \phi_k^* \text{ for } k = 1, \cdots, K\}$. We refer to elements in $\widetilde{\mathcal{A}}$ as incomparable auxiliary survival rates and the remaining elements in $\mathcal{A} - \widetilde{\mathcal{A}}$ as comparable ones. Using incomparable ϕ_k directly may result in substantial estimation bias, and consequently any apparent gain in efficiency could be spurious. We aim to fit the additive risk model by incorporating such external information appropriately to maintain estimation consistency and improve the efficiency of estimation.

3 | ESTIMATION METHOD

3.1 | Generalized method of moments with penalization

We introduce a K-dimensional parameter vector $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)^T$ to accommodate the potential incomparability by setting $\tau_k = \phi_k^* - \phi_k$, for $k = 1, \ldots, K$. Then, equation (3) can be re-written as

$$P(T > t^* | \boldsymbol{Z} \in \Omega_k) = \phi_k + \tau_k, \tag{4}$$

for $k = 1, \ldots, K$. There are situations where, for certain external studies, the vector $\boldsymbol{\tau}$ is not zero for all components. From equation (4), we can see that if all the auxiliary subgroup survival probabilities from the external study are comparable with the internal study population, that is, $\widetilde{\mathcal{A}}^* = \emptyset$, we have $\tau_k = \phi_k^* - \phi_k = 0$ for all $k = 1, \ldots, K$. Otherwise, the bias parameter $\tau_k = \phi_k^* - \phi_k \neq 0$ for some k. Thus, detecting the sparsity structure in $\boldsymbol{\tau}$ could be an effective way to check the comparability assumption.

By the use of the double expectation formula, the t^*-year survival probabilities in (4) can be expressed as a system of K unbiased estimating equations

$$E\left[I(\boldsymbol{Z} \in \Omega_k)\{\exp\{-\Lambda_0(t^*) - \boldsymbol{\beta}^T \boldsymbol{Z} t^*\} - (\phi_k + \tau_k)\}\right] = 0,$$

for $k = 1, \ldots, K$, under the additive risk model in (1). Write $\alpha = \Lambda_0(t^*)$, and define

$$\psi^{(k)}(\boldsymbol{Z}; \boldsymbol{\beta}, \alpha, \boldsymbol{\tau}) = I(\boldsymbol{Z} \in \Omega_k)\{\exp\{-\alpha - \boldsymbol{\beta}^T \boldsymbol{Z} t^*\} - (\phi_k + \tau_k)\},$$

for $k = 1, \ldots, K$. The population moment equations can be approximated by the sample moment estimating equations

$$\Psi_n^{(k)}(\boldsymbol{\xi}) = \frac{1}{n}\sum_{i=1}^{n} \psi^{(k)}(\boldsymbol{Z}_i; \boldsymbol{\beta}, \alpha, \boldsymbol{\tau}) = 0, \quad k = 1, \ldots, K, \tag{5}$$

where $\boldsymbol{\xi} = (\boldsymbol{\beta}^T, \alpha, \boldsymbol{\tau}^T)^T$. We need to estimate $\boldsymbol{\beta}$ and identify non-zero components of $\boldsymbol{\tau}$ to properly incorporate the auxiliary information from the external study.

We can combine both the estimating function $\boldsymbol{\Phi}_n(\boldsymbol{\beta})$ in (2) and the estimating functions $\boldsymbol{\Psi}_n(\boldsymbol{\xi}) = (\Psi_n^{(1)}(\boldsymbol{\xi}), \ldots, \Psi_n^{(k)}(\boldsymbol{\xi}))^T$ derived from the auxiliary survival information in the following way. Specifically, we consider the penalized GMM estimator that minimizes the following loss function

$$Q_n(\boldsymbol{\xi}) = \boldsymbol{U}_n(\boldsymbol{\xi})^T \boldsymbol{W}_n \boldsymbol{U}_n(\boldsymbol{\xi}) + \lambda_n \sum_{k=1}^{K} w_k |\tau_k|, \tag{6}$$

where $\boldsymbol{U}_n(\boldsymbol{\xi}) = (\boldsymbol{\Psi}_n(\boldsymbol{\xi})^T, \boldsymbol{\Phi}_n(\boldsymbol{\beta})^T)^T$, $\boldsymbol{W}_n$ is a specified positive definite weight matrix, λ_n is a tuning parameter that controls the strength of the penalty, and $\boldsymbol{w} = (w_1, \ldots, w_k)^T$ is a positive weight vector that is chosen adaptively by data.

The specification for $\boldsymbol{W}_n$ and $\boldsymbol{w}$ is crucial to achieve an optimal solution. In this paper, we recommend to set $\boldsymbol{W}_n$ as the inverse of an estimate of $\boldsymbol{\Sigma}(\boldsymbol{\xi}_0)$, where $\boldsymbol{\xi}_0$ is the true value of $\boldsymbol{\xi}$ and $\boldsymbol{\Sigma}(\boldsymbol{\xi}_0)$ is the asymptotic covariance matrix of $\sqrt{n}\boldsymbol{U}_n(\boldsymbol{\xi}_0)$. This is an optimal choice of weight matrix that yields an efficient estimator when penalty is absent in (6). Let $\hat{\boldsymbol{\xi}}^1 = (\hat{\boldsymbol{\beta}}_{\mathrm{LY}}^T, \hat{\alpha}_{\mathrm{LY}}, \tilde{\boldsymbol{\tau}}^T)^T$ be an initial estimator, where $\hat{\boldsymbol{\beta}}_{\mathrm{LY}}$ denotes Lin and Ying (1994)'s estimator, $\hat{\alpha}_{\mathrm{LY}} = \hat{\Lambda}_0(t^*, \hat{\boldsymbol{\beta}}_{\mathrm{LY}})$ is an estimator for the cumulative hazard function evaluated at t^* and $\tilde{\tau}_k$ is given by

$$\tilde{\tau}_k = \frac{\sum_{i=1}^{n} I(\boldsymbol{Z}_i \in \Omega_k) \exp\left\{-\hat{\alpha}_{LY} - \hat{\boldsymbol{\beta}}_{\mathrm{LY}}^T \boldsymbol{Z}_i t^*\right\}}{\sum_{i=1}^{n} I(\boldsymbol{Z}_i \in \Omega_k)} - \phi_k. \tag{7}$$

We note that $\tilde{\tau}_k$ is obtained by setting (5) to be equal to 0 and then solving for τ_k directly. Such an initial estimator is consistent but not efficient since the external information is not utilized for estimating the regression coefficients. We

next write $\boldsymbol{A}_n(\boldsymbol{\xi}) = \text{diag}\left\{\frac{1}{n}\sum_{i=1}^n \psi^{(k)}(\boldsymbol{Z}_i;\boldsymbol{\beta},\alpha,\boldsymbol{\tau})^2\right\}_{k=1,\ldots,K}$ and

$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{\xi}) = \begin{pmatrix} \boldsymbol{A}_n(\boldsymbol{\xi}) & \mathbf{0} \\ \mathbf{0} & \frac{1}{n}\sum_{i=1}^n \int_0^{+\infty} \{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t)\}^{\otimes 2} dN_i(t) \end{pmatrix},$$

where $\mathbf{0}$ represents a matrix of 0s with conformable dimension. Note that we have set the off-diagonal elements of the block matrix $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\xi}}^1)$ to be $\mathbf{0}$ directly. The reason lies in the fact that the two sets of estimating equations in $\boldsymbol{U}_n(\boldsymbol{\xi})$ are asymptotically uncorrelated. See Lemma 1 in the Appendix. Meanwhile, $\boldsymbol{\Sigma}(\boldsymbol{\xi}_0)$ can be consistently estimated by $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\xi}}^1)$ and we set $\boldsymbol{W}_n = \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\xi}}^1)^{-1}$. It can be easily shown that, $\tilde{\boldsymbol{\tau}} = (\tilde{\tau}_1, \ldots, \tilde{\tau}_K)^T$ is indeed a $\sqrt{n}$-consistent estimator of $\boldsymbol{\tau}$. For $k = 1, \ldots, K$, we propose to set $w_k = 1/\tilde{\tau}_k$.

Finally, by substituting the recommended $\boldsymbol{W}_n$ and $\boldsymbol{w}$ into the target objective function $Q_n(\boldsymbol{\xi})$ in (6), our proposed estimator is given by

$$\hat{\boldsymbol{\xi}} = \arg\min_{\boldsymbol{\xi}} \left\{ \boldsymbol{U}_n(\boldsymbol{\xi})^T \boldsymbol{W}_n \boldsymbol{U}_n(\boldsymbol{\xi}) + \lambda_n \sum_{k=1}^K |\tau_k|/|\tilde{\tau}_k| \right\},$$

where $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\beta}}^T, \hat{\alpha}, \hat{\boldsymbol{\tau}}^T)^T$. We shall refer to this proposed estimator as the adaptive generalized method of moments (AGMM) estimator in the sequel.

Although we present our estimation method using the adaptive lasso penalty proposed by Zou,[31] many other familiar choices for penalty function $P_{\lambda_n}(\cdot)$ can also be considered, for example, the SCAD penalty by Fan and Li.[30] Comparing to those non-convex penalty functions, the adaptive Lasso is relatively easy to implement and generally converges very fast. Later we shall provide more details for a computationally efficient algorithm for optimizing $Q_n(\boldsymbol{\xi})$.

Note that if we are sure that all the auxiliary t^*-year survival probabilities $\mathcal{A}$ obtained from the external study are comparable to the internal data, that is, $\phi_k^* = \phi_k$ for all $k = 1, \ldots, K$, we can directly set the nuisance parameters $\boldsymbol{\tau} = \mathbf{0}$ and $\tilde{\boldsymbol{\xi}}^1 = (\hat{\boldsymbol{\beta}}_{\text{LY}}^T, \hat{\alpha}_{\text{LY}}, \mathbf{0}^T)^T$ in the minimization of $Q_n(\boldsymbol{\xi})$. This estimation procedure does not involve any unknown $\boldsymbol{\tau}$ to accommodate the potential violation of the comparability assumption. We refer to the resulting estimator of $\boldsymbol{\beta}$ as the GMM estimator and denote it by $\hat{\boldsymbol{\beta}}_{\text{GMM}}$. When all the auxiliary subgroup survival rates are comparable, $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ may perform better than our proposed AGMM estimator $\hat{\boldsymbol{\beta}}$ due to its simplicity. Otherwise, if some of the auxiliary subgroup survival rates indeedly violate the comparability assumption, $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ is not even consistent, as will be demonstrated in our numerical study.

Throughout the derivation of our proposed estimator, we have implicitly assume that probability values in $\mathcal{A}$ come from an external study with sufficiently large sample so that the uncertainty in ϕ_k ($k = 1, \ldots, K$) is negligible compared to the internal study. In another word, the ϕ_k values are treated as fixed instead of random. Such an assumption has been adopted in similar studies.[11,24,25] We have more discussions in Section 6.

3.2 | Computations

In this subsection, we discuss some computational issues in implementing the proposed estimation method. For our objective function $Q_n(\boldsymbol{\xi})$ in (6), there are totally $p + 1 + K$ components in $\boldsymbol{\xi}$. When the number of auxiliary statistics K is large, directly minimizing $Q_n(\boldsymbol{\xi})$ may be computationally intensive.

Denote $\boldsymbol{D}_n = \frac{1}{n}\sum_{i=1}^n \int_0^{\infty} \{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t)\}^{\otimes 2} dN_i(t)$ and $\hat{\psi}_{\psi}^{(k)} = \frac{1}{n}\sum_{i=1}^n \psi^{(k)}(\boldsymbol{Z}_i;\hat{\boldsymbol{\beta}}_{LY},\hat{\alpha}_{\text{LY}},\tilde{\boldsymbol{\tau}})^2$ for $k = 1, \ldots, K$. Using the fact that the matrix $\boldsymbol{W}_n$ is a block diagonal matrix composed of $\boldsymbol{A}_n(\hat{\boldsymbol{\xi}}^1)^{-1}$ and $\boldsymbol{D}_n^{-1}$, and $\boldsymbol{A}_n(\hat{\boldsymbol{\xi}}^1)$ is a diagonal matrix whose diagonal elements are $\hat{\psi}_{\psi}^{(k)}, k = 1, \ldots, K$, we can rewrite $Q_n(\boldsymbol{\xi})$ in the following form

$$Q_n(\boldsymbol{\xi}) = \boldsymbol{\Phi}_n(\boldsymbol{\beta})^T \boldsymbol{D}_n^{-1} \boldsymbol{\Phi}_n(\boldsymbol{\beta}) + \sum_{k=1}^K \left\{ (a_k(\boldsymbol{\beta},\alpha) - b_k(\boldsymbol{\beta},\alpha)\cdot\tau_k)^2 + \lambda_n |\tau_k|/|\tilde{\tau}_k| \right\}, \tag{8}$$

where $a_k(\boldsymbol{\beta},\alpha) = \sum_{i=1}^n \left[I(\boldsymbol{Z}_i \in \Omega_k) \exp\{-\alpha - \boldsymbol{\beta}^T \boldsymbol{Z}_i t^*\} - \phi_k \right] / (n^2 \hat{\psi}_{\psi}^{(k)})^{1/2}$ and $b_k(\boldsymbol{\beta},\alpha) = \sum_{i=1}^n I(\boldsymbol{Z}_i \in \Omega_k)/(n^2 \hat{\psi}_{\psi}^{(k)})^{1/2}$ for $k = 1, \ldots, K$. Based on this sort of "quadratic" representation, it can be easily shown that the objective function $Q_n(\boldsymbol{\xi})$ is convex with respect to the parameters $\boldsymbol{\tau}$.

Note that for any given $\boldsymbol{\beta}$ and α, $Q_n(\boldsymbol{\xi})$ in (8) can be uniquely minimized at $\hat{\boldsymbol{\tau}}^P(\boldsymbol{\beta},\alpha) = (\hat{\tau}_1^P(\boldsymbol{\beta},\alpha), \ldots, \hat{\tau}_K^P(\boldsymbol{\beta},\alpha))^T$ which has an explicit expression

$$\begin{aligned}\hat{\tau}_k^P(\boldsymbol{\beta},\alpha) &= \arg\min_{\tau_k}\left\{(a_k(\boldsymbol{\beta},\alpha) - b_k(\boldsymbol{\beta},\alpha)\cdot\tau_k)^2 + \lambda_n|\tau_k|/|\tilde{\tau}_k|\right\}\\ &= \frac{1}{2b_k^2(\boldsymbol{\beta},\alpha)}\eta_S\left(2a_k(\boldsymbol{\beta},\alpha)b_k(\boldsymbol{\beta},\alpha), \lambda_n/|\tilde{\tau}_k|\right),\end{aligned}$$

for all $k = 1, \ldots, K$, where

$$\eta_S(\omega,\lambda) = \text{sign}(\omega)\cdot(|\omega| - \lambda)_+.$$

The function $\eta_S(\omega,\lambda)$ is in fact the well-known soft-thresholding function in machine learning.[50] Thus, we minimize the profiled function $Q_n^P(\boldsymbol{\beta},\alpha) = Q_n\left(\boldsymbol{\beta},\alpha,\hat{\boldsymbol{\tau}}^P(\boldsymbol{\beta},\alpha)\right)$ to obtain the estimator $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}$ by the Nelder-Mead algorithm and update $\hat{\boldsymbol{\tau}} = \hat{\boldsymbol{\tau}}^P(\hat{\boldsymbol{\beta}},\hat{\alpha})$.[51] The resulting minimizer of this profiled minimization procedure is equivalent to that of minimizing $Q_n(\boldsymbol{\xi})$ directly by the convexity of the function. This profiled estimation algorithm is efficient since no matter how many components of the vector $\boldsymbol{\tau}$ we have, they will not be included in the minimization step of $Q_n^P(\boldsymbol{\beta},\alpha)$. The computation of estimating $p+1$ regular regression parameters is much more affordable than the original optimization problem. We find very stable performance of this method in our simulation studies.

3.3 | Selection of the tuning parameter

The selection consistency of adaptive lasso depends on an appropriate choice of the tuning parameter. To choose the optimal tuning parameter λ_n, we propose to use the following Bayesian information criterion (BIC)

$$\text{BIC}(\lambda_n) = \boldsymbol{U}(\hat{\boldsymbol{\xi}}_{\lambda_n})^T\boldsymbol{W}_n^{-1}\boldsymbol{U}(\hat{\boldsymbol{\xi}}_{\lambda_n}) + \frac{\log(n)}{n}|\mathcal{A}|,$$

where $|A|$ is the cardinality of the set A and $\mathcal{A} = \{k : \hat{\tau}_k \neq 0, k = 1, \ldots, K\}$.[52] The BIC balances the model fits and the model complexity and enjoys a solid theoretical support in the literature. Thus, the best λ_n is chosen as $\hat{\lambda}_n = \arg\min_{\lambda_n\in\Upsilon}\text{BIC}(\lambda_n)$ where Υ is a pre-defined set of candidate values of λ_n.

3.4 | Asymptotic properties

We establish asymptotic properties of our proposed AGMM estimator $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\beta}}^T, \hat{\alpha}, \hat{\boldsymbol{\tau}}^T)^T$ in this subsection, where true values of these parameters are denoted as $\boldsymbol{\xi}_0 = (\boldsymbol{\beta}_0^T, \alpha_0, \boldsymbol{\tau}_0^T)^T$. Write the true parameter vector of $\boldsymbol{\tau}$ as $\boldsymbol{\tau}_0 = (\boldsymbol{\tau}_{10}^T, \boldsymbol{\tau}_{20}^T)^T$, where $\boldsymbol{\tau}_{10}$ consists of $s < K$ nonzero components and $\boldsymbol{\tau}_{20}$ consists of the remaining $K - s$ zero components. Correspondingly, we write the estimator of $\boldsymbol{\tau}$ from (6) as $\hat{\boldsymbol{\tau}} = (\hat{\boldsymbol{\tau}}_1^T, \hat{\boldsymbol{\tau}}_2^T)^T$. We need to assume the regularity conditions in Lin and Ying[33] so that the asymptotic properties of $\hat{\boldsymbol{\beta}}_{\text{LY}}$ can hold. To obtain the asymptotic results of our AGMM estimator, we impose the following technical assumptions:

(C1) The true regression parameter value $\boldsymbol{\beta}_0$ and $\alpha_0 = \Lambda_0(t^*)$ are both interior points in their respective compact parameter spaces.
(C2) The estimating equation for each auxiliary subgroup survival probability has nonzero variance, that is, the value $E\{\boldsymbol{\Psi}(\boldsymbol{Z};\boldsymbol{\beta}_0,\alpha_0,\boldsymbol{\tau}_0)^2\}$ is strictly positive for $k = 1, \ldots, K$.
(C3) For $k = 1, \ldots, K$, $E\{\|I(\boldsymbol{Z}\in\Omega_k)\boldsymbol{Z}^{\otimes 2}\|\}$ are bounded, where $\|\boldsymbol{A}\| = \max\left\{\sum_{i=1}^n|a_{i1}|, \ldots, \sum_{i=1}^n|a_{in}|\right\}$ denotes the matrix maximum absolute column sum 1−norm of an $n\times n$ matrix $\boldsymbol{A} = (a_{ij})$.

Assumption (C1) is necessary to ensure the existence of the estimator we are interested in. From Assumption (C2), it is easy to derive that for $k = 1, \ldots, K$, $P(\boldsymbol{Z}\in\Omega_k) > 0$ implies that for sufficiently large sample each subgroup Ω_k will be allocated with some observations. Assumption (C3) is needed to guarantee that the second partial derivatives of the estimating functions are uniformly bounded.

The following theorem shows that $\hat{\boldsymbol{\xi}}$ is $\sqrt{n}$−consistent if $\lambda_n \to 0$ at an appropriate rate.

Theorem 1. *Under the regularity assumptions* $(C1)-(C3)$*, if* $\sqrt{n}\lambda_n = O_p(1)$*, then there exists a local minimizer* $\hat{\boldsymbol{\xi}}$ *of* $Q_n(\boldsymbol{\xi})$ *such that* $\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0\| = O_p(n^{-1/2})$.

Next we show that, when λ_n is chosen properly, the estimator of $\boldsymbol{\tau}$ has the oracle property, that is, as n goes to infinity, $\hat{\boldsymbol{\tau}}$ can perform as well as if the sparse structure was known in advance.

Theorem 2. *Assume that* $\sqrt{n}\lambda_n \to 0$, $n\lambda_n \to \infty$, *and there exists at least one k, such that*, $\tau_k = 0$. *Then, under the regularity assumptions* $(C1)-(C3)$*, with probability tending to 1, the* $\sqrt{n}-$*consistent estimator* $\hat{\boldsymbol{\tau}}$ *must satisfy the following conditions*:

(i) (Sparsity) $\hat{\boldsymbol{\tau}}_2 = 0$;
(ii) (Asymptotic normality) $\sqrt{n}(\hat{\boldsymbol{\tau}}_1 - \boldsymbol{\tau}_{10}) \to N(\mathbf{0}, (\boldsymbol{M}_{22} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12})^{-1})$ *in distribution as n tends to infinity, where* $\boldsymbol{M}_{11}, \boldsymbol{M}_{12}, \boldsymbol{M}_{21}$, *and* $\boldsymbol{M}_{22}$ *are specified in the Appendix.*

Define $\boldsymbol{\psi}_{\beta} = E\left\{\frac{\partial \boldsymbol{\Psi}(\xi)}{\partial \beta}|_{\xi=\xi_0}\right\}$ and $\boldsymbol{\psi}_{\boldsymbol{\tau}_1} = E\left\{\frac{\partial \boldsymbol{\Psi}(\xi)}{\partial \boldsymbol{\tau}_1}|_{\xi=\xi_0}\right\}$. The large sample properties of our proposed AGMM estimator $\hat{\boldsymbol{\beta}}$ is provided in the following theorem.

Theorem 3. *Let* $\boldsymbol{\Sigma}_{\mathrm{LY}}$ *be the asymptotic variance-covariance of* $\hat{\boldsymbol{\beta}}_{\mathrm{LY}}$ *and define*

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\psi}_{\beta}^T\{\boldsymbol{H}_1 - \boldsymbol{H}_1\boldsymbol{\psi}_{\boldsymbol{\tau}_1}(\boldsymbol{\psi}_{\boldsymbol{\tau}_1}^T\boldsymbol{H}_1\boldsymbol{\psi}_{\boldsymbol{\tau}_1})^{-1}\boldsymbol{\psi}_{\boldsymbol{\tau}_1}^T\boldsymbol{H}_1\}\boldsymbol{\psi}_{\beta},$$

where $\boldsymbol{H}_1$ *is a positive semi-definite matrix specified in the Appendix. Then, under the same regularity assumptions of Theorem 2, we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, (\boldsymbol{\Sigma}_{\mathrm{LY}}^{-1} + \boldsymbol{\Sigma}_0)^{-1}).$$

Remark 1. Assuming that we know which of the available auxiliary subgroup survival probabilities in $\mathcal{A}$ are comparable ones in advance, we can use their ϕ_k values in the framework of the GMM procudure directly without including additional nuisance parameters $\boldsymbol{\tau}$ and the resulting estimator may be denoted as the oracle GMM estimator $\hat{\boldsymbol{\beta}}_{\mathrm{OGMM}}$. Theorem 2 states that our proposed AGMM estimation procedure can consistently identify all potentially auxiliary subgroup survival rates $\widetilde{\mathcal{A}}$ that are derived from undesirable subgroups that produce biased values. After tedious calculation given in the Appendix, we can show that the asymptotic variance-covariance of $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{OGMM}} - \boldsymbol{\beta}_0)$ coincides with that of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ in Theorem 3, that is, $(\boldsymbol{\Sigma}_{\mathrm{LY}}^{-1} + \boldsymbol{\Sigma}_0)^{-1}$. This implies that our proposed AGMM estimator $\hat{\boldsymbol{\beta}}$ achieves the same asymptotic efficiency as if no discordant auxiliary information is involved.

Remark 2. Note that if we set $\boldsymbol{\Sigma}_1 = \boldsymbol{\psi}_{\beta}^T\boldsymbol{H}_1\boldsymbol{\psi}_{\beta}$, the asymptotic variance-covariance of $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}} - \boldsymbol{\beta}_0)$ can be expressed as $(\boldsymbol{\Sigma}_{\mathrm{LY}}^{-1} + \boldsymbol{\Sigma}_1)^{-1}$. Using the fact that matrices $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0$ are all positive semi-definite, we have

$$(\boldsymbol{\Sigma}_{\mathrm{LY}}^{-1} + \boldsymbol{\Sigma}_1)^{-1} \leq (\boldsymbol{\Sigma}_{\mathrm{LY}}^{-1} + \boldsymbol{\Sigma}_0)^{-1} \leq \boldsymbol{\Sigma}_{\mathrm{LY}},$$

immediately suggesting that the proposed AGMM estimator $\hat{\boldsymbol{\beta}}$ is more efficient than $\hat{\boldsymbol{\beta}}_{\mathrm{LY}}$ which uses the internal data only, but may be less efficient than $\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}$. When $\boldsymbol{\tau}_0 = 0$, that is, all the auxiliary subgroup survival probabilities are correctly specified and homogeneity assumption is satisfied, the matrix $\boldsymbol{\Sigma}_0$ reduces to $\boldsymbol{\Sigma}_1$, indicating that our proposed AGMM estimator $\hat{\boldsymbol{\beta}}$ is as efficient as $\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}$ despite the inclusion of nuisance parameters. However, when $\boldsymbol{\tau}_0 \neq 0$, that is, incomparability exists, although the first inequality still holds, the seemingly efficiency gain for $\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}$ is meaningless since the estimating equations have nonzero expectations and the corresponding estimator is inconsistent, as will be clearly demonstrated in our simulation studies.

For the purpose of risk predication, we need to study the asymptotic property of $\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}})$. Denote $\hat{S}(t|\boldsymbol{Z}, \hat{\boldsymbol{\beta}}) = \exp(-\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}) - t\hat{\boldsymbol{\beta}}^T\boldsymbol{Z})$, which serves as an estimator for $S(t|\boldsymbol{Z}) = \exp(-\Lambda_0(t) - t\boldsymbol{\beta}_0^T\boldsymbol{Z})$.

Theorem 4. *Suppose that the same regularity assumptions of Theorem 2 holds as well, and for a* $\zeta > 0$, *these exists continuous and bounded functions* $g(t)$ *and* $\boldsymbol{\Gamma}(t)$ *such that* $\sup_{t\in(0,\zeta]} \left|\sum_{i=1}^n \int_0^t n/\left\{\sum_{j=1}^n Y_j(u)\right\}^2 dN_i(u) - g(t)\right| \xrightarrow{p} 0$, *and* $\sup_{t\in(0,\zeta]} \|\sum_{i=1}^n \int_0^t \{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(u)\}/\{\sum_{j=1}^n Y_j(u)\}dN_i(u) - \boldsymbol{\Gamma}(t)\| \xrightarrow{p} 0$. *Then, given* $\boldsymbol{Z}$, *we have* $n^{1/2}\{\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}) - \Lambda_0(t)\}$ *and* $n^{1/2}\{\hat{S}(t|\boldsymbol{Z}, \hat{\boldsymbol{\beta}}) - S(t|\boldsymbol{Z})\}$ *converge weakly to mean zero Gaussian processes with covariance functions at* (t_1, t_2) *given by* $C_{\Lambda}(t_1, t_2)$ *and* $C_S(t_1, t_2)$, *respectively, where* $C_{\Lambda}(t_1, t_2)$ *and* $C_S(t_1, t_2)$ *are specified in the Appendix.*

3.5 | Extension to auxiliary information from multiple time points

We have focused on the scenario in which the external information is aggregate survival rates across different subgroups at a single time point. In practice, inspired by the fact that the survival probabilities of subjects change over time, aggregate survival information may be available at multiple time points (eg, 1-year and 3-year survival rates). The proposed method above can be readily generalized to accommodate such information. In this subsection, we formulate the subgroup survival probabilities at more than one survival time point, which promotes the utilization of the auxiliary time-varying survival information. More precisely, for the internal data we are interested in, the true underlying survival rates can be formulated in the form

$$P(T > t_j^* | \boldsymbol{Z} \in \Omega_{jk}) = \phi_{jk}^*, \tag{9}$$

for $j = 1, \ldots, L$ and $k = 1, \ldots, K_j$, where Ω_{jk} denotes the kth subgroup at the prespecified jth time point t_j^*. The observed auxiliary survival rates from external study are $\{\phi_{jk}, j = 1, \ldots, L \text{ and } k = 1, \ldots, K_j\}$. Similarly, for each of these values, we should be cautious about whether $\phi_{jk}^* = \phi_{jk}$ or not.

Denote $K = \sum_{j=1}^{L} K_j$ and $\alpha_j = \Lambda_0(t_j^*)$ for $j = 1, \ldots, L$. After introducing an unknown K-dimensional parameter vector $\boldsymbol{\tau} = (\tau_{11}, \ldots, \tau_{1K_1}, \ldots, \tau_{L1} \ldots, \tau_{LK_L})^T$, we can write equation (9) into $P(T > t_j^* | \boldsymbol{Z} \in \Omega_{jk}) = \phi_{jk} + \tau_{jk}$, which immediately suggest the following estimation equations

$$\Psi_n^{(jk)}(\boldsymbol{\xi}) = \frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{Z}_i \in \Omega_{jk}) \left\{ \exp\left\{ -\alpha_j - \boldsymbol{\beta}^T \boldsymbol{Z} t_j^* \right\} - (\phi_{jk} + \tau_{jk}) \right\}, \tag{10}$$

for $j = 1, \ldots, L$ and $k = 1, \ldots, K_j$, where $\boldsymbol{\xi} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}, \boldsymbol{\tau}^T)^T$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_L)^T$. Then, we consider the penalized GMM estimator that minimizes the following loss function

$$Q_n(\boldsymbol{\xi}) = \boldsymbol{U}_n(\boldsymbol{\xi})^T \boldsymbol{W}_n \boldsymbol{U}_n(\boldsymbol{\xi}) + \lambda_n \sum_{j=1}^{L}\sum_{k=1}^{K_j} w_{jk} |\tau_{jk}|, \tag{11}$$

where $\boldsymbol{\Psi}_n(\boldsymbol{\xi}) = (\Psi_n^{(11)}(\boldsymbol{\xi}), \ldots, \Psi_n^{(1K_1)}(\boldsymbol{\xi}), \ldots, \Psi_n^{(L1)}(\boldsymbol{\xi}), \ldots, \Psi_n^{(LK_L)}(\boldsymbol{\xi}))^T$ and $\boldsymbol{U}_n(\boldsymbol{\xi}) = (\boldsymbol{\Psi}_n(\boldsymbol{\xi})^T, \boldsymbol{\Phi}_n(\boldsymbol{\beta})^T)^T$. The specifications of $\boldsymbol{W}_n$, λ_n and $\boldsymbol{w} = (w_{11}, \ldots, w_{1K_1}, \ldots, w_{L1}, \ldots, w_{LK_L})^T$ can follow a similar construction as in the preceding subsections.

4 | NUMERICAL STUDIES

We next conduct simulation studies to evaluate the finite sample performance of our proposed AGMM estimator. In the first part, we generate data from a pre-specified distribution for simulation. In the second part, we use a real data set with moderate scale sample size, and randomly select a small subsample from it as the internal research.

4.1 | Simulated data sets

In this subsection, we considered two covariates Z_1 and Z_2 in our simulation studies, with Z_1 generated from a uniform distribution on $[0, 1]$ and Z_2 generated from a Bernoulli distribution with the success probability of 0.5. Given the values of $Z_1 = z_1$ and $Z_2 = z_2$, we considered the additive risk model for the potential failure time as $\lambda(t|z_1, z_2) = 2t + \beta_1 z_1 + \beta_2 z_2$, where we set the true values of the regression parameters as $(\beta_1, \beta_2) = (0.5, 0.8)$. The potential censoring time variable C was generated from a uniform distribution on $[0, c]$. We set $c = 2.2$ and 1.3 respectively, and the resulting censoring rates were approximately 30% and 50%. Sample sizes of 300 and 500 were used, and each scenario had 300 repetitions. We considered three settings for possible relationships between the internal study and external aggregate information.

Setting 1. We derived the auxiliary survival probabilities at $t^* = 0.5$ for 6 subgroups: $\Omega_1 = \{(Z_1, Z_2) : 0.67 \le Z_1 \le 1.00, Z_2 = 0\}$, $\Omega_2 = \{(Z_1, Z_2) : 0.33 \le Z_1 < 0.67, Z_2 = 0\}$, $\Omega_3 = \{(Z_1, Z_2) : 0.00 \le Z_1 \le 0.33, Z_2 = 0\}$, $\Omega_4 = \{(Z_1, Z_2) : 0.67 \le Z_1 \le 1.00, Z_2 = 1\}$, $\Omega_5 = \{(Z_1, Z_2) : 0.33 \le Z_1 < 0.67, Z_2 = 1\}$, and $\Omega_6 = \{(Z_1, Z_2) : 0.00 \le Z_1 \le 0.33, Z_2 = 1\}$. The true underlying survival information of survival probabilities for these six subgroups were $\phi_1^* = P(T >$

$t^*|\boldsymbol{Z} \in \Omega_1) = 0.632$, $\phi_2^* = P(T > t^*|\boldsymbol{Z} \in \Omega_2) = 0.687$, $\phi_3^* = P(T > t^*|\boldsymbol{Z} \in \Omega_3) = 0.747$, $\phi_4^* = P(T > t^*|\boldsymbol{Z} \in \Omega_4) = 0.424$, $\phi_5^* = P(T > t^*|\boldsymbol{Z} \in \Omega_5) = 0.461$, and $\phi_6^* = P(T > t^*|\boldsymbol{Z} \in \Omega_6) = 0.501$, respectively. The observed survival rates we obtained were set to be $(\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6) = (0.482, 0.587, 0.747, 0.424, 0.461, 0.701)$, indicating that there were 3 undesirable subgroups $\Omega_1, \Omega_2, \Omega_6$ that produce discordant survival probabilities, and thus $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \tau_6) = (0.15, 0.1, 0, 0, 0, -0.2)$.

Setting 2. We considered the case where there were more auxiliary survival probabilities available from external studies and more discordant ones, which was obviously much more complicated than the first setting of simulation experiments. We derived the auxiliary survival probabilities at $t^* = 0.5$ for 10 subgroups: $\Omega_1 = \{(Z_1, Z_2) : 0.8 \le Z_1 \le 1.0, Z_2 = 0\}$, $\Omega_2 = \{(Z_1, Z_2) : 0.6 \le Z_1 < 0.8, Z_2 = 0\}$, $\Omega_3 = \{(Z_1, Z_2) : 0.4 \le Z_1 \le 0.6, Z_2 = 0\}$, $\Omega_4 = \{(Z_1, Z_2) : 0.2 \le Z_1 \le 0.4, Z_2 = 0\}$, $\Omega_5 = \{(Z_1, Z_2) : 0.0 \le Z_1 < 0.2, Z_2 = 0\}$, $\Omega_6 = \{(Z_1, Z_2) : 0.8 \le Z_1 \le 1.0, Z_2 = 1\}$, $\Omega_7 = \{(Z_1, Z_2) : 0.6 \le Z_1 < 0.8, Z_2 = 1\}$, $\Omega_8 = \{(Z_1, Z_2) : 0.4 \le Z_1 \le 0.6, Z_2 = 1\}$, $\Omega_9 = \{(Z_1, Z_2) : 0.2 \le Z_1 \le 0.4, Z_2 = 1\}$, and $\Omega_{10} = \{(Z_1, Z_2) : 0.0 \le Z_1 < 0.2, Z_2 = 1\}$. The true underlying survival information of survival probabilities for these ten subgroups were $\phi_1^* = P(T > t^*|\boldsymbol{Z} \in \Omega_1) = 0.622$, $\phi_2^* = P(T > t^*|\boldsymbol{Z} \in \Omega_2) = 0.654$, $\phi_3^* = P(T > t^*|\boldsymbol{Z} \in \Omega_3) = 0.688$, $\phi_4^* = P(T > t^*|\boldsymbol{Z} \in \Omega_4) = 0.723$, $\phi_5^* = P(T > t^*|\boldsymbol{Z} \in \Omega_5) = 0.760$, $\phi_6^* = P(T > t^*|\boldsymbol{Z} \in \Omega_6) = 0.417$, $\phi_7^* = P(T > t^*|\boldsymbol{Z} \in \Omega_7) = 0.438$, $\phi_8^* = P(T > t^*|\boldsymbol{Z} \in \Omega_8) = 0.461$, $\phi_9^* = P(T > t^*|\boldsymbol{Z} \in \Omega_9) = 0.484$ and $\phi_{10}^* = P(T > t^*|\boldsymbol{Z} \in \Omega_{10}) = 0.509$, respectively. The observed survival rates we obtained were set to be $(\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6, \phi_7, \phi_8, \phi_9, \phi_{10}) = (0.422, 0.454, 0.588, 0.723, 0.760, 0.417, 0.438, 0.461, 0.334, 0.659)$, indicating that there were 5 undesirable subgroups Ω_1, Ω_2, Ω_3, Ω_9, Ω_{10} that produce discordant survival probabilities, and $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \tau_6, \tau_7, \tau_8, \tau_9, \tau_{10}) = (0.2, 0.2, 0.1, 0, 0, 0, 0, 0, 0.15, -0.15)$.

Setting 3. A case with auxiliary subgroup survival probabilities at 2 time points, that is, $t_1^* = 0.5$ and $t_2^* = 0.7$, was considered. The auxiliary survival probabilities derived at $t_1^* = 0.5$ were the same as those in setting 1, and we denoted the corresponding 6 groups as $\Omega_{11}, \Omega_{12}, \Omega_{13}, \Omega_{14}, \Omega_{15}$ and Ω_{16}. At $t_2^* = 0.7$, we derived the auxiliary survival probabilities for another 4 subgroups: $\Omega_{21} = \{(Z_1, Z_2) : Z_1 \ge 0.5, Z_2 = 0\}$, $\Omega_{22} = \{(Z_1, Z_2) : Z_1 < 0.5, Z_2 = 0\}$, $\Omega_{23} = \{(Z_1, Z_2) : Z_1 \ge 0.5, Z_2 = 1\}$ and $\Omega_{24} = \{(Z_1, Z_2) : Z_1 < 0.5, Z_2 = 1\}$. The true underlying survival information of survival probabilities for these subgroups were $\phi_{11}^* = P(T > t^*|\boldsymbol{Z} \in \Omega_{11}) = 0.632$, $\phi_{12}^* = P(T > t^*|\boldsymbol{Z} \in \Omega_{12}) = 0.687$, $\phi_{13}^* = P(T > t^*|\boldsymbol{Z} \in \Omega_{13}) = 0.747$, $\phi_{14}^* = P(T > t^*|\boldsymbol{Z} \in \Omega_{14}) = 0.424$, $\phi_{15}^* = P(T > t^*|\boldsymbol{Z} \in \Omega_{15}) = 0.461$, $\phi_{16}^* = P(T > t^*|\boldsymbol{Z} \in \Omega_{16}) = 0.501$, $\phi_{21}^* = P(T > t^*|\boldsymbol{Z} \in \Omega_{21}) = 0.472$, $\phi_{22}^* = P(T > t^*|\boldsymbol{Z} \in \Omega_{22}) = 0.562$, $\phi_{23}^* = P(T > t^*|\boldsymbol{Z} \in \Omega_{23}) = 0.269$ and $\phi_{24}^* = P(T > t^*|\boldsymbol{Z} \in \Omega_{24}) = 0.321$, respectively. The observed survival rates we obtained were set to be $(\phi_{11}, \phi_{12}, \phi_{13}, \phi_{14}, \phi_{15}, \phi_{16}, \phi_{21}, \phi_{22}, \phi_{23}, \phi_{24}) = (0.482, 0.587, 0.747, 0.424, 0.461, 0.701, 0.472, 0.562, 0.119, 0.171)$, indicating that there were 5 undesirable subgroups Ω_{11}, Ω_{12}, Ω_{16}, Ω_{23}, Ω_{24} that produce discordant survival probabilities, and $(\tau_{11}, \tau_{12}, \tau_{13}, \tau_{14}, \tau_{15}, \tau_{16}, \tau_{21}, \tau_{22}, \tau_{23}, \tau_{24}) = (0.15, 0.1, 0, 0, 0, -0.2, 0, 0, 0.15, 0.15)$.

In addition to reporting the performance of our proposed AGMM method, we also considered various competing methods, including Lin and Ying's[33] estimator, the GMM estimator and the oracle GMM estimator. Note that the oracle estimator was usually unavailable in a practical setting and was only used in this simulation as a reference. For comparison, we also conducted simulation studies of the estimator derived using the empirical likelihood method[15] under the comparability assumption, referred to as EL. Before examining the estimated regression coefficients $\boldsymbol{\beta}$, we first checked whether our method can consistently estimate $\boldsymbol{\tau}$ and identified nonzero components correctly. In Table 1, we reported the frequency with which each τ_k was estimated as nonzero, and also the average numbers of correct and incorrect zero estimates over 300 Monte Carlo replicates. To measure prediction accuracy of $\boldsymbol{\tau}$, the mean squared errors $(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau})^T(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau})$ were also summarized. Standard errors were given in parentheses. Overall, we can see from Table 1 that our proposed AGMM method performed well, being able to correct undesirable auxiliary information from the external study.

Next, we summarized the simulation results of the estimated regression coefficients $\boldsymbol{\beta}$ in Tables 2–4 for the three settings. The summary statistics were the average biases (Bias), the sample standard deviation (SD), the average estimated standard errors (SE), the average square root of mean squared errors (RMSE) and the empirical coverage probability (CP) of 95% confidence intervals. We further calculated the relative efficiency (RE), which was computed as the ratio of the RMSE from each of the three methods with that from the oracle GMM method. Note that we only presented the simulation results of the EL estimator under setting 1, which is relatively simple. For setting 2 and 3, subtantial biases were observed due to much more complicated incomparability settings, and in a large fraction of simulations, the program fails to converge for the empirical likelihood method. We thus did not present those results.

From Table 2, we can observe a quite large efficiency gain in the standard error estimates for those methods using subgroup survival rates as auxiliary information under setting 1. By using the correct auxiliary subgroup survival rates only, the oracle GMM estimator has the smallest SDs and RMSEs, as expected. As we can see, since part of the external survival probabilities did not match well with that of the internal data, the biases of the GMM and EL estimators were quite substantial. These approaches may generate misleading inferences with coverage probabilities less than 10%, and it did not

TABLE 1 Performance of the AGMM method in sparse estimation of $\boldsymbol{\tau}$ under different settings

N	CR	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}	Corr.	Incorr.	MSE
Setting 1														
300	30%	78	83	13	19	13	98	—	—	—	—	2.55	0.41	0.0125(0.021)
	50%	76	81	10	18	19	97	—	—	—	—	2.53	0.46	0.0147(0.023)
500	30%	96	94	6	5	7	100	—	—	—	—	2.81	0.11	0.0038(0.011)
	50%	89	90	10	9	8	99	—	—	—	—	2.73	0.22	0.0073(0.016)
Setting 2														
300	30%	98	100	90	21	19	18	21	20	100	89	4.01	0.23	0.0164(0.032)
	50%	97	100	92	23	18	18	22	24	99	86	3.95	0.26	0.0192(0.035)
500	30%	100	100	94	13	10	7	9	6	100	98	4.56	0.08	0.0048(0.016)
	50%	100	100	94	15	11	9	15	13	100	95	4.37	0.11	0.0075(0.020)
Setting 3														
300	30%	92	84	14	9	11	100	16	4	100	98	4.46	0.26	0.0139(0.031)
	50%	91	85	13	8	12	99	20	2	99	98	4.45	0.28	0.0155(0.037)
500	30%	98	93	7	3	3	100	4	1	100	100	4.83	0.09	0.0034(0.012)
	50%	97	92	8	4	4	100	6	1	100	100	4.78	0.11	0.0053(0.017)

Note: Column G_1 to Column G_{10} summarize the frequency of nonzero estimates for the groups. CR, the censoring rates; MSE, mean squared error; Corr., average number of correct zeros; Incorr., average number of incorrect zeros.

improve with increasing sample sizes and decreasing censoring rates. In contrast, our proposed AGMM method could adaptively detect discordant survival rates from heterogeneous subgroups and incorporate the external data appropriately, resulting in more reasonable estimators and inference results. Although the AGMM estimation involved additional parameters $\boldsymbol{\tau}$ to account for the potential disparities, it was still more efficient than the classical LY method, with smaller SDs and RMSEs. And it is worth noting that our method was only slightly inferior to the oracle estimator, indicating that it can perform well as if all the incorrect auxiliary survival rates have been excluded from consideration. We could also find that the SEs and SDs of the AGMM estimator were similar to each other, and the coverage probabilities were close to the nominal level, which affirmed our theoretical results.

Under setting 2 (Table 3), the proposed AGMM method still performed well in terms of estimating the regression coefficients $\boldsymbol{\beta}$ and efficiency gain. In contrast, the biases of the GMM estimator which used auxiliary information directly without considering the potential heterogeneity problem were not ignorable, and even larger than that in setting 1, due to the larger number of heterogeneous subgroups that provide auxiliary survival rates. Hence, any efficiency gain for the GMM method was spurious. In comparison, even in this more complicated case, our AGMM estimator could still perform as well as the oracle procedure, similar to the simulation results in the first setting. Under setting 3 (Table 4), similar results can be observed as before.

To evaluate the performances of risk predication, we also presented the estimates of the baseline cumulative hazard function under settings 1-3. We summarized the results for estimating $\Lambda_0(t)$ at four pre-specified points $t \in \{0.3, 0.6, 0.9, 1.2\}$ in Table 5, where 0.6 is roughly the median and 1.2 is the 90th percentile of the survival time. For GMM and AGMM methods, we calculated the relative efficiency (RE), which was computed as the ratio of the RMSE from either method over that from the LY method. These results indicated that the AGMM method can further improve the efficiency for the estimation of the nonparametric component.

In summary, our proposed AGMM maintains a steady performance under all the settings considered in this simulation study. It is desirable to implement this approach to achieve consistent and efficient estimation.

4.2 | Auxiliary information from a moderate scale study: AIDS data

In the second part of our simulation studies, we apply the proposed approach to a data from the AIDS Clinical Trials Group protocol 175, a study that equally randomized 2139 patients to four antiretroviral regimens.[53] We fitted an additive

TABLE 2 Simulation results for regression coefficients under setting 1

		β_1						β_2					
CR	**Method**	**Bias**	**SD**	**SE**	**RMSE**	**CP**	**RE**	**Bias**	**SD**	**SE**	**RMSE**	**CP**	**RE**
Sample Size = 300													
30%	Oracle	0.002	0.0311	0.0289	0.0311	0.920	—	−0.002	0.0173	0.0167	0.0175	0.927	—
	LY	0.001	0.3134	0.3304	0.3128	0.970	10.1	0.018	0.1850	0.1997	0.1856	0.963	10.6
	EL	0.768	0.3144	0.0348	0.8292	0.023	26.7	−0.414	0.1174	0.0196	0.4302	0.000	24.6
	GMM	0.795	0.1061	0.0316	0.8020	0.000	25.8	−0.481	0.0778	0.0187	0.4870	0.000	27.9
	AGMM	0.005	0.0320	0.0292	0.0323	0.930	1.0	−0.005	0.0176	0.0168	0.0182	0.940	1.0
50%	Oracle	0.002	0.0291	0.0290	0.0292	0.943	—	−0.003	0.0161	0.0167	0.0163	0.960	—
	LY	−0.011	0.3462	0.3461	0.3458	0.947	11.9	0.004	0.2094	0.2071	0.2091	0.973	12.8
	EL	0.770	0.2996	0.0354	0.8264	0.003	28.3	−0.419	0.1260	0.0198	0.4371	0.000	26.7
	GMM	0.803	0.0745	0.0315	0.8062	0.000	27.6	−0.485	0.0664	0.0189	0.4900	0.000	30.0
	AGMM	0.005	0.0320	0.0299	0.0323	0.927	1.1	−0.005	0.0183	0.0179	0.0189	0.940	1.2
Sample Size = 500													
30%	Oracle	0.000	0.0214	0.0221	0.0214	0.973	—	−0.003	0.0119	0.0129	0.0122	0.960	—
	LY	−0.005	0.2452	0.2555	0.2449	0.960	11.5	−0.016	0.1502	0.1533	0.1508	0.960	12.4
	EL	0.714	0.3527	0.0265	0.7963	0.013	37.3	−0.406	0.1269	0.0150	0.4253	0.000	35.0
	GMM	0.797	0.0646	0.0244	0.8000	0.000	37.4	−0.479	0.0648	0.0148	0.4835	0.000	39.8
	AGMM	0.002	0.0226	0.0222	0.0227	0.960	1.1	−0.004	0.0126	0.0129	0.0132	0.950	1.1
50%	Oracle	−0.001	0.0230	0.0220	0.0230	0.933	—	−0.001	0.0136	0.0128	0.0137	0.943	—
	LY	−0.008	0.2698	0.2667	0.2695	0.957	11.7	−0.014	0.1601	0.1596	0.1604	0.953	11.7
	EL	0.770	0.2806	0.0272	0.8197	0.003	35.6	−0.431	0.1009	0.0151	0.4426	0.000	32.4
	GMM	0.795	0.0874	0.0246	0.7999	0.000	34.8	−0.483	0.0683	0.0147	0.4875	0.000	35.7
	AGMM	0.002	0.0248	0.0223	0.0248	0.903	1.1	−0.003	0.0145	0.0130	0.0149	0.920	1.1

Note: Oracle, the Oracle GMM estimator which use comparable auxiliary survival rates only; LY, Lin and Ying (1994)'s estimator; GMM, the GMM estimator which uses available auxiliary survival rates directly; AGMM, our proposed estimator which uses available auxiliary survival rates adaptively by penalization techniques. CR, the censoring rates; Bias, the average point estimates subtracted by the true values; SD, the sample standard deviation; SE, the average estimated standard errors; RMSE, the average square root of mean squared errors; CP, 95% level empirical coverage rate; RE, the RMSE from each of the three methods divided by that of the Oracle GMM method.

risk model to study the effects of two regressors: Z_1, the baseline CD4 count (divided by 100) on the first time a patient had a decline of CD4 T cell to at least 50%, which indicates a progression to AIDS or death; Z_2, the treatment (taking a value of 0 if the therapy is the monotherapy and 1 for the other combined therapies). Following Huang et al[11] and Han et al,[16] we randomly select a subset of 300 patients from the complete dataset as the internal individual level dataset for our analysis. For 2139 patients in the complete dataset, the survival time ranges from 0.04 to 3.37 years with median 2.73 years, and subjects who receive the monotherapy account for 25% of the sample. The minimum, median, and maximum of CD4 T cell count at baseline are 0, 340, and 1199, respectively. For the cases in the individual level dataset, the range of the survival time in years is [0.12, 3.35], close to that of the complete dataset. The median is 2.73 years and the percentage of people who receive the monotherapy is 24%. The censoring rates for the complete and individual level datasets are 75.64% and 75.00%, respectively.

We cross-classified the patients into 4 subgroups according to the value of Z_1 dichotomized at 3.4, the median of the baseline CD4 count and the value of Z_2. The four subgroups were given as $\Omega_1 = \{(Z_1, Z_2) : Z_1 < 3.4, Z_2 = 1\}$, $\Omega_2 = \{(Z_1, Z_2) : Z_1 \geq 3.4, Z_2 = 1\}$, $\Omega_3 = \{(Z_1, Z_2) : Z_1 < 3.4, Z_2 = 0\}$ and $\Omega_4 = \{(Z_1, Z_2) : Z_1 \geq 3.4, Z_2 = 0\}$. The resulting survival probabilities at the median survival time 2.73-year in the four subgroups, according to the Kaplan-Meier estimator based on the complete dataset, were 0.6428, 0.8332, 0.4636 and 0.6853, respectively. We treated these probabilities as known auxiliary information for illustrating our method. Table 6 displays the point estimates (Est), the estimated

TABLE 3 Simulation results for regression coefficients under setting 2

		β_1						β_2					
CR	**Method**	**Bias**	**SD**	**SE**	**RMSE**	**CP**	**RE**	**Bias**	**SD**	**SE**	**RMSE**	**CP**	**RE**
Sample size = 300													
30%	Oracle	0.001	0.0175	0.0166	0.0175	0.937	—	0.000	0.0102	0.0096	0.0102	0.920	—
	LY	0.010	0.3420	0.3314	0.3416	0.940	19.5	0.000	0.2058	0.1996	0.2054	0.970	20.1
	GMM	0.608	0.3481	0.0339	0.7000	0.060	39.9	−0.346	0.1756	0.0195	0.3882	0.043	38.0
	AGMM	0.007	0.0187	0.0168	0.0200	0.907	1.1	−0.003	0.0102	0.0097	0.0107	0.947	1.0
50%	Oracle	0.000	0.0162	0.0166	0.0162	0.953	—	0.001	0.0099	0.0096	0.0099	0.920	—
	LY	0.062	0.3388	0.3467	0.3439	0.973	21.2	0.005	0.2040	0.2078	0.2038	0.970	20.5
	GMM	0.641	0.3415	0.0331	0.7261	0.053	44.9	−0.368	0.1794	0.0190	0.4089	0.033	41.1
	AGMM	0.006	0.0187	0.0167	0.0195	0.917	1.2	−0.002	0.0113	0.0097	0.0116	0.910	1.2
Sample size = 500													
30%	Oracle	0.000	0.0129	0.0128	0.0128	0.957	—	0.001	0.0072	0.0074	0.0072	0.957	—
	LY	−0.027	0.2440	0.2550	0.2451	0.967	19.1	−0.005	0.1433	0.1535	0.1432	0.977	19.8
	GMM	0.494	0.3322	0.0264	0.5947	0.047	46.3	−0.286	0.1696	0.0154	0.3325	0.047	45.9
	AGMM	0.004	0.0134	0.0129	0.0140	0.930	1.1	−0.002	0.0078	0.0078	0.0079	0.943	1.1
50%	Oracle	0.000	0.0128	0.0128	0.0128	0.950	—	0.001	0.0077	0.0074	0.0077	0.950	—
	LY	−0.004	0.2612	0.2676	0.2608	0.953	20.4	−0.008	0.1698	0.1602	0.1697	0.937	22.0
	GMM	0.553	0.3474	0.0261	0.6529	0.037	51.2	−0.318	0.1813	0.0153	0.3661	0.047	47.6
	AGMM	0.003	0.0130	0.0128	0.0133	0.947	1.0	−0.001	0.0073	0.0074	0.0073	0.953	0.9

Note: The interpretations for CR, Bias, SD, SE, RMSE, CP, RE, Orace, LY, GMM and AGMM are the same as for Table 2.

standard errors (SE) and *P*-values for the fitted additive risk model. The estimation results of the complete dataset using classical LY estimator were also reported as a reference (oracle). While all procedures produced similar results for the regression parameter estimates, a large difference was observed in the standard errors. We can see that the LY estimator was the least efficient since it did not account for information from external data. Furthermore, we find that the estimation results for GMM, EL and AGMM methods are quite similar. The results are reasonable since the comparability assumption is satisfied for this randomly generated sample.

Our proposed AGMM method detected a $\hat{\tau}_3 = 0.085$ discrepancy for the auxiliary survival probability for the third group Ω_3, leading to a slight efficiency loss. This may be because the external data in this example is only of moderate size, resulting in a noticeable sampling variability.[28] Our method seems to be rather sensitive to such sampling non-comparability. It may be more robust than other methods compared in this section. However, more theoretical studies on how to account for the external information with uncertainty are needed, but beyond the scope of this work.

For the purpose of illustration, we next changed the observed values of auxiliary information manually to simulate a heterogeneous situation. Specifically, we also fitted EL, GMM, and AGMM methods when the survival rate ϕ_1 of the first subgroup Ω_1 was intentionally replaced by a different value 0.8428, that is, there was a roughly 0.2 discrepancy from the true survival rate ϕ_1^*. Table 6 reported our estimated regression results denoted by EL_H, GMM_H and AGMM_H. When heterogeneous auxiliary information was present, the estimation results from the EL and GMM methods could be quite inconsistent, for example, the sign for CD40 was even reversed. On the other hand, the AGMM method can correctly quantify the incomparability of the survival rate for the first subgroup to be 0.14 and lead to satisfactory results.

5 | REAL CASE STUDY

Neoplasms of the lungs are the leading cause of cancer incidence and mortality worldwide.[54] There are many well acknowledged risk factors of lung cancer, such as air pollution, personal characteristics and genetics. A leading cause is

TABLE 4 Simulation results for regression coefficients under setting 3

		β_1						β_2					
CR	**Method**	**Bias**	**SD**	**SE**	**RMSE**	**CP**	**RE**	**Bias**	**SD**	**SE**	**RMSE**	**CP**	**RE**
Sample size = 300													
30%	Oracle	0.000	0.0180	0.0183	0.0179	0.950	—	−0.002	0.0106	0.0106	0.0107	0.967	—
	LY	0.002	0.3442	0.3318	0.3437	0.937	19.2	0.001	0.2074	0.1996	0.2071	0.967	19.3
	GMM	−0.344	1.7675	0.2191	1.7978	0.140	100.2	0.217	0.7886	0.0959	0.8166	0.127	76.2
	AGMM	0.001	0.0179	0.0183	0.0180	0.963	1.0	−0.003	0.0107	0.0107	0.0110	0.963	1.0
50%	Oracle	−0.001	0.0186	0.0182	0.0186	0.943	—	−0.002	0.0107	0.0106	0.0109	0.940	—
	LY	0.056	0.3394	0.3467	0.3435	0.973	18.5	0.008	0.2053	0.2080	0.2051	0.967	18.8
	GMM	−0.183	1.4399	0.1802	1.4491	0.143	77.9	0.220	0.6153	0.0656	0.6527	0.187	59.9
	AGMM	0.002	0.0199	0.0184	0.0199	0.940	1.1	−0.003	0.0120	0.0107	0.0125	0.923	1.1
Sample size = 500													
30%	Oracle	−0.002	0.0146	0.0140	0.0147	0.933	—	−0.001	0.0080	0.0082	0.0081	0.963	—
	LY	−0.033	0.2456	0.2548	0.2474	0.963	16.8	−0.008	0.1453	0.1534	0.1453	0.967	18.0
	GMM	−0.652	1.9270	0.2161	2.0312	0.090	138.3	0.257	0.8278	0.0739	0.8653	0.137	107.3
	AGMM	−0.001	0.0149	0.0141	0.0149	0.933	1.0	−0.002	0.0081	0.0082	0.0082	0.953	1.0
50%	Oracle	−0.001	0.0151	0.0141	0.0151	0.923	—	−0.002	0.0081	0.0082	0.0084	0.970	—
	LY	−0.013	0.2725	0.2654	0.2724	0.923	18.0	0.008	0.1596	0.1605	0.1596	0.963	19.0
	GMM	−0.393	1.2937	0.1344	1.3498	0.050	89.4	0.248	0.5729	0.0624	0.6234	0.150	74.3
	AGMM	0.000	0.0149	0.0142	0.0149	0.933	1.0	−0.003	0.0081	0.0083	0.0086	0.960	1.0

Note: The interpretations for CR, Bias, SD, SE, RMSE, CP, RE, Orace, LY, GMM and AGMM are the same as for Table 2.

tobacco smoking, which accounts for more than 80% of lung cancer death.[55,56] We studied the association between a set of clinical covariates and the survival of a lung adenocarcinoma cohort (LUAD) available on The Cancer Genome Atlas (TCGA), which is a landmark cancer genomics program offering various kinds of data such as clinical data, mRNA, and copy numbers.[57,58] For the purpose of illustrating our methodology, we only focused on clinical data in this paper, so that diagnosis age, sex and smoking history were taken into account. We used the TCGA GDC API to download the latest clinical follow-up information. All data were collected on November 8, 2021 from the website of National Cancer Institute (https://cancergenome.nih.gov/). We focused on patients whose smoking history is clear, that is, we exclude any subject whose smoking history has not been documented or duration of quitting is unknown. After further removing patients who had missing information regarding risk factors of interest, the internal data included 484 lung adenocarcinoma patients.

The observed survival time ranged from 0.12 to 18.66 years with mean 1.44, and males account for 45.9%. The minimum, median, and maximum age of patients at initial pathologic diagnosis were 33, 66, and 88, respectively. For smoking, we categorized patients into two main groups: smoking = 1 represents non-smokers who have never smoked during lifetime or have reformed for greater than 15 years, and smoking = 0 represents the remaining patients. Non-smokers accounted for 40.9% of the sample. The censoring rate was approximately 77%.

We analyzed the data using the additive risk model (1) with three covariates age, sex and smoking. Table 7 summarizes the estimated regression coefficients using classical Lin-Ying estimator.[33] It appears that older age, female and smoking behavior were associated with a higher mortality risk. However, although the signs of these parameter estimates are consistent with medical literature, none of these risk factors was identified as statistically significant. Such results may be due to inefficiency since the sample size is only moderate for this study.

To improve efficiency in the estimation of the additive risk model, we sought to borrow information from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program, which collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 30% of the US population. As reported by SEER using patients from 2011-2017 (https://seer.cancer.gov/explorer/): the 5 year survival rates among lung adenocarcinoma patients were 0.232 for male Ω_1 and 0.332 for female Ω_2, and were 0.278, 0.289, 0.309, and 0.252 for

TABLE 5 Simulation results for the baseline distribution function under different settings

		LY			GMM				AGMM			
CR	$\Lambda_0(t)$	**Bias**	**SD**	**RMSE**	**Bias**	**SD**	**RMSE**	**RE**	**Bias**	**SD**	**RMSE**	**RE**
Setting 1												
30%	$\Lambda_0(0.3)$	0.004	0.0538	0.0538	−0.042	0.0312	0.0521	1.0	0.000	0.0310	0.0309	0.6
	$\Lambda_0(0.6)$	0.001	0.1099	0.1097	−0.101	0.0674	0.1214	1.1	−0.013	0.0538	0.0552	0.5
	$\Lambda_0(0.9)$	0.003	0.1724	0.1722	−0.161	0.1165	0.1984	1.2	−0.010	0.1017	0.1020	0.6
	$\Lambda_0(1.2)$	0.017	0.2548	0.2550	−0.215	0.2055	0.2972	1.2	−0.002	0.1923	0.1920	0.8
50%	$\Lambda_0(0.3)$	0.006	0.0581	0.0583	−0.044	0.0307	0.0535	0.9	−0.002	0.0302	0.0302	0.5
	$\Lambda_0(0.6)$	0.004	0.1237	0.1235	−0.105	0.0731	0.1277	1.0	−0.020	0.0602	0.0633	0.5
	$\Lambda_0(0.9)$	0.009	0.1821	0.1820	−0.164	0.1335	0.2115	1.2	−0.011	0.1239	0.1242	0.7
	$\Lambda_0(1.2)$	−0.012	0.3459	0.3455	−0.246	0.3025	0.3895	1.1	−0.036	0.2940	0.2958	0.9
Setting 2												
30%	$\Lambda_0(0.3)$	0.011	0.0578	0.0589	−0.034	0.0385	0.0510	0.9	−0.001	0.0289	0.0289	0.5
	$\Lambda_0(0.6)$	0.016	0.1157	0.1166	−0.084	0.0790	0.1153	1.0	−0.005	0.0494	0.0495	0.4
	$\Lambda_0(0.9)$	0.032	0.1822	0.1846	−0.127	0.1461	0.1932	1.0	0.001	0.1042	0.1040	0.6
	$\Lambda_0(1.2)$	0.038	0.2864	0.2884	−0.183	0.2498	0.3094	1.1	0.003	0.1963	0.1960	0.7
50%	$\Lambda_0(0.3)$	0.005	0.0607	0.0607	−0.033	0.0409	0.0527	0.9	−0.001	0.0292	0.0292	0.5
	$\Lambda_0(0.6)$	−0.007	0.1285	0.1285	−0.094	0.0899	0.1300	1.0	−0.017	0.0521	0.0546	0.4
	$\Lambda_0(0.9)$	−0.008	0.1992	0.1990	−0.145	0.1574	0.2139	1.1	−0.017	0.1241	0.1251	0.6
	$\Lambda_0(1.2)$	−0.009	0.3879	0.3874	−0.195	0.3623	0.4110	1.1	−0.042	0.3225	0.3247	0.8
Setting 3												
30%	$\Lambda_0(0.3)$	0.003	0.0566	0.0566	−0.037	0.0351	0.0511	0.9	−0.007	0.0296	0.0303	0.5
	$\Lambda_0(0.6)$	0.000	0.1158	0.1156	−0.088	0.0801	0.1192	1.0	−0.020	0.0505	0.0543	0.5
	$\Lambda_0(0.9)$	0.017	0.1809	0.1814	−0.122	0.1295	0.1776	1.0	−0.018	0.0943	0.0958	0.5
	$\Lambda_0(1.2)$	0.026	0.2685	0.2693	−0.166	0.2316	0.2848	1.1	−0.018	0.2015	0.2021	0.8
50%	$\Lambda_0(0.3)$	0.002	0.0549	0.0549	−0.036	0.0351	0.0502	0.9	−0.007	0.0301	0.0309	0.6
	$\Lambda_0(0.6)$	0.004	0.1132	0.1131	−0.078	0.0803	0.1120	1.0	−0.020	0.0508	0.0546	0.5
	$\Lambda_0(0.9)$	0.017	0.1917	0.1921	−0.114	0.1418	0.1818	0.9	−0.019	0.1224	0.1237	0.6
	$\Lambda_0(1.2)$	0.027	0.3621	0.3625	−0.147	0.3329	0.3635	1.0	−0.003	0.3218	0.3213	0.9

Note: The interpretations for CR, Bias, SD, RMSE, LY, GMM and AGMM are the same as for Table 2; RE, the RMSE from each of the two methods divided by that of the LY method

patients aged between <50, 50–64, 65-74 and >74 ($\Omega_3, \ldots, \Omega_6$), respectively. We consider adopting survival information from these 6 groups. The Sargan-Hansen J-test was used to check the conformity assumption of such external information. The null hypothesis was rejected at a type I error rate of 5% ($p = 1.9 \times 10^{-9}$), suggesting that part of subgroups where the auxiliary survival information came from were highly incomparable with the individual level data. The regression results of EL and GMM methods ignoring the heterogeneity are included in Table 7 too. By incorporating the external information, the EL and GMM estimators obtained a substantial efficiency gain in estimating the covariate effects. However, the estimation results are suggesting quite unreasonable direction for the estimated regression coefficients. For example, the positive sign of smoking indicated that non-smokers have a higher mortality risk, and similarly, the negative sign of age indicating that the elders could survival better. These results were misleading and the seemingly gains in efficiency became rather meaningless.

Finally, we applied the proposed AGMM methods to synthesize the auxiliary subgroup survival information from heterogeneous subgroups. Specifically, we first identified a possible range for the turning parameter λ_n and then used a

TABLE 6 Estimation results for the AIDS simulated data

	cd40			treat		
Method	**Est**	**SE**	***P*-value**	**Est**	**SE**	***P*-value**
Complete	−0.0360	0.0039	<0.0001	−0.0759	0.0124	<0.0001
LY	−0.0318	0.0088	0.0003	−0.0408	0.0315	0.1953
EL	−0.0497	0.0023	<0.0001	−0.0787	0.0060	<0.0001
GMM	−0.0480	0.0025	<0.0001	−0.0771	0.0061	<0.0001
AGMM	−0.0450	0.0025	<0.0001	−0.0651	0.0064	<0.0001
EL_H	0.0020	0.0001	<0.0001	−0.0719	0.0003	<0.0001
GMM_H	0.0021	0.0001	<0.0001	−0.0716	0.0003	<0.0001
$AGMM_H$	−0.0309	0.0097	0.0014	−0.0682	0.0042	<0.0001

TABLE 7 Estimation results for the TCGA lung cancer data with auxiliary information from SEER

	Age			Sex			Smoking		
Method	**Est**	**SE**	***P*-value**	**Est**	**SE**	***P*-value**	**Est**	**SE**	***P*-value**
LY	0.0029	0.0018	0.1053	−0.0112	0.0324	0.7306	−0.0008	0.0377	0.9823
EL	−0.0003	0.0004	0.4739	0.0711	0.0015	<0.0001	0.0200	0.0261	0.4436
GMM	−0.0010	0.0005	0.0431	0.0750	0.0041	<0.0001	0.0921	0.0305	0.0025
AGMM	0.0048	0.0008	<0.0001	−0.0228	0.0090	0.0109	−0.0264	0.0371	0.4762

grid search for the minimizer of BIC within the range. Our proposed method identified Ω_1, Ω_5 and Ω_6 as heterogeneous subgroups, and the other three as heterogeneous ones. The estimated values of τ_2, τ_3 and τ_4 by our proposed method were 0.13, 0.27, and 0.10, respectively. As shown in Table 7, the estimated covariate effects using AGMM were of the same direction as the LY method. We can see that age and sex became statistically significant predictors for survival after applying the proposed estimator, reflecting the efficiency gain clearly in this case. Although smoking was still not significant, its *P*-value was relatively smaller than that from LY method. This real case study demonstrated the clinical value of utilizing auxiliary information to produce more efficient estimation results in cancer survival research.

6 | DISCUSSION

The present study combines internal individual-level data with auxiliary subgroup survival probabilities that may be discordant under the additive risk model. With the addition of nuisance parameters to characterize the potential biases between subgroups of internal and external studies, we can better synthesize the information in our survival analysis. The extensive numerical studies in this paper show promising performance of the proposed AGMM estimator. All of the numerical work in this study was implemented on R, and we have created an R package, ARAuxSP, which is publicly available (http://github.com/Stat-WangXG/ARAuxSP). We remark here that in addition to the violation of comparability assumption, another complicated problem is the variability or non-negligible uncertainty in the auxiliary information.[26,27] We have assumed external information is obtained without variability in this paper but many medical studies (eg. AIDS data) may not be large enough in real applications. A possible solution is to consider the joint likelihood approach,[28] which treats the auxiliary information as observed random variable from their asymptotic distribution. We will investigate such a non-trivial issue in a future project. Instead of utilizing auxiliary statistics in the form of survival probabilities, there may be other aggregated information such as hazard ratios and correlations, among others. Our methods can be modified to incorporate such parameters as well. These problems will be addressed in further research.

ACKNOWLEDGEMENTS

The authors are grateful to the Joint Editor, Associate Editor and reviewers for their constructive suggestions and insightful comments. Xiaoguang Wang's work is partially supported by Dalian High-level Talent Innovation Project

(Grant 2020RD09). Jialiang Li's work is partially supported by Academic Research Funds A-8000016-0-0 in Singapore. Ian McKeague's work is partially supported by NSF Grant DMS-2112938 and NIH Grant R01AG062401. The authors acknowledge the efforts of The Cancer Genome Atlas (TCGA) Program of the National Cancer Institute in the creation of the TCGA database.

DATA AVAILABILITY STATEMENT

The dataset that supports the findings in this article is available from the website of National Cancer Institute at https://cancergenome.nih.gov/. The efforts of the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute in the creation of the SEER database are also acknowledged by the authors.

ORCID

Jie Ding https://orcid.org/0000-0002-6083-7529
Jialiang Li https://orcid.org/0000-0002-9704-4135
Xiaoguang Wang https://orcid.org/0000-0002-5598-062X

REFERENCES

1. Imbens GW, Lancaster T. Combining micro and macro data in microeconometric models. *Rev Econ Stud*. 1994;61(4):655-680.
2. Chaudhuri S, Handcock MS, Rendall MS. Generalized linear models incorporating population level information: an empirical likelihood based approach. *J R Stat Soc Ser B*. 2008;70(2):311-328.
3. Chen J, Sitter RR. A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Stat Sin*. 1999;9(2):385-406.
4. 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.
5. Kundu P, Tang R, Chatterjee N. Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika*. 2019;106(3):567-585.
6. Liu D, Zheng Y, Prentice RL, Hsu L. Estimating risk with time-to-event data: an application to the Women's health initiative. *J Am Stat Assoc*. 2014;109(506):514-524.
7. 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.
8. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons; 1980.
9. Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer-Verlag; 2003.
10. Li J, Ma S. *Survival Analysis in Medicine and Genetics*. New York: Chapman ς Hall CRC Press; 2013.
11. 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.
12. Cox DR. Regression models and life tables (with discussion). *J R Stat Soc Ser B*. 1972;34(2):187-202.
13. Aalen OO. A linear regression model for the analysis of life times. *Stat Med*. 1989;8(8):907-925.
14. Yakovlev AY, Tsodikov AD, Asselain B. *Stochastic Models of Tumor Latency and their Biostatistical Applications*. Singapore: World Scientific; 1996.
15. He J, Li H, Zhang S, Duan X. Additive hazards model with auxiliary subgroup survival information. *Lifetime Data Anal*. 2019;25(1):128-149.
16. Han B, Keilegom IV, Wang X. Semiparametric estimation of the nonmixture cure model with auxiliary survival information. *Biometrics*. 2022;78(2):448-459.
17. Owen AB. *Empirical Likelihood*. New York: Chapman and Hall/CRC; 2001.
18. Hansen LP. Large sample properties of generalized method of moments estimators. *Econometrica*. 1982;50(4):1029-1054.
19. Qin J, Lawless J. Empirical likelihood and general estimating equations. *Ann Stat*. 1994;22(1):300-325.
20. Chamberlain G. Asymptotic efficiency in estimation with conditional moment restrictions. *J Econ*. 1987;34(3):305-334.
21. Shang W, Wang X. The generalized moment estimation of the additive–multiplicative hazard model with auxiliary survival information. *Comput Stat Data Anal*. 2017;112:154-169.
22. Sheng Y, Sun Y, Deng D, Huang CY. Censored linear regression in the presence or absence of auxiliary survival information. *Biometrics*. 2020;76(3):734-745.
23. Han P, Lawless JF. Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Stat Sin*. 2019;29(3):1321-1342.
24. Chen Z, Ning J, Shen Y, Qin J. Combining primary cohort data with external aggregate information without assuming comparability. *Biometrics*. 2021;77(3):1024-1036.
25. Zhai Y, Han P. Data integration with oracle use of external information from heterogeneous populations. *J Comput Graph Stat*. 2022;31(4):1001-1012.
26. Huang C, Qin J. A unified approach for synthesizing population-level covariate effect information in semiparametric estimation with survival data. *Stat Med*. 2020;39(10):1573-1590.

27. Sheng Y, Sun Y, Huang CY, Kim MO. Synthesizing external aggregated information in the presence of population heterogeneity: a penalized empirical likelihood approach. *Biometrics*. 2022;78(2):679-690.
28. Zhang H, Deng L, Schiffman M, Qin J, Yu K. Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika*. 2020;107(3):689-703.
29. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B*. 1996;58(1):267-288.
30. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96(456):1348-1360.
31. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101(476):1418-1429.
32. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat*. 2010;38(2):894-942.
33. Lin DY, Ying Z. Semiparametric analysis of the additive risk model. *Biometrika*. 1994;81(1):61-71.
34. McKeague IW, Sasieni PD. A partly parametric additive risk model. *Biometrika*. 1994;81(3):501-514.
35. Lin D, Oakes D, Ying Z. Additive hazards regression with current status data. *Biometrika*. 1998;85(2):289-298.
36. Shen Y, Cheng SC. Confidence bands for cumulative incidence curves under the additive risk model. *Biometrics*. 1999;55(4):1093-1100.
37. Martinussen T, Scheike TH. Efficient estimation in additive hazards regression with current status data. *Biometrika*. 2002;89(3):649-658.
38. Wang X, Wang Z. EM algorithm for the additive risk mixture cure model with interval-censored data. *Lifetime Data Anal*. 2021;27(1):91-130.
39. Yin G, Cai J. Additive hazards model with multivariate failure time data. *Biometrika*. 2004;91(4):801-818.
40. Li H, Duan X, Yin G. Generalized method of moments for additive hazards model with clustered dental survival data. *Scand J Stat*. 2016;43(4):1124-1139.
41. Martinussen T, Scheike TH. Covariate selection for the semiparametric additive risk model. *Scand J Stat*. 2009;36(4):602-619.
42. Li J, Fine J, Brookhart A. Instrumental variable additive hazards model. *Biometrics*. 2015;71:122-130.
43. Jiang B, Li J, Fine J. On two-step residual inclusion estimator for instrument variable additive hazards model. *Biostat Epidemiol*. 2018;2:47-60.
44. Ying A, Xu R, Murphy J. Two-stage residual inclusion for survival data and competing risks—an instrumental variable approach with application to SEER-Medicare linked data. *Stat Med*. 2019;38:1775-1801.
45. Rava D, Xu R. Explained variation under the additive hazards model. *Stat Med*. 2021;40:85-100.
46. Hall AR. *Generalized Method of Moments*. New York: Oxford University Press; 2005.
47. Caner M. Lasso-type GMM estimator. *Economet Theor*. 2009;25:270-290.
48. Caner M, Han X, Lee Y. Adaptive elastic net GMM estimation with many invalid moment conditions: simultaneous model and moment selection. *J Bus Econ Stat*. 2018;36:24-46.
49. Xiao Z. Efficient GMM estimation with singular system of moment conditions. *Stat Theory Related Fields*. 2020;4:2475-2477.
50. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22.
51. Nelder JA, Mead R. A simplex method for function minimization. *Comput J*. 1965;7(4):308-313.
52. Andrews DWK, Lu B. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *J Econ*. 2001;101(1):123-164.
53. Hammer SM, Katzenstein DA, Hughes MD, et al. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *N Engl J Med*. 1996;335(15):1081-1090.
54. Fitzmaurice C, Akinyemiju TF, Lami FHA, et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2017: a systematic analysis for the global burden of disease study. *JAMA Oncol*. 2018;3(4):524-548.
55. Furrukh M. Tobacco smoking and lung cancer: perception-changing facts. *Sultan Qaboos Univ Med J*. 2013;13(3):345-358.
56. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424.
57. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511:543-550.
58. Campbell JD, Alexandrov A, Kim J, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat Genet*. 2016;48(6):607-616.
59. Lin DY, Ying Z. Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *Ann Stat*. 1995;23(5):1712-1734.

How to cite this article: Ding J, Li J, Han Y, McKeague IW, Wang X. Fitting additive risk models using auxiliary information. *Statistics in Medicine*. 2023;42(6):894-916. doi: 10.1002/sim.9649

APPENDIX . THEORETICAL PROOFS

All the proofs (Theorem) are provided in this section.

The derivation of equation (8). Actually, we only need to do some tedious algebraic operations to prove equation (8). Using the fact that the matrix $\boldsymbol{W}_n$ is defined as a block diagonal matrix composed of $\boldsymbol{A}_n(\hat{\boldsymbol{\xi}}^1)^{-1}$ and $\boldsymbol{D}_n^{-1}$, and further, $\boldsymbol{A}_n(\hat{\boldsymbol{\xi}}^1)$

is a diagonal matrix whose diagonal elements are $\hat{\psi}_{\psi}^{(k)}, k = 1, \ldots, K$, we have

$$\begin{aligned} Q_n(\xi) &= \boldsymbol{U}_n(\xi)^T \boldsymbol{W}_n \boldsymbol{U}_n(\xi) + \lambda_n \sum_{k=1}^{K} |\tau_k|/|\tilde{\tau}_k| \\ &= \boldsymbol{\Phi}_n(\beta)^T \boldsymbol{D}_n^{-1} \boldsymbol{\Phi}_n(\beta) + \boldsymbol{\Psi}_n(\beta)^T \hat{\psi}_{\psi}(\hat{\xi}^1)^{-1} \boldsymbol{\Psi}_n(\beta) + \lambda_n \sum_{k=1}^{K} |\tau_k|/|\tilde{\tau}_k| \\ &= \boldsymbol{\Phi}_n(\beta)^T \boldsymbol{D}_n^{-1} \boldsymbol{\Phi}_n(\beta) + \sum_{k=1}^{K} \frac{1}{\hat{\psi}_{\psi}^{(k)}} \Psi^{(k)}(\xi)^2 + \lambda_n \sum_{k=1}^{K} |\tau_k|/|\tilde{\tau}_k| \\ &= \boldsymbol{\Phi}_n(\beta)^T \boldsymbol{D}_n^{-1} \boldsymbol{\Phi}_n(\beta) + \sum_{k=1}^{K} \left\{ (a_k(\boldsymbol{\beta}, \alpha) - b_k(\boldsymbol{\beta}, \alpha) \cdot \tau_k)^2 + \lambda_n |\tau_k|/|\tilde{\tau}_k| \right\}, \end{aligned}$$

where $a_k(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} \left[I(\boldsymbol{Z}_i \in \Omega_k) \exp\{-\alpha - \boldsymbol{\beta}^T \boldsymbol{Z}_i t^*\} - \phi_k \right] / (n^2 \hat{\psi}_{\psi}^{(k)})^{1/2}$ and $b_k(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} I(\boldsymbol{Z}_i \in \Omega_k)/(n^2 \hat{\psi}_{\psi}^{(k)})^{1/2}$ for $k = 1, \ldots, K$. Thus, the result is proved.

Next, we prove theorems 1 − 3. To complete the proofs, we need the following lemma.

Lemma 1. *Under the regularity assumptions (C1)-(C3), the following results hold*:

(i) $\sqrt{n}\boldsymbol{U}(\xi_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma})$ *where*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{pmatrix},$$

$\boldsymbol{A} = \text{diag}\left\{ \text{var}[\psi^{(1)}(\boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0, \boldsymbol{\tau}_0)], \ldots, \text{var}[\psi^{(K)}(\boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0, \boldsymbol{\tau}_0)] \right\}$, $\boldsymbol{D} = E\left[\frac{1}{n} \sum_{i=1}^{n} \int_0^{\infty} \{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t)\}^{\otimes 2} dN_i(t) \right]$ *and the blank part is zero. The matrix* $\boldsymbol{\Sigma}$ *is of full rank in probability.*

(ii) $\hat{\boldsymbol{\Sigma}}(\hat{\xi}^1)$ *converges in probability to* $\boldsymbol{\Sigma}$*, where* $\hat{\xi}^1$ *is a consistent estimator of* ξ_0 *specified in Section 3.*

Proof of Lemma 1. Denote the filtration as $\mathcal{F}_t = \sigma\{N_i(s), Y_i(s+), \boldsymbol{Z}_i, 0 \le s \le t, i = 1, \ldots, n\}$. Define $M_i(t, \boldsymbol{\beta}) = N_i(t) - \int_0^t Y_i(s)(\lambda_0(s) + \boldsymbol{\beta}^T \boldsymbol{Z}_i) ds$ for $i = 1, \ldots, n$. It is easy to see that $\{M_i(t, \boldsymbol{\beta}_0), i = 1, \ldots, n\}$ are all mean-zero martingales with respect to $\mathcal{F}_t$. From the definition of $\boldsymbol{\Psi}_n(\xi)$, the estimating equations related to auxiliary information, we can verify straightforwardly that $E\{\boldsymbol{\Psi}_n(\xi_0)\} = 0$ and $\text{cov}\{n^{1/2} \boldsymbol{\Psi}_n(\xi_0)\} = \boldsymbol{A}$. Further, by plugging the true value of regression coefficients $\boldsymbol{\beta}_0$ into the classical estimating equations defined in (2), we can obtain the following equivalent reformulation

$$\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0) = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\infty} \{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t)\} dM_i(t, \boldsymbol{\beta}_0).$$

Since $\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t)$ is a predictable stochastic process, the stochastic integral $I(t) = \sum_{i=1}^{n} \int_0^t \{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(s)\} dM_i(s, \boldsymbol{\beta}_0)$ is also a mean-zero martingale with respect to $\mathcal{F}_t$. Hence, we have $E\{\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)\} = 0$. The variance-covariance matrix for $n^{1/2} \boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)$ is $nE(\boldsymbol{\Phi}_n^T(\boldsymbol{\beta}_0) \boldsymbol{\Phi}_n(\boldsymbol{\beta}_0))$, and another straightforward calculation shows that

$$\text{cov}\left\{ n^{1/2} \boldsymbol{\Phi}_n(\boldsymbol{\beta}_0) \right\} = nE(\boldsymbol{\Phi}_n^T(\boldsymbol{\beta}_0) \boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)) = E\left[\frac{1}{n} \sum_{i=1}^{n} \int_0^{\infty} \{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t)\}^{\otimes 2} dN_i(t) \right] = \boldsymbol{D}.$$

Since $E\{\boldsymbol{\Psi}_n(\xi_0)\} = 0$ and $E\{\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)\} = 0$, we can see that

$$\text{cov}\{\boldsymbol{\Psi}_n(\xi_0), \boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)\} = E\left\{ \boldsymbol{\Psi}_n^T(\xi_0) \boldsymbol{\Phi}_n(\boldsymbol{\beta}_0) \right\} = E\left\{ \boldsymbol{\Psi}_n^T(\xi_0) E\{\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0) | \mathcal{F}_0\} \right\} = 0$$

where the second equality is due to the fact that the calculation of $\boldsymbol{\Psi}_n(\xi_0)$ defined in (5) only depends on the covariates $\{\boldsymbol{Z}_i, i = 1, \ldots, n\}$. Finally, using the standard counting process arguments and the martingale central limit theorem, we

can show that (i) in Lemma 1 holds. As for (ii), note that $\hat{\boldsymbol{\beta}}_{\mathrm{LY}}^T$ and $\hat{\alpha}_{\mathrm{LY}} = \hat{\Lambda}_0(t^*, \hat{\boldsymbol{\beta}}_{\mathrm{LY}})$ are classical estimators of $\boldsymbol{\beta}_0$ and $\Lambda_0(t^*)$, respectively, whose consistencies are established by Lin and Ying.[33] Also, it can be shown that $\tilde{\boldsymbol{\tau}}$ converges in probability to $\boldsymbol{\tau}_0$ by applying the law of large numbers and assumption (C2). Since $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\xi}_0) \xrightarrow{p} \boldsymbol{\Sigma}$, we obtain that $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\xi}}^1) \xrightarrow{p} \boldsymbol{\Sigma}$. ∎

Proof of Theorem 1. Below are the partial derivatives of the estimating functions

$$\begin{aligned}
\frac{\partial \Psi_n^{(k)}(\boldsymbol{\xi})}{\partial \boldsymbol{\beta}} &= -\frac{t^*}{n}\sum_{i=1}^n I(\boldsymbol{Z}_i \in \Omega_k)\exp\{-\alpha - \boldsymbol{\beta}^T\boldsymbol{Z}_i t^*\}\boldsymbol{Z}_i^T,\ k = 1, \ldots, K,\\
\frac{\partial \Psi_n^{(k)}(\boldsymbol{\xi})}{\partial \alpha} &= -\frac{1}{n}\sum_{i=1}^n I(\boldsymbol{Z}_i \in \Omega_k)\exp\{-\alpha - \boldsymbol{\beta}^T\boldsymbol{Z}_i t^*\},\ k = 1, \ldots, K,\\
\frac{\partial \boldsymbol{\Psi}_n(\boldsymbol{\xi})}{\partial \boldsymbol{\tau}} &= \mathrm{diag}\left\{-\frac{1}{n}\sum_{i=1}^n I(\boldsymbol{Z}_i \in \Omega_1), \ldots, -\frac{1}{n}\sum_{i=1}^n I(\boldsymbol{Z}_i \in \Omega_K)\right\},\\
\frac{\partial \boldsymbol{\Phi}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\xi}} &= \left(\frac{\partial \boldsymbol{\Phi}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \frac{\partial \boldsymbol{\Phi}_n(\boldsymbol{\beta})}{\partial \alpha}, \frac{\partial \boldsymbol{\Phi}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\tau}}\right)\\
&= \left(-\frac{1}{n}\sum_{i=1}^n \int_0^\infty Y_i(t)\{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t)\}\boldsymbol{Z}_i^T dt, \mathbf{0}, \mathbf{0}\right),
\end{aligned}$$

and thus

$$\frac{\partial \boldsymbol{U}_n(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \begin{pmatrix} \frac{\partial \boldsymbol{\Psi}_n(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} \\ \frac{\partial \boldsymbol{\Phi}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\xi}} \end{pmatrix} = \begin{pmatrix} \frac{\partial \boldsymbol{\Psi}_n(\boldsymbol{\xi})}{\partial \boldsymbol{\beta}} & \frac{\partial \boldsymbol{\Psi}_n(\boldsymbol{\xi})}{\partial \alpha} & \frac{\partial \boldsymbol{\Psi}_n(\boldsymbol{\xi})}{\partial \boldsymbol{\tau}} \\ \frac{\partial \boldsymbol{\Phi}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} & \frac{\partial \boldsymbol{\Phi}_n(\boldsymbol{\beta})}{\partial \alpha} & \frac{\partial \boldsymbol{\Phi}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\tau}} \end{pmatrix}.$$

Define $\boldsymbol{\psi}_\beta = E\left\{\frac{\partial \boldsymbol{\Psi}_n(\boldsymbol{\xi})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0}\right\}$, $\boldsymbol{\psi}_\alpha = E\left\{\frac{\partial \boldsymbol{\Psi}_n(\boldsymbol{\xi})}{\partial \alpha}|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0}\right\}$, $\boldsymbol{\psi}_{\boldsymbol{\tau}} = E\left\{\frac{\partial \boldsymbol{\Psi}_n(\boldsymbol{\xi})}{\partial \boldsymbol{\tau}}|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0}\right\}$, $\boldsymbol{B} = E\left\{\frac{\partial \boldsymbol{\Phi}_n(\boldsymbol{\xi})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0}\right\}$ and

$$\boldsymbol{\psi}_{\boldsymbol{\xi}} = \begin{pmatrix} \boldsymbol{\psi}_\beta & \boldsymbol{\psi}_\alpha & \boldsymbol{\psi}_{\boldsymbol{\tau}} \\ \boldsymbol{B} & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Then, due to the classical central limit theorem and the martingale central limit theorem, we know that $\boldsymbol{G}_n(\boldsymbol{\xi}_0) = \frac{\partial \boldsymbol{U}_n(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0}$ converges to $\boldsymbol{\psi}_{\boldsymbol{\xi}}$ in probability and consequently, $\boldsymbol{G}_n(\boldsymbol{\xi}_0) = O_p(1)$.

Consider the C-ball $B_n(C) = \{\boldsymbol{\xi} : \boldsymbol{\xi} = \boldsymbol{\xi}_0 + n^{-1/2}\boldsymbol{u}, ||\boldsymbol{u}|| \le C\}, C > 0$, and denote its boundary by $\partial B_n(C)$. It is sufficient to show that, for any given $\epsilon > 0$, there exists a large constant C so that

$$P\left\{\inf_{\boldsymbol{\xi}\in\partial B_n(C)} Q_n(\boldsymbol{\xi}) > Q_n(\boldsymbol{\xi}_0)\right\} \ge 1 - \epsilon. \tag{A1}$$

This implies that, with probability at least $1 - \epsilon$, there exists a local minimizer $\hat{\boldsymbol{\xi}}$ of $Q_n(\boldsymbol{\xi})$ in the ball $B_n(C)$. Hence, there exists a local miminizer such that $||\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0|| = O_p(n^{-1/2})$.

Denote $\boldsymbol{W}_n = \hat{\boldsymbol{\Sigma}}(\tilde{\boldsymbol{\xi}})^{-1}$ and $l_n(\boldsymbol{\xi}) = \boldsymbol{U}_n(\boldsymbol{\xi})^T\boldsymbol{W}_n\boldsymbol{U}_n(\boldsymbol{\xi})$. By the second-order Taylor expansion, we know that $\sqrt{n}\boldsymbol{U}_n(\boldsymbol{\xi}_0 + n^{-1/2}\boldsymbol{u}) = \sqrt{n}\boldsymbol{U}_n(\boldsymbol{\xi}_0) + \boldsymbol{G}_n(\boldsymbol{\xi}_0)\boldsymbol{u} + o_p(1)$. Let $\boldsymbol{u} = (u_1, \ldots, u_K)^T$. Hence, for any $\boldsymbol{\xi} \in \partial B_n(C)$, we have

$$\begin{aligned}
R_n(\boldsymbol{u}) &= n\{Q_n(\boldsymbol{\xi}_0 + n^{-1/2}\boldsymbol{u}) - Q_n(\boldsymbol{\xi}_0)\}\\
&= n\{l_n(\boldsymbol{\xi}_0 + n^{-1/2}\boldsymbol{u}) - l_n(\boldsymbol{\xi}_0)\} + n\lambda_n\sum_{k=1}^K(|\tau_{k0} + n^{-1/2}u_k| - |\tau_{k0}|)/|\tilde{\tau}_k|\\
&\ge n\{l_n(\boldsymbol{\xi}_0 + n^{-1/2}\boldsymbol{u}) - l_n(\boldsymbol{\xi}_0)\} + n\lambda_n\sum_{k=1}^s(|\tau_{k0} + n^{-1/2}u_k| - |\tau_{k0}|)/|\tilde{\tau}_k|\\
&\ge 2\sqrt{n}\boldsymbol{U}_n^T(\boldsymbol{\xi}_0)\boldsymbol{W}_n\boldsymbol{G}_n(\boldsymbol{\xi}_0)\boldsymbol{u} + \boldsymbol{u}^T\boldsymbol{G}_n^T(\boldsymbol{\xi}_0)\boldsymbol{W}_n\boldsymbol{G}_n(\boldsymbol{\xi}_0)\boldsymbol{u} - \sqrt{n}\lambda_n\sum_{k=1}^s|u_k|/|\tilde{\tau}_k| + o_p(1),
\end{aligned} \tag{A2}$$

where s is the number of nonzero components in $\boldsymbol{\tau}_0$. Since we have assumed that the estimator $\tilde{\boldsymbol{\tau}} = (\tilde{\tau}_1, \ldots, \tilde{\tau}_K)^T$ satisfies $||\tilde{\boldsymbol{\tau}} - \boldsymbol{\tau}_0|| = O_p(n^{-1/2})$, we have, for $1 \le k \le s$,

$$\frac{1}{|\tilde{\tau}_k|} = \frac{1}{|\tau_{k0}|} - \frac{\text{sign}(\tau_{k0})}{\tau_{k0}^2}(\tilde{\tau}_k - \tau_{k0}) + o_p(|\tilde{\tau}_k - \tau_{k0}|) = \frac{1}{|\tau_{k0}|} + \frac{O_p(1)}{\sqrt{n}}.$$

In addition, since $\sqrt{n}\lambda_n = O_p(1)$, we have

$$\sqrt{n}\lambda_n \sum_{k=1}^{s} |u_k|/|\tilde{\tau}_k| = \sqrt{n}\lambda_n \sum_{k=1}^{s} \left(\frac{|u_k|}{|\tau_{k0}|} + \frac{|u_k|}{\sqrt{n}} O_p(1) \right) \le C O_p(1).$$

Furthermore, we know that $\sqrt{n}\boldsymbol{U}_n(\boldsymbol{\xi}_0) = O_p(1)$, $\boldsymbol{W}_n = O_p(1)$ and $\boldsymbol{G}_n(\boldsymbol{\xi}_0) = O_p(1)$. Therefore in (A2), if we choose a sufficiently large C, the first term is of the order $CO_p(1)$. The second is of the order $C^2O_p(1)$. The first term and the third term are all dominated by the second term. Therefore, by choosing a sufficiently large constant C, we have $R_n(\boldsymbol{u}) > 0$ and consequently, (A1) holds, which completes the proof. ∎

Proof of Theorem 2. (i) Let $\boldsymbol{\xi}_{10} = (\boldsymbol{\beta}_0^T, \alpha_0, \boldsymbol{\tau}_{10}^T)^T$. We show that with probability tending to 1, for any given $\boldsymbol{\xi}_1 = (\boldsymbol{\beta}^T, \alpha, \boldsymbol{\tau}_1^T)^T$ satisfying that $||\boldsymbol{\xi}_1 - \boldsymbol{\xi}_{10}|| = O_p(n^{-1/2})$ and for any constant C,

$$Q_n((\boldsymbol{\xi}_1^T, \boldsymbol{0})^T) = \min_{||\boldsymbol{\tau}_2|| \le Cn^{-1/2}} Q_n((\boldsymbol{\xi}_1^T, \boldsymbol{\tau}_2)^T).$$

It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any given $\boldsymbol{\xi}_1$ satisfying that $||\boldsymbol{\xi}_1 - \boldsymbol{\xi}_{10}|| = O_p(n^{-1/2})$, and $||\boldsymbol{\tau}_2|| \le Cn^{-1/2}$, $\partial Q_n(\boldsymbol{\xi})/\partial\tau_k$ and τ_k have same signs for $k = s+1, \ldots, K$. For each $\boldsymbol{\xi}$ in a neighborhood of $\boldsymbol{\xi}_0$ and for $k = s+1, \ldots, K$, we have

$$\begin{aligned} \frac{\partial Q_n(\boldsymbol{\xi})}{\partial \tau_k} &= \frac{\partial l_n(\boldsymbol{\xi})}{\partial \tau_k} + \lambda_n \frac{\text{sign}(\hat{\tau}_k)}{|\tilde{\tau}_k|} \\ &= 2\boldsymbol{U}_n^T(\boldsymbol{\xi}) \boldsymbol{W}_n \frac{\partial \boldsymbol{U}_n(\boldsymbol{\xi})}{\partial \tau_k} + (n\lambda_n) n^{-1/2} \frac{\text{sign}(\hat{\tau}_k)}{|\sqrt{n}\tilde{\tau}_k|} \\ &= n^{-1/2} \left\{ O_p(1) + n\lambda_n \frac{\text{sign}(\hat{\tau}_k)}{|O_p(1)|} \right\}. \end{aligned} \tag{A3}$$

Since $n\lambda_n \to \infty$, the sign of $\partial Q_n(\boldsymbol{\xi})/\partial\tau_k$ in (A3) is completely determined by the sign of τ_k when n is large, and they always have same signs.

(ii) We show the asymptotic normality of $\boldsymbol{\xi}_1$ when there exists at least one k, such that, $\tau_k = 0$. Define $\boldsymbol{G}_n^1(\boldsymbol{\xi}_0) = \frac{\partial \boldsymbol{U}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}_1}|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0}$, $\boldsymbol{\psi}_{\boldsymbol{\tau}_1} = E\left\{ \frac{\partial \boldsymbol{\Psi}(\boldsymbol{\xi})}{\partial \boldsymbol{\tau}_1}|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0} \right\}$ and

$$\boldsymbol{\psi}_{\boldsymbol{\xi}_1} = \begin{pmatrix} \boldsymbol{\psi}_\beta & \boldsymbol{\psi}_\alpha & \boldsymbol{\psi}_{\boldsymbol{\tau}_1} \\ \boldsymbol{B} & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}.$$

We know that $\boldsymbol{G}_n^1(\boldsymbol{\xi}_0)$ converges to $\boldsymbol{\psi}_{\boldsymbol{\xi}_1}$ in probability. And based on the assumption (C3), we also have that $\frac{\partial^2 l_n(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}_1 \partial \boldsymbol{\xi}_1^T}|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0}$ converges to $\boldsymbol{\psi}_{\boldsymbol{\xi}_1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\psi}_{\boldsymbol{\xi}_1}$ in probability by Taylor's expansion. Before proceeding to the following discussion, we first prove that the matrix $\boldsymbol{\psi}_{\boldsymbol{\xi}_1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\psi}_{\boldsymbol{\xi}_1}$ is invertible in probability, which is equivalent to show that $\boldsymbol{\psi}_{\boldsymbol{\xi}_1}$ is a column full rank matrix in probability. From the assumption (C2), we see that $P(\boldsymbol{Z} \in \Omega_k) > 0$ for $k = 1, \ldots, K$. The survival probability of T given $\boldsymbol{Z}$ is equal to $\exp\{-\alpha - \boldsymbol{\beta}^T \boldsymbol{Z} t^*\}$ at t^*, so we have that $0 < \exp\{-\alpha - \boldsymbol{\beta}^T \boldsymbol{Z}_i t^*\} < 1$ for $i = 1, \ldots, n$. Thus the elements of $\boldsymbol{\psi}_\alpha$ and the diagonal elements of $\boldsymbol{\psi}_{\boldsymbol{\tau}}$ are less than zero in probability. Since we have assume that there exists at least one k, such that, $\tau_k = 0$, $(\boldsymbol{\psi}_\alpha, \boldsymbol{\psi}_{\boldsymbol{\tau}_1})$ is of full column rank. Then, together with the fact that $\boldsymbol{B}$ is a nonsingular matrix in probability, $\boldsymbol{\psi}_{\boldsymbol{\xi}_1}$ is of full column rank.

From the proof of Theorem 1 and the statement in (i), it is easy to show that there exists $\hat{\boldsymbol{\xi}}_1$ that is a $\sqrt{n}-$consistent local minimizer of $Q_n((\boldsymbol{\xi}_1^T, \boldsymbol{0})^T)$ satisfying

$$\boldsymbol{0} = \frac{\partial Q_n(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}_1}\Big|_{\boldsymbol{\xi}=(\hat{\boldsymbol{\xi}}_1^T, \boldsymbol{0}^T)^T}$$

$$= \frac{\partial l_n(\xi)}{\partial \xi_1}\Big|_{\xi=(\hat{\xi}_1^T,\mathbf{0}^T)^T} + \lambda_n\left(\mathbf{0}^T, \frac{\text{sign}(\hat{\tau}_1)}{|\tilde{\tau}_1|}, \dots, \frac{\text{sign}(\hat{\tau}_s)}{|\tilde{\tau}_s|}\right)^T$$
$$= 2\boldsymbol{G}_n^{1T}(\xi_0)\boldsymbol{W}_n\boldsymbol{U}_n(\xi_0) + \left(\frac{\partial^2 l_n(\xi)}{\partial \xi_1 \partial \xi_1^T}|_{\xi=\xi_0} + O_p(1)\right)(\hat{\xi}_1 - \xi_{10})$$
$$+ \lambda_n\left(\mathbf{0}^T, \frac{\text{sign}(\tau_{10})}{|\tilde{\tau}_1|}, \dots, \frac{\text{sign}(\tau_{s0})}{|\tilde{\tau}_s|}\right)^T.$$

The first two parts in the last equation are based on the Taylor's expansion around ξ_0 and the third part is implied by $\text{sign}(\hat{\tau}_k) = \text{sign}(\tau_{k0})$ when n is large, since $\hat{\tau}_k$ is a $\sqrt{n}-$consistent estimator of τ_{k0}. Furthermore, since $\sqrt{n}\lambda_n \to 0$, $\boldsymbol{W}_n$ converges in probability to $\boldsymbol{\Sigma}^{-1}$ and $\sqrt{n}\boldsymbol{U}_n(\xi) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$, we have the following equation

$$\sqrt{n}(\hat{\xi}_1 - \xi_{10}) = (\boldsymbol{\psi}_{\xi_1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\psi}_{\xi_1})^{-1} \boldsymbol{\psi}_{\xi_1}^T \boldsymbol{\Sigma}^{-1} (\sqrt{n}\boldsymbol{U}_n(\xi_0)) + o_p(1),$$

and as a direct result, we have

$$\sqrt{n}(\hat{\xi}_1 - \xi_{10}) \xrightarrow{d} N\left(\mathbf{0}, \left(\boldsymbol{\psi}_{\xi_1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\psi}_{\xi_1}\right)^{-1}\right).$$

By the definition of $\boldsymbol{\psi}_{\xi_1}$ and $\boldsymbol{\Sigma}$, we can express

$$\boldsymbol{\psi}_{\xi_1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\psi}_{\xi_1} = \begin{pmatrix} \boldsymbol{\psi}_\beta^T & \boldsymbol{B}^T \\ \boldsymbol{\psi}_\alpha^T & \mathbf{0}^T \\ \boldsymbol{\psi}_{\tau_1}^T & \mathbf{0}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{A}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\psi}_\beta & \boldsymbol{\psi}_\alpha & \boldsymbol{\psi}_{\tau_1} \\ \boldsymbol{B} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$
$$= \begin{pmatrix} \boldsymbol{B}^T\boldsymbol{D}^{-1}\boldsymbol{B} + \boldsymbol{\psi}_\beta^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_\beta & \boldsymbol{\psi}_\beta^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_\alpha & \boldsymbol{\psi}_\beta^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_{\tau_1} \\ \boldsymbol{\psi}_\alpha^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_\beta & \boldsymbol{\psi}_\alpha^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_\alpha & \boldsymbol{\psi}_\alpha^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_{\tau_1} \\ \boldsymbol{\psi}_{\tau_1}^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_\beta & \boldsymbol{\psi}_{\tau_1}^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_\alpha & \boldsymbol{\psi}_{\tau_1}^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_{\tau_1} \end{pmatrix}.$$

Since $\hat{\boldsymbol{\tau}}_1$ is the third part of $\hat{\boldsymbol{\xi}}_1 = (\hat{\boldsymbol{\beta}}^T, \hat{\alpha}, \hat{\boldsymbol{\tau}}_1^T)^T$, we know that

$$\sqrt{n}(\hat{\boldsymbol{\tau}}_1 - \boldsymbol{\tau}_{10}) \xrightarrow{d} N(\mathbf{0}, (\boldsymbol{M}_{22} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12})^{-1}),$$

where $\boldsymbol{M}_{11}$ is the 2×2 block element in the upper left corner of $\boldsymbol{\psi}_{\xi_1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\psi}_{\xi_1}$, $\boldsymbol{M}_{12}$ is the 2×1 block element in the upper right corner, $\boldsymbol{M}_{21}$ is the 1×2 block element in the lower left corner and $\boldsymbol{M}_{22}$ is the 1×1 block element in the lower right corner. Thus we complete the proof. ∎

Proof of Theorem 3. Since $\hat{\boldsymbol{\beta}}$ is the first part of $\hat{\boldsymbol{\xi}}_1 = (\hat{\boldsymbol{\beta}}^T, \hat{\alpha}, \hat{\boldsymbol{\tau}}_1^T)^T$, we can obtain its asymptotic covariance–variance matrix in a similar way to that of $\hat{\boldsymbol{\tau}}_1$. Define $\boldsymbol{J}_1 = \boldsymbol{\psi}_\beta^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_\alpha, \boldsymbol{J}_2 = \boldsymbol{\psi}_\beta^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_{\tau_1}, \boldsymbol{J} = (\boldsymbol{J}_1, \boldsymbol{J}_2)$ and

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{\psi}_\alpha^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_\alpha & \boldsymbol{\psi}_\alpha^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_{\tau_1} \\ \boldsymbol{\psi}_{\tau_1}^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_\alpha & \boldsymbol{\psi}_{\tau_1}^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_{\tau_1} \end{pmatrix} \triangleq \begin{pmatrix} V_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}.$$

Then, the asymptotic covariance–variance matrix of $\hat{\boldsymbol{\beta}}$ is actually the upper left corner of $(\boldsymbol{\psi}_{\xi_1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\psi}_{\xi_1})^{-1}$, which can be expressed as $(\boldsymbol{\Sigma}_{LY}^{-1} + \boldsymbol{\Sigma}_0)^{-1}$, where $\boldsymbol{\Sigma}_{LY} = (\boldsymbol{B}^T\boldsymbol{D}^{-1}\boldsymbol{B})^{-1}$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{\psi}_\beta^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_\beta - \boldsymbol{J}\boldsymbol{V}^{-1}\boldsymbol{J}^T$. It is indeed the asymptotic covariance–variance matrix of $\hat{\boldsymbol{\beta}}_{\text{LY}}$ in Lin and Ying.[33] Notice that if we define $\boldsymbol{\Sigma}_1 = \boldsymbol{\psi}_\beta^T\boldsymbol{A}^{-1}\boldsymbol{\psi}_\beta - \boldsymbol{J}_1 V_{11}^{-1}\boldsymbol{J}_1^T$, the asymptotic covariance–variance matrix of $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ is $(\boldsymbol{\Sigma}_{LY}^{-1} + \boldsymbol{\Sigma}_1)^{-1}$, which will be shown to be highly correlated with our proposed estimator $\hat{\boldsymbol{\beta}}$.

Next, we will show that $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0$ are all positive semi-definite matrices, and investigate the relationship between $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. By using inductive method, we can know that $\boldsymbol{H}_1 = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{\psi}_\alpha^T V_{11}^{-1}\boldsymbol{\psi}_\alpha\boldsymbol{A}^{-1}$ is a positive semi-definite

matrix. Then, since we have $\boldsymbol{\Sigma}_1 = \boldsymbol{\psi}_\beta^T \boldsymbol{H}_1 \boldsymbol{\psi}_\beta$, $\boldsymbol{\Sigma}_1$ is positive semi-definite as well. Our main goal is to prove the following equation

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\psi}_\beta^T \{\boldsymbol{H}_1 - \boldsymbol{H}_1 \boldsymbol{\psi}_{\tau_1}^T (\boldsymbol{\psi}_{\tau_1}^T \boldsymbol{H}_1 \boldsymbol{\psi}_{\tau_1})^{-1} \boldsymbol{\psi}_{\tau_1}^T \boldsymbol{H}_1\} \boldsymbol{\psi}_\beta. \tag{A4}$$

Note that

$$\begin{aligned}
\boldsymbol{\Sigma}_0 &= \boldsymbol{\psi}_\beta^T \boldsymbol{A}^{-1} \boldsymbol{\psi}_\beta - \boldsymbol{J} \boldsymbol{V}^{-1} \boldsymbol{J}^T \\
&= \boldsymbol{\psi}_\beta^T \left\{ \boldsymbol{A}^{-1} - \begin{pmatrix} \boldsymbol{A}^{-1}\boldsymbol{\psi}_\alpha & \boldsymbol{A}^{-1}\boldsymbol{\psi}_{\tau_1} \end{pmatrix} \begin{pmatrix} V_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\psi}_\alpha^T \boldsymbol{A}^{-1} \\ \boldsymbol{\psi}_{\tau_1}^T \boldsymbol{A}^{-1} \end{pmatrix} \right\} \boldsymbol{\psi}_\beta \\
&\overset{\Delta}{=} \boldsymbol{\psi}_\beta^T \boldsymbol{H}_0 \boldsymbol{\psi}_\beta.
\end{aligned}$$

Furthermore, define $\boldsymbol{P} = (\boldsymbol{V}_{21} - V_{11}^{-1}\boldsymbol{V}_{22}\boldsymbol{V}_{12})^{-1} = (\boldsymbol{\psi}_{\tau_1}^T \boldsymbol{H}_1 \boldsymbol{\psi}_{\tau_1})^{-1}$ and $\boldsymbol{C} = \boldsymbol{A}^{-1}\boldsymbol{\psi}_\alpha V_{11}^{-1} \boldsymbol{\psi}_\alpha^T \boldsymbol{A}^{-1}$, then we have

$$\begin{aligned}
\boldsymbol{H}_0 &= \boldsymbol{A}^{-1} - \begin{pmatrix} \boldsymbol{A}^{-1}\boldsymbol{\psi}_\alpha & \boldsymbol{A}^{-1}\boldsymbol{\psi}_{\tau_1} \end{pmatrix} \begin{pmatrix} V_{11}^{-1} + V_{11}^{-1}\boldsymbol{V}_{12}\boldsymbol{P}\boldsymbol{V}_{21}V_{11}^{-1} & -V_{11}^{-1}\boldsymbol{V}_{12}\boldsymbol{P} \\ -\boldsymbol{P}\boldsymbol{V}_{21}V_{11}^{-1} & \boldsymbol{P} \end{pmatrix} \begin{pmatrix} \boldsymbol{\psi}_\alpha^T \boldsymbol{A}^{-1} \\ \boldsymbol{\psi}_{\tau_1}^T \boldsymbol{A}^{-1} \end{pmatrix} \\
&= \boldsymbol{A}^{-1} - \boldsymbol{C} - \boldsymbol{C}\boldsymbol{\psi}_{\tau_1}\boldsymbol{P}\boldsymbol{\psi}_{\tau_1}^T\boldsymbol{C} + \boldsymbol{A}^{-1}\boldsymbol{\psi}_{\tau_1}\boldsymbol{P}\boldsymbol{\psi}_{\tau_1}^T\boldsymbol{C} + \boldsymbol{C}\boldsymbol{\psi}_{\tau_1}\boldsymbol{P}\boldsymbol{\psi}_{\tau_1}^T\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{\psi}_{\tau_1}\boldsymbol{P}\boldsymbol{\psi}_{\tau_1}^T\boldsymbol{A}^{-1} \\
&= \boldsymbol{H}_1 + \boldsymbol{H}_1\boldsymbol{\psi}_{\tau_1}\boldsymbol{P}\boldsymbol{\psi}_{\tau_1}^T\boldsymbol{C} - \boldsymbol{H}_1\boldsymbol{\psi}_{\tau_1}\boldsymbol{P}\boldsymbol{\psi}_{\tau_1}\boldsymbol{A}^{-1} \\
&= \boldsymbol{H}_1 - \boldsymbol{H}_1\boldsymbol{\psi}_{\tau_1}\boldsymbol{P}\boldsymbol{\psi}_{\tau_1}^T\boldsymbol{H}_1 \\
&= \boldsymbol{H}_1 - \boldsymbol{H}_1\boldsymbol{\psi}_{\tau_1}(\boldsymbol{\psi}_{\tau_1}^T\boldsymbol{H}_1\boldsymbol{\psi}_{\tau_1})^{-1}\boldsymbol{\psi}_{\tau_1}^T\boldsymbol{H}_1,
\end{aligned}$$

which confirms that equation (A4) is correct. Since H_1 is positive semi-definite, we can decompose it into two equally positive semi-definite matrices $\boldsymbol{H}_1 = \boldsymbol{H}_1^{\frac{1}{2}}\boldsymbol{H}_1^{\frac{1}{2}}$ and as a result, we have

$$\boldsymbol{H}_0 = \boldsymbol{H}_1^{\frac{1}{2}} \left\{ I - \boldsymbol{H}_1^{\frac{1}{2}}\boldsymbol{\psi}_{\tau_1}(\boldsymbol{\psi}_{\tau_1}^T\boldsymbol{H}_1\boldsymbol{\psi}_{\tau_1})^{-1}\boldsymbol{\psi}_{\tau_1}^T\boldsymbol{H}_1^{\frac{1}{2}} \right\} \boldsymbol{H}_1^{\frac{1}{2}}.$$

It is easy to show that the matrix in the brace is idempotent. Hence, H_0 is positive semi-definite, and $\boldsymbol{\Sigma}_0$ is also positive semi-definite. In addition, by $\boldsymbol{H}_0 - \boldsymbol{H}_1 = \boldsymbol{H}_1\boldsymbol{\psi}_{\tau_1}(\boldsymbol{\psi}_{\tau_1}^T\boldsymbol{H}_1\boldsymbol{\psi}_{\tau_1})^{-1}\boldsymbol{\psi}_{\tau_1}^T\boldsymbol{H}_1$, we know that $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0$ is also a positive semi-definite matrix.

In summary, we have shown that $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0$ are all positive semi-definite matrices and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, (\boldsymbol{\Sigma}_{LY}^{-1} + \boldsymbol{\Sigma}_0)^{-1}),$$

and thus, we complete the proof. ∎

Proof of Theorem 4. We can decompose $n^{1/2}(\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}) - \Lambda_0(t))$ as

$$n^{1/2}(\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}) - \hat{\Lambda}_0(t, \boldsymbol{\beta}_0)) + n^{1/2}(\hat{\Lambda}_0(t, \boldsymbol{\beta}_0) - \Lambda_0^*(t)) + n^{1/2}(\Lambda_0^*(t) - \Lambda_0(t)), \tag{A5}$$

where $\Lambda_0^*(t) = \int_0^t I(\sum_{i=1}^n Y_i(s) > 0)\lambda_0(s)ds$. Following the proof of Theorem 3.1 in Lin and Ying,[59] the third term in (A5) is asymptotically negligible. Furthermore, by using Taylor expansion and the asymptotic expression for $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, the first term becomes $C^T(t)\boldsymbol{\Sigma}_{\xi_1}(\sqrt{n}\boldsymbol{U}_n(\boldsymbol{\xi}_0)) + o_p(1)$, where $\boldsymbol{\Sigma}_{\xi_1} = \Pi(\boldsymbol{\psi}_{\xi_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\psi}_{\xi_1})^{-1}\boldsymbol{\psi}_{\xi_1}^T\boldsymbol{\Sigma}^{-1}$, Π is a projection matrix that extracts the first p rows of $(\boldsymbol{\psi}_{\xi_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\psi}_{\xi_1})^{-1}\boldsymbol{\psi}_{\xi_1}^T\boldsymbol{\Sigma}^{-1}$, and $C(t)$ is the probability limit of $-\int_0^t \overline{\boldsymbol{Z}}(u)du$. Consequently, expression $n^{1/2}(\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}) - \Lambda_0(t))$ can be reformulated into

$$n^{1/2}(\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}) - \Lambda_0(t)) = n^{1/2}\int_0^t \frac{\sum_{i=1}^n dM_i(s, \boldsymbol{\beta}_0)}{\sum_{j=1}^n Y_j(s)} + C^T(t)\boldsymbol{\Sigma}_{\xi_1} n^{1/2}\boldsymbol{U}_n(\boldsymbol{\xi}_0) + o_p(1). \tag{A6}$$

Next, a straightforward covariance calculation leads to

$$C_\Lambda(t_1,t_2)=g(\min(t_1,t_2))+C^T(t_1)(\boldsymbol{\Sigma}_{\mathrm{LY}}^{-1}+\boldsymbol{\Sigma}_0)^{-1}C(t_2)+C^T(t_1)\boldsymbol{\Sigma}_{\xi_1}\tilde{\Gamma}(t_2)+C^T(t_2)\boldsymbol{\Sigma}_{\xi_1}\tilde{\Gamma}(t_1),$$

where we denote $\tilde{\Gamma}(t)=(\mathbf{0}_K^T,\Gamma^T(t))^T$. We can calculate the covariance between $\hat{S}(t_1|\boldsymbol{Z},\hat{\boldsymbol{\beta}})$ and $\hat{S}(t_2|\boldsymbol{Z},\hat{\boldsymbol{\beta}})$ to be

$$\begin{aligned}
&\mathrm{cov}(\hat{S}(t_1|\boldsymbol{Z},\hat{\boldsymbol{\beta}}),\hat{S}(t_2|\boldsymbol{Z},\hat{\boldsymbol{\beta}}))\\
&\quad=S(t_1|\boldsymbol{Z})S(t_2|\boldsymbol{Z})E[\{\exp[-(\hat{\Lambda}_0(t_1,\hat{\boldsymbol{\beta}})-\Lambda_0(t_1))-t_1(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\boldsymbol{Z}]-1\}\\
&\qquad\times\{\exp[-(\hat{\Lambda}_0(t_2,\hat{\boldsymbol{\beta}})-\Lambda_0(t_2))-t_2(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\boldsymbol{Z}]-1\}]\\
&\quad=S(t_1|\boldsymbol{Z})S(t_2|\boldsymbol{Z})E[\{(\hat{\Lambda}_0(t_1,\hat{\boldsymbol{\beta}})-\Lambda_0(t_1))+t_1(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\boldsymbol{Z}\}\{(\hat{\Lambda}_0(t_2,\hat{\boldsymbol{\beta}})-\Lambda_0(t_2))+t_2(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\boldsymbol{Z}\}]+o_p(1),
\end{aligned}$$

where the second equality is due to that fact that e^x-1 and x are equivalent infinitesimal. Then, using (A6) and the asymptotic expression for $n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$, we can obtain the covariance function as

$$C_S(t_1,t_2)=S(t_1|\boldsymbol{Z})S(t_2|\boldsymbol{Z})\{g(\min(t_1,t_2))+D^T(t_1)(\boldsymbol{\Sigma}_{\mathrm{LY}}^{-1}+\boldsymbol{\Sigma}_0)^{-1}D(t_2)+D^T(t_1)\boldsymbol{\Sigma}_{\xi_1}\tilde{\Gamma}(t_2)+D^T(t_2)\boldsymbol{\Sigma}_{\xi_1}\tilde{\Gamma}(t_1)\},$$

where $D(t)=C(t)+t\boldsymbol{Z}$. The finite dimensional convergence of $n^{1/2}\{\hat{S}(t|\boldsymbol{Z},\hat{\boldsymbol{\beta}})-S(t|\boldsymbol{Z})\}$ and $n^{1/2}\{\hat{\Lambda}_0(t,\hat{\boldsymbol{\beta}})-\Lambda_0(t)\}$ to a multivariate normal with the above covariance functions can be established by applying Theorem 3 and the martingale central limit theorem. The standard counting process techniques can be employed to prove the tightness of the process. We omit the lengthy argument in this paper. ■