The Annals of Applied Statistics
2020, Vol. 14, No. 1, 51–73
https://doi.org/10.1214/19-AOAS1287

INTEGRATIVE SURVIVAL ANALYSIS WITH UNCERTAIN EVENT TIMES IN APPLICATION TO A SUICIDE RISK STUDY

BY WENJIE WANG[1,*], ROBERT ASELTINE[2], KUN CHEN[1,**] AND JUN YAN[1,†]

[1]*Department of Statistics, University of Connecticut,* [*]*wenjie.2.wang@uconn.edu;* [**]*kun.chen@uconn.edu;* [†]*jun.yan@uconn.edu*

[2]*Division of Behavioral Science and Community Health, Center for Population Health, UConn Health, aseltine@uchc.edu*

The concept of integrating data from disparate sources to accelerate scientific discovery has generated tremendous excitement in many fields. The potential benefits from data integration, however, may be compromised by the uncertainty due to incomplete/imperfect record linkage. Motivated by a suicide risk study, we propose an approach for analyzing survival data with uncertain event times arising from data integration. Specifically, in our problem deaths identified from the hospital discharge records together with reported suicidal deaths determined by the Office of Medical Examiner may still not include all the death events of patients, and the missing deaths can be recovered from a complete database of death records. Since the hospital discharge data can only be linked to the death record data by matching basic patient characteristics, a patient with a censored death time from the first dataset could be linked to multiple potential event records in the second dataset. We develop an integrative Cox proportional hazards regression in which the uncertainty in the matched event times is modeled probabilistically. The estimation procedure combines the ideas of profile likelihood and the expectation conditional maximization algorithm (ECM). Simulation studies demonstrate that under realistic settings of imperfect data linkage the proposed method outperforms several competing approaches including multiple imputation. A marginal screening analysis using the proposed integrative Cox model is performed to identify risk factors associated with death following suicide-related hospitalization in Connecticut. The identified diagnostics codes are consistent with existing literature and provide several new insights on suicide risk, prediction and prevention.

1. Introduction. In many fields of science, engineering and medicine, combining multiple datasets from disparate sources has made it possible to tackle important problems at an accelerated rate through integrative statistical learning. These datasets cover overlapped or interrelated measurements from individuals. In an ideal situation the multisource data should pertain to the same set of fully identified individuals. For example, in a cancer study multiplatform genetic data such as mRNA gene expression, DNA methylation and copy number variation are available from each patient (Zhao et al. (2015)); an integrative analysis then ensures a comprehensive coverage of genetic perspectives to understand the disease mechanism. In practice, however, more than often a unique identifier is not provided or does not even exist to link multi-source or multi-platform datasets. This gives rise to the so-called "data/record linkage" problem, that is, matching records from different sources that belong to the same person or entity based on available characteristics of the entity (e.g., Winglee, Valliant and Scheuren (2005)); see Harron, Goldstein and Dibben (2015) for a recent review. Matching errors are bound to occur (Bohensky et al. (2010)), and the potential benefits from data integration may be compromised. Therefore, in statistical analysis with integrated data it is important to take into account the uncertainty due to imperfect linkage.

Received November 2017; revised May 2019.
Key words and phrases. Cox model, data linkage, ECM algorithm, integrative learning, suicide prevention.

Our research was motivated by the survival analysis of youth and young adult patients in the State of Connecticut who were at elevated risk of suicide because of having been hospitalized for suicide attempt or intentional self-injury. Data from diagnosis were available from the Connecticut Hospital Inpatient Discharge Data (HIDD). Deaths by suicide were determined from the Office of the Connecticut Medical Examiner (OCME). It has been revealed, however, that suicidal death is often underreported in key Western countries (Pritchard and Hansen (2015), Tøllefsen et al. (2016)). The death records identified from the OCME for this group are incomplete because, first, suicide deaths may be underreported and, second, they do not include deaths due to other causes. Hence, some patients with censored suicide times might have died. While the missing deaths may possibly be recovered from a complete mortality database of the state, the HIDD data can only be linked to the death records by matching basic patient characteristics such as date of birth, gender, race and residential zip code because there is no unique identifier to join the two datasets even before the data were deidentified in order to protect patient privacy. Consequently, in the integrated data a censored death time before matching could be linked to multiple possible death times in the mortality data; see details in Section 2.

Figure 1 illustrates the data matching patterns in a general integrated survival analysis setup similar to that in our suicide risk study. In Dataset I a positive number of subjects' event times are observed and known to be accurate (Case 1). For those subjects whose event times are censored in Dataset I, their event times might be captured in Dataset II. After the linkage process with partial identifiers, the event time of any subject who does not find a match in dataset II is still censored (Case 2). As such, Case 1 and Case 2 consist of noncensored and censored subjects, respectively, in a standard right censored data setting. Challenges are brought by those subjects with one or more matches (Case 3); we are not sure which one, if any, of the matched event times is the truth. The subjects in Case 3 can be further classified into two types: Case 3a contains subjects whose true event time is included in the matched records, and Case 3b contains subjects whose true event time is not included in the matched records and, hence, is actually censored. This classification is unknown and has to be inferred from the data. The task can be regarded as a missing data problem in which the indicators of whether each matched record is true are missing.

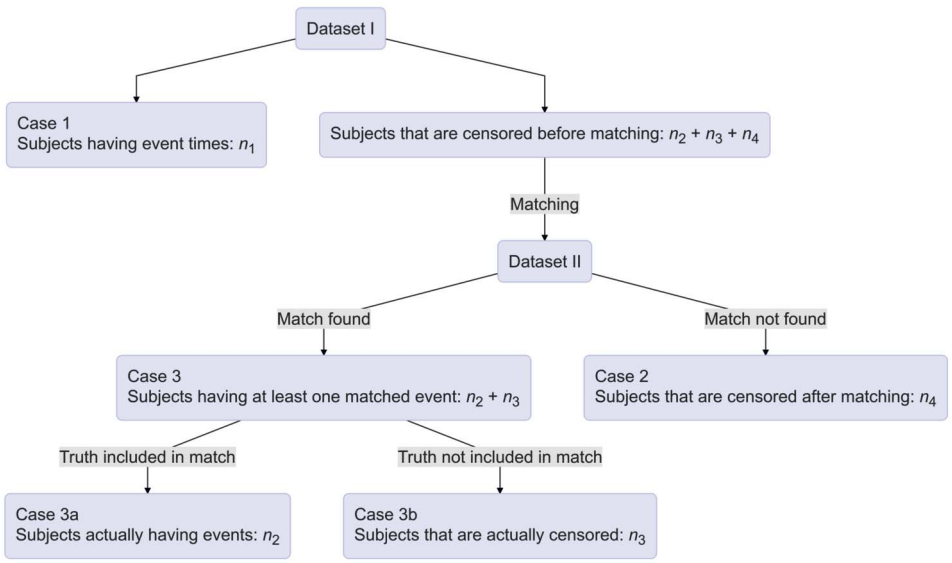

FIG. 1. *Illustration of the data matching patterns for studies with event time outcomes.*

Some efforts have been made to similar problems of mismeasured outcomes or uncertain endpoints. Snapinn (1998) proposed a modification of the Cox proportional hazard model (Cox (1972)) for nonfatal uncertain endpoints by assigning weights that represent the likelihood of each potential endpoint being true. The determination of the weights, however, requires an additional diagnostic score and depends on a subjective estimation of the relative frequency of true endpoints to false endpoints suggested by the endpoint committee or experts in the therapeutic area. Richardson and Hughes (2000) proposed an estimation procedure for the product limit estimate of survival function with no covariate based on the expectation maximization (EM) algorithm (Dempster, Laird and Rubin (1977)) when a binary diagnosis outcome was measured with uncertainty. The method was designed for discrete-time contexts where the time points of outcome testing were predetermined. Meier, Richardson and Hughes (2003) extended the discrete proportional hazard model (Kalbfleisch and Prentice (2002)) to mismeasured outcomes under a setting similar to Richardson and Hughes (2000) but allowed covariate effects. In a more general setting regression methods have been developed for linked data where the response and covariates come from two databases (e.g., Hof and Zwinderman (2012, 2015), Tancredi and Liseo (2015)). None of the existing works was designed to handle the data integration problem in a survival analysis like ours.

We propose an integrative Cox proportional hazard model for data with uncertain event time points. The uncertainty in the integrated survival data is modeled probabilistically where the probabilities depend only on the relative hazards from the Cox model itself. The model reduces to the regular Cox model when there is no uncertain record. In contrast to the method of Snapinn (1998), our method does not require any extra diagnostic variable or prior knowledge on the initial probabilities indicating the true outcomes. The estimation procedure combines the ideas of profile likelihood and the expectation conditional maximization (ECM) algorithm. The proposed method is shown to outperform naive approaches in simulation studies under realistic settings similar to the real data example. Using data obtained by integrating the HIDD/OCME data and the mortality record data of the period 2005–2012 in Connecticut, we apply the proposed approach to identifying risk factors associated with patient survival after suicide-related hospitalization. The identified diagnostic codes are mostly consistent with existing results and provide several new insights on suicide risk prediction and prevention.

The rest of this paper is organized as follows: The settings for integrated survival data for the Connecticut suicide risk analysis and the associated challenges are presented in Section 2. In Section 3 we present the integrative Cox regression modeling framework. The estimation procedure is developed in Section 4. The simulation studies are presented in Section 5. A marginal screening analysis using the proposed integrative model for the Connecticut suicide risk study is reported in Section 6. Section 7 concludes with a discussion. Implementation of the proposed methods is available in a package named `intsurv` for R (R Development Core Team (2017)) which can be accessed at https://github.com/wenjie2wang/intsurv.

2. Integrated survival data of a patient group with elevated suicide risk. Suicide is a serious public health problem in the U.S. Death by suicide is increasing among all age groups in the U.S. with a 24% increase in suicide rates observed from 1999 to 2014. There is a strong tendency for suicide attempters to make additional attempts after the initial suicide attempt (Suominen et al. (2004)), and suicide attempt is a strong predictor of suicidal death (Bostwick et al. (2015)). Understanding factors associated with suicide for patients hospitalized due to suicide attempt is critical to a better allocation of selected prevention efforts among those at elevated risk. An immediate challenge in statistical modeling is that attributing death to suicide is not easy as suicidal death is often under-reported. For example, Pritchard and Hansen (2015) showed that undetermined and accidental death was a main source of the under-reported-suicides across different countries, including the US; Tøllefsen

et al. (2016) reported that, from re-evaluations of 1800 deaths in Scandinavia, 9% of the natural deaths and accidents were reclassified as suicides in the Norwegian data, and 21% of the undetermined deaths were reclassified as suicides in the Swedish data.

We focused on patients of age 15–30 with high suicide risk in Connecticut. This group of patients consisted of those who were admitted to a hospital in Connecticut during fiscal years 2005–2012, due to suicide attempt or self-inflicted injury, survived and were discharged. The entry time of each patient into the study is the time of last such discharge. The event time is the time to death from all causes, including suicide, since the entry time. The cutoff date of the HIDD is September 30, the end of fiscal year of 2012 which means that the patients were followed up until this time. The OCME provided data on suicide deaths of this period, which included a field for reporting source, that allowed accurate identification of the corresponding patients in HIDD. Since the HIDD and OCME data only captured reported suicide deaths, we acquired the complete mortality data of the same period from the Connecticut Department of Public Health, aiming to recover the missing deaths through record linkage using basic patient characteristics. The HIDD and OCME data lead to Dataset I while the mortality data is Dataset II in Figure 1. We stress that here we set the terminal event as death from all causes rather than only due to suicide. This is mainly because the cause of death is not available in the mortality data so that it can not be recovered from data integration. On the other hand, without data integration ignoring unreported suicidal deaths and deaths due to other causes would jeopardize the validity of statistical results. Because suicide is a major cause of death among young suicide attempters, death due to all causes stands as a valid terminal event to study in our problem.

A total of 7304 patients were followed up until September 30, 2012. Among them, 4981 were white (2775 female and 2206 male) and 2323 were nonwhite (1304 female and 1019 male). Before matching, Case 1 consisted of 133 patients with confirmed suicide death from the OCME, a censoring rate of 98.2%. For the 7171 patients with censored event times, we made record linkage with the Connecticut state mortality database by date of birth, gender and race. Since the death time had to happen after the discharge, we excluded any matched event before the discharge date of each patient during the matching process. After matching, Case 2 consisted of 6546 patients with no matched record while Case 3 consisted of 625 patients with at least one matched records. In Case 3, 584 patients had one match, 39 patients had two matches and two patients had four matches; it was possible for each patient to be still alive on September 30, 2012, in which case, the true death time is censored.

The HIDD data contained a large number of records on the characteristics of patients and their previous hospital admissions. The research interest was to identify important diagnostic categories associated with patient death. The diagnostics were recorded as ICD-9 diagnosis codes or, more formally, ICD-9-CM (International Classification of Diseases, 9th Revision, Clinical Modification). We grouped the ICD-9 codes by their three leading characters that define the major diagnosis categories. Suicide attempts were identified by both ICD-9 external cause of injury codes and other ICD-9 code combinations indicative of suicidal behavior (Chen and Aseltine (2017), Patrick et al. (2010)). Other ICD-9 codes during the inpatient hospitalization fell into 167 major diagnosis categories which led to 167 indicator variables. Not all 167 indicators, however, can be used as covariates. Among them, 51 ICD-9 indicators had quasicomplete separation (Albert and Anderson (1984)) in our data; that is, there was no death event among those whose diagnosis included any of these ICD-9 categories. Although they could be potentially useful in predicting survival and thus merit further investigation, they cannot be considered as covariates in a Cox regression framework adopted in this work since their coefficient estimates would tend to be negative infinite. To focus on the main idea, we further filtered out another 58 ICD-9 indicators by restricting every cell of the cross table of the diagnosis indicator and event indicator to be at least three. The remaining 58 ICD-9 codes were used in a marginal screening analysis; see Section 6.

3. Integrative Cox model. Consider a random sample of n subjects who fall into the three cases as illustrated in Figure 1. Let I_1, I_2 and I_3 be the indices of the subjects in Cases 1, 2 and 3, respectively. For subject $j \in I_1$, we observe the event time V_j. For subject $j \in I_2$, we observe the censoring time C_j. For subject $j \in I_3$, the true event time V_j has $s_j \geq 2$ possibilities, $0 < V_{j,1} < \cdots < V_{j,s_j-1} < V_{j,s_j}$, but we only observe $0 < V_{j,1} < \cdots < V_{j,s_j-1} < C_j$ where C_j is the censoring time such that $C_j < V_{j,s_j}$. The reason for $C_j < V_{j,s_j}$ is Case 3b in Figure 1, where none of the matches is correct, so the actual death time must be after C_j. Regarding subjects in Cases 1–2 as having only $s_j = 1$ possibility with $V_{j,1} = V_j$, we use a unified notation for the observed data from subject j

$$(T_{j,k}, \Delta_{j,k}, \boldsymbol{x}_j) : k \in \{1, \ldots, s_j\},$$

where $\boldsymbol{x}_j$ is a p-dimensional vector of predictors, $T_{j,k} = \min(V_{j,k}, C_j)$, $\Delta_{j,k} = \mathbf{1}(V_{j,k} \leq C_j)$ and C_j is the censoring time. For Cases 1–2, $\Delta_{j,1}$ is the event indicator, and the notation is the same as in standard right-censored data. For Case 3, we have $s_j \geq 2$; $\Delta_{j,1} = \cdots = \Delta_{j,s_j-1} = 1$ and $\Delta_{j,s_j} = 0$ are indicators denoting that all the matches before C_j are possible events and the last possibility is always censored. These notations will be used in the estimation procedure.

The true event time V_j of subject j, $j \in \{1, \ldots, n\}$, is assumed to follow a Cox model with hazard function

$$h_j(t) = h_0(t) \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}), \tag{1}$$

where $h_0(\cdot)$ is an unspecified baseline function and $\boldsymbol{\beta}$ is a vector of unknown coefficient of the covariate vector $\boldsymbol{x}_j$. Let $S_j(t) = \exp\{-H_0(t) \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})\}$, where $H_0(t) = \int_0^t h(s)\,\mathrm{d}s$, be the survival function of subject j. The density function is then $f_j(t) = h_j(t) S_j(t)$. In addition, we assume that the censoring time C_j has an unknown density function $g(t)$, distribution function $G(t)$, survival function $\overline{G}(t) = 1 - G(t)$, does not depend on the covariates $\boldsymbol{x}_j$, and is independent of the event times conditional on the covariates $\boldsymbol{x}_j$. The conditional independence assumption of the censoring time is justified for our study because the censoring was administrative.

We propose to model the uncertain records in a probabilistic way by introducing a vector of truth indicator for each subject. For subject j, let $\boldsymbol{Z}_j = (Z_{j,1}, \ldots, Z_{j,s_j})$ be a random vector from multinomial distribution Multi$(1, \boldsymbol{\pi}_j)$,

$$Z_{j,k} = \begin{cases} 1, & V_j = V_{j,k}, \text{ or } (T_{j,k}, \Delta_{j,k}) \text{ is the truth,} \\ 0, & \text{otherwise,} \end{cases}$$

where $k \in \{1, \ldots, s_j\}$, $\sum_{k=1}^{s_j} Z_{j,k} = 1$, $0 \leq \pi_{j,k} \leq 1$ and $\sum_{k=1}^{s_j} \pi_{j,k} = 1$. As such, for each subject j, $j \in \{1, \ldots, n\}$, $\boldsymbol{\pi}_j = (\pi_{j,1}, \ldots, \pi_{j,s_j})$ is the probability vector where $\pi_{j,k} = \Pr(V_j = V_{j,k})$ (i.e., probability of the kth record being true). Clearly, for $j \in I_1 \cup I_2$, we have $s_j = 1$ and $\pi_{j,1} = 1$, that is, $Z_{j,1} = 1$ with probability 1. For $j \in I_3$, however, the truth indicators can be regarded as missing. That $Z_{j,k} = 1$, $k \in \{1, \ldots, s_j - 1\}$ corresponds to Case 3a, $Z_{j,s_j} = 1$ suggests Case 3b.

Let $\boldsymbol{T}_j = (T_{j,1}, \ldots, T_{j,s_j})$ and $\boldsymbol{\Delta}_j = (\Delta_{j,1}, \ldots, \Delta_{j,s_j})$ with realizations $\boldsymbol{t}_j = (t_{j,1}, \ldots, t_{j,s_j})$ and $\boldsymbol{\delta}_j = (\delta_{j,1}, \ldots, \delta_{j,s_j})$, respectively. Let the set of all model parameters be $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\pi}, h_0(\cdot), g(\cdot)\}$ where $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n)$. Let $\boldsymbol{z}_j$ be a realization of $\boldsymbol{Z}_j$. Given the truth indicators, we assume that the distribution of the fake records is independent of the true record and degenerates to a point mass at the point of the observed fake records. This assumption allows us to get away with modeling the intractable distribution of the fake records (e.g., the fake death times produced from imperfect data matching in our suicide risk study), so that

the likelihood of $(\boldsymbol{T}_j, \boldsymbol{\Delta}_j)$ given $\boldsymbol{Z}_k$ only depends on the likelihood of the true record. The complete-data likelihood of $(\boldsymbol{T}_j, \boldsymbol{\Delta}_j, \boldsymbol{Z}_j)$ from subject j turns out to be

$$L_j^C(\boldsymbol{\theta}) = \prod_{k=1}^{s_j} \{\pi_{j,k}[f_j(t_{j,k})\overline{G}(t_{j,k})]^{\delta_{j,k}}[g(t_{j,k})S_j(t_{j,k})]^{1-\delta_{j,k}}\}^{z_{j,k}}. \tag{2}$$

The derivation detail is available in Section 1 of the Supplementary Materials (Wang et al. (2020)). All the possible realizations of $\boldsymbol{Z}_j$ are $\boldsymbol{z}_j = (1, 0, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0), \ldots,$ $(0, 0, \ldots, 0, 1)$. The observed-data likelihood contribution from subject j is then obtained by summing out $\boldsymbol{z}_j$ in (2):

$$L_j^O(\boldsymbol{\theta}) = \sum_{k=1}^{s_j} \pi_{j,k}[f_j(t_{j,k})\overline{G}(t_{j,k})]^{\delta_{j,k}}[g(t_{j,k})S_j(t_{j,k})]^{1-\delta_{j,k}}. \tag{3}$$

Let $\boldsymbol{Y}_{\text{obs}} = \{(\boldsymbol{t}_1, \boldsymbol{\delta}_1, \boldsymbol{x}_1), \ldots, (\boldsymbol{t}_n, \boldsymbol{\delta}_n, \boldsymbol{x}_n)\}$ denote the observed data of the n independent subjects. The likelihood for the observed data is then given by $L^O(\boldsymbol{\theta}) = \prod_{j=1}^n L_j^O(\boldsymbol{\theta})$.

Thus far the observed-date likelihood in (3) is derived from a missing data perspective, but it can also be understood in several different ways. Intuitively, for subject j each of its s_j records leads to a likelihood of the event time and the censoring time, that is, $[f_j(t_{j,k})\overline{G}(t_{j,k})]^{\delta_{j,k}}[g(t_{j,k})S_j(t_{j,k})]^{1-\delta_{j,k}}$ for $k \in \{1, \ldots, s_j\}$, and the $L_j^O(\boldsymbol{\theta})$, the contribution of subject j to $L^O(\boldsymbol{\theta})$, is then constructed as a weighted sum with weights $\pi_{j,k}$ satisfying $0 \le \pi_{j,k} \le 1$ and $\sum_{k=1}^{s_j} \pi_{j,k} = 1$. From the perspective of finite mixture model, the $\pi_{j,k}$'s are the mixing probabilities, and the above likelihood form of each mixture component is a direct consequence of our assumption that given the truth indicator the distribution of the fake records degenerates such that the distribution of $(\boldsymbol{T}_j, \boldsymbol{\Delta}_j)$ only depends on the true record. Interestingly, the proposed method is also connected to a trimmed likelihood approach (e.g., Hadi and Luceño (1997), Neykov et al. (2007)), for which, however, the optimization problem is combinatorial in nature and a naive exhaustive search is not feasible; see Section 4.4 for details. In contrast, the proposed probabilistic formulation allows us to develop an ECM algorithm to conduct maximum likelihood estimation. We remark that our approach may allow potential incorporation of certain known missing mechanism of the true label through imposing more structures on $\boldsymbol{\pi}_j$ or modeling them using covariates. For instance, in some applications it may be reasonable to assume that the prior probability of being censored is the same for all the subjects with uncertain records. In this work, however, we focus on the unconstrained situation.

4. Model estimation via an ECM algorithm.

4.1. *Estimation procedure.* The ECM algorithm is a variation of the powerful EM algorithm for dealing with incomplete data (Meng and Rubin (1993)). It replaces the M-step of an EM algorithm with multiple conditional maximization (CM) steps which are often computationally easier to handle. We propose a maximum likelihood estimation procedure for the integrative Cox model following the architecture of the ECM in which the CM-steps utilize a profile likelihood similar to the partial likelihood (Cox (1975)).

The complete-data loglikelihood can be decomposed into two parts which involve two exclusive sets of parameters, respectively. Let $\boldsymbol{Y}_{\text{mis}} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ and $\boldsymbol{Y} = \{\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}\}$. From (2) the complete-data loglikelihood is

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \ell\big(\boldsymbol{\beta}, \boldsymbol{\pi}, h_0(\cdot) \mid \boldsymbol{Y}\big) + \ell_c\big(g(\cdot) \mid \boldsymbol{Y}\big), \tag{4}$$

where

$$\begin{aligned}&\ell(\boldsymbol{\beta}, \boldsymbol{\pi}, h_0(\cdot) \mid \boldsymbol{Y})\\ &\quad= \sum_{j=1}^{n}\sum_{k=1}^{s_j} z_{j,k}\{\log \pi_{j,k} + \delta_{j,k}\log f_j(t_{j,k}) + (1-\delta_{j,k})\log S_j(t_{j,k})\},\end{aligned} \tag{5}$$

and

$$\ell_c(g(\cdot) \mid \boldsymbol{Y}) = \sum_{j=1}^{n}\sum_{k=1}^{s_j} z_{j,k}\{\delta_{j,k}\log \overline{G}(t_{j,k}) + (1-\delta_{j,k})\log g(t_{j,k})\}. \tag{6}$$

The second part $\ell_c(g(\cdot) \mid \boldsymbol{Y})$ only involves the nuisance distribution of the censoring time.

We compute the conditional expectations of the complete-data loglikelihood (4) given the observed data $\boldsymbol{Y}_{\text{obs}}$ and the set of parameter estimates $\boldsymbol{\theta}^{(i)} = \{\boldsymbol{\beta}^{(i)}, \boldsymbol{\pi}^{(i)}, h_0^{(i)}(\cdot), g^{(i)}(\cdot)\}$ at ith iteration ($i = 0, 1, \ldots$) where $\boldsymbol{\theta}^{(0)}$ is the initial/starting estimate. Define

$$\begin{aligned}w_{j,k}(\boldsymbol{\theta}^{(i)}) &:= \mathrm{P}(Z_{j,k} = 1, \boldsymbol{T}_j, \boldsymbol{\Delta}_j \mid \boldsymbol{\theta}^{(i)})\\ &= \pi_{j,k}^{(i)}\big(h_{j,k}^{(i)} S_{j,k}^{(i)} \overline{G}_{j,k}^{(i)}\big)^{\delta_{j,k}}\big(g_{j,k}^{(i)} S_{j,k}^{(i)}\big)^{1-\delta_{j,k}},\end{aligned}$$

where $h_{j,k}^{(i)} = h_0^{(i)}(t_{j,k})\exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}^{(i)})$ and $S_{j,k}^{(i)} = \exp\{-H_0^{(i)}(t_{j,k})\exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}^{(i)})\}$, $\overline{G}_{j,k}^{(i)} = \overline{G}^{(i)}(t_{j,k})$ and $g_{j,k}^{(i)} = g^{(i)}(t_{j,k})$. By Bayes' rule we have

$$\begin{aligned}p_{j,k}(\boldsymbol{\theta}^{(i)}) &:= \mathrm{P}(Z_{j,k} = 1 \mid \boldsymbol{T}_j, \boldsymbol{\Delta}_j, \boldsymbol{\theta}^{(i)})\\ &= \frac{w_{j,k}(\boldsymbol{\theta}^{(i)})}{\sum_{k=1}^{s_j} w_{j,k}(\boldsymbol{\theta}^{(i)})}.\end{aligned} \tag{7}$$

Plugging (7) into (5) and (6), we obtain the E-step that involves two separate parts:

$$\begin{aligned}&\mathbb{E}\ell\{\boldsymbol{\beta}, \boldsymbol{\pi}, h_0(\cdot) \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(i)}\}\\ &\quad= \sum_{j=1}^{n}\sum_{k=1}^{s_j} p_{j,k}(\boldsymbol{\theta}^{(i)})\{\log(\pi_{j,k}) + \delta_{j,k}\log f_j(t_{j,k})\\ &\qquad+ (1-\delta_{j,k})\log S(t_{j,k})\}\end{aligned} \tag{8}$$

and

$$\begin{aligned}&\mathbb{E}\ell_c\{g(\cdot) \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\theta}^{(i)}\}\\ &\quad= \sum_{j=1}^{n}\sum_{k=1}^{s_j} p_{j,k}(\boldsymbol{\theta}^{(i)})\{\delta_{j,k}\log \overline{G}(t_{j,k}) + (1-\delta_{j,k})\log g(t_{j,k})\}.\end{aligned} \tag{9}$$

The separation of the two terms in parameters facilitates the M-step. The first term (8) can be handled by profiling out the nuisance parameters. Note that, for fixed $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$, the $h_0(t)$ maximizing the conditional expectation (8) is a discrete function that is positive only at possible event times and zero anywhere else. Let $Y_{j,k}(t) = \mathbf{1}(t_{j,k} \geq t)$ and $N_{j,k}(t) = z_{j,k}\mathbf{1}(t_{j,k} \leq t, \delta_{j,k} = 1)$. Then, the true number of events by time t is $N(t) = \sum_{j=1}^{n}\sum_{k=1}^{s_j} N_{j,k}(t)$. Let $\mathrm{d}N(t)$ denote the number of true events at time t. Let $\widetilde{N}_{j,k}(t; \boldsymbol{\theta}^{(i)}) = p_{j,k}(\boldsymbol{\theta}^{(i)})\mathbf{1}(t_{j,k} \leq t, \delta_{j,k} = 1)$ and $\widetilde{N}(t; \boldsymbol{\theta}^{(i)}) = \sum_{j=1}^{n}\sum_{k=1}^{s_j}\widetilde{N}_{j,k}(t; \boldsymbol{\theta}^{(i)})$ which are the conditional expectation of $N_{j,k}(t)$ and $N(t)$ given $\boldsymbol{Y}_{\text{obs}}$, evaluated at $\boldsymbol{\theta}^{(i)}$, respectively. Then,

$\mathrm{d}\widetilde{N}(t;\boldsymbol{\theta}^{(i)}) = \mathbb{E}\{\mathrm{d}N(t)|\boldsymbol{Y}_{\text{obs}},\boldsymbol{\theta}^{(i)}\}$ is the jump size of $\widetilde{N}(t;\boldsymbol{\theta}^{(i)})$ at time t. Equation (8) can be rewritten to allow tied event times as follows:

$$\begin{aligned}
&\mathbb{E}\ell\{\boldsymbol{\beta},\boldsymbol{\pi},h_0(\cdot)\mid \boldsymbol{Y}_{\text{obs}},\boldsymbol{\theta}^{(i)}\}\\
&\quad=\sum_{t\in\mathcal{T}}\Bigg[-h_0(t)\sum_{j=1}^{n}\sum_{k=1}^{s_j}Y_{j,k}(t)p_{j,k}\big(\boldsymbol{\theta}^{(i)}\big)\exp\big(\boldsymbol{x}_j^{\top}\boldsymbol{\beta}\big)+\mathrm{d}\widetilde{N}\big(t;\boldsymbol{\theta}^{(i)}\big)\log h_0(t)\Bigg] \qquad (10)\\
&\qquad+\sum_{j=1}^{n}\sum_{k=1}^{s_j}p_{j,k}\big(\boldsymbol{\theta}^{(i)}\big)\big[\delta_{j,k}\boldsymbol{x}_j^{\top}\boldsymbol{\beta}+\log\pi_{j,k}\big],
\end{aligned}$$

where $\mathcal{T}=\{t_{j,k}\mid \delta_{j,k}=1,k\in\{1,\dots,s_j\},j\in\{1,\dots,n\}\}$ is the collection of all observed possible event times.

Given $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$, the baseline hazard h_0 only appears in the first term of (10), and the maximizer is

$$\hat{h}_0(t)=\frac{\mathrm{d}\widetilde{N}(t;\boldsymbol{\theta}^{(i)})}{\sum_{j=1}^{n}\sum_{k=1}^{s_j}Y_{j,k}(t)p_{j,k}(\boldsymbol{\theta}^{(i)})\exp(\boldsymbol{x}_j^{\top}\boldsymbol{\beta})},$$

which is nonzero only for those $t\in\mathcal{T}$, similar to the "Breslow estimator" (Breslow (1974)). Further, for fixed $\boldsymbol{\beta}$ it is easy to check that $\pi_{j,k}^{(i+1)}=p_{j,k}(\boldsymbol{\theta}^{(i)})$ maximizes (10) by Lagrange multipliers method. Plugging these estimators back into (10), we get a profile likelihood in terms of $\boldsymbol{\beta}$

$$\begin{aligned}
&\mathbb{E}\ell\{\boldsymbol{\beta},\hat{\boldsymbol{\pi}},\hat{h}_0\mid \boldsymbol{Y}_{\text{obs}},\boldsymbol{\theta}^{(i)}\}\\
&\quad=\sum_{t\in\mathcal{T}}\{-\mathrm{d}\widetilde{N}(t;\boldsymbol{\theta}^{(i)})[1-\log\mathrm{d}\widetilde{N}(t;\boldsymbol{\theta}^{(i)})]\}+\sum_{j=1}^{n}\sum_{k=1}^{s_j}p_{j,k}(\boldsymbol{\theta}^{(i)})\log p_{j,k}(\boldsymbol{\theta}^{(i)})\\
&\qquad+p\ell(\boldsymbol{\beta}\mid\boldsymbol{\theta}^{(i)}),
\end{aligned}$$

where

$$\begin{aligned}
p\ell(\boldsymbol{\beta}\mid\boldsymbol{\theta}^{(i)})&=\sum_{j=1}^{n}\sum_{k=1}^{s_j}\int_0^{\infty}I(\boldsymbol{\beta},t\mid\boldsymbol{\theta}^{(i)})\,\mathrm{d}\widetilde{N}_{j,k}(t;\boldsymbol{\theta}^{(i)}),\\
I(\boldsymbol{\beta},t\mid\boldsymbol{\theta}^{(i)})&=\boldsymbol{x}_j^{\top}\boldsymbol{\beta}-\log\Bigg(\sum_{l=1}^{n}\sum_{m=1}^{s_l}Y_{l,m}(t)p_{l,m}(\boldsymbol{\theta}^{(i)})\exp(\boldsymbol{x}_l^{\top}\boldsymbol{\beta})\Bigg),
\end{aligned}\qquad (11)$$

is the only part involving $\boldsymbol{\beta}$. This profiling approach is similar to the partial likelihood of Cox (1975) except that the distribution of the censoring time comes into play through $p_{j,k}$'s and $\mathrm{d}\widetilde{N}_{j,k}$'s. The estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by maximizing (11). Once $\hat{\boldsymbol{\beta}}$ has converged, $\hat{h}_0(\cdot)$ and $\hat{\pi}_{j,k}$'s can be updated.

Maximizing the second part (9) involves nonparametric maximum likelihood estimator of the censoring distribution function $G(\cdot)$. We characterize the censoring time by its hazard function $h_c(\cdot)$. Similar to $h_0(t)$, the $h_c(\cdot)$ that maximizes (9) is nonzero only at the observed censoring times. By the assumption we made, the only possible censoring time for each subject is its last record time. For $j\in\{1,\dots,n\}$, define $C_j(t;\boldsymbol{\theta}^{(i)})=p_{j,s_j}(\boldsymbol{\theta}^{(i)})\mathbf{1}(t_{j,s_j}\le t,\delta_{j,s_j}=0)$ and $C(t;\boldsymbol{\theta}^{(i)})=\sum_{j=1}^{n}C_j(t;\boldsymbol{\theta}^{(i)})$. Let $\mathrm{d}C(t;\boldsymbol{\theta}^{(i)})$ be the jump size of $C(t;\boldsymbol{\theta}^{(i)})$ at time t. Then, we may rewrite (9) to allow tied censoring times as follows:

$$\begin{aligned}
&\mathbb{E}\ell_c(g(\cdot)\mid\boldsymbol{Y}_{\text{obs}},\boldsymbol{\theta}^{(i)})\\
&\quad=\sum_{t\in\mathcal{C}}\Bigg\{\mathrm{d}C\big(t;\boldsymbol{\theta}^{(i)}\big)\log h_c(t)-h_c(t)\sum_{j=1}^{n}\sum_{k=1}^{s_j}p_{j,k}\big(\boldsymbol{\theta}^{(i)}\big)Y_{j,k}(t)\Bigg\},
\end{aligned}$$

Algorithm 1 Estimation procedure for integrative Cox model with uncertain event records. (The dependence of $\pi_{j,k}$'s, $\widetilde{N}_{j,k}$'s, $\mathrm{d}\widetilde{N}(t)$, and $\mathrm{d}C(t)$ on $\boldsymbol{\theta}$ is dropped for ease of notation)

initialize $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$;
repeat
  for $j = 1, 2, \ldots, n$ **do** ▷ Update $\widetilde{N}_{j,k}(t)$'s
    for $k = 1, 2, \ldots, s_j$ **do**

$$\widetilde{N}_{j,k}(t) \leftarrow \pi_{j,k}\mathbf{1}(t_{j,k} \le t, \delta_{j,k} = 1);$$

    end for
  end for
  for each $t \in \mathcal{T}$ **do** ▷ Update $\hat{h}_0(\cdot)$

$$h_0(t) \leftarrow \frac{\mathrm{d}\widetilde{N}(t)}{\sum_{j=1}^{n}\sum_{k=1}^{s_j} Y_{j,k}(t)\pi_{j,k}\exp(\boldsymbol{x}_j^\top\boldsymbol{\beta})}; \qquad H_0(t) \leftarrow \sum_{s\le t} h_0(s);$$

  end for
  for each $t \in \mathcal{C}$ **do** ▷ Update $\hat{h}_c(\cdot)$

$$h_c(t) \leftarrow \frac{\mathrm{d}C(t)}{\sum_{j=1}^{n}\sum_{k=1}^{s_j} Y_{j,k}(t)\pi_{j,k}}; \qquad H_c(t) \leftarrow \sum_{s\le t} h_c(s),$$

  end for
  for $j = 1, 2, \ldots, n$ **do** ▷ Update $\hat{\pi}_{j,k}$'s
    for $k = 1, 2, \ldots, s_j$ **do**

$$S_{j,k} \leftarrow \exp\{-H_0(t_{j,k})\exp(\boldsymbol{x}_j^\top\boldsymbol{\beta})\}; \qquad \overline{G}_{j,k} \leftarrow \exp\{-H_c(t_{j,k})\};$$

$$w_{j,k} \leftarrow \pi_{j,k}[h_{j,k}S_{j,k}\overline{G}_{j,k}]^{\delta_{j,k}}[g_{j,k}S_{j,k}]^{1-\delta_{j,k}}; \qquad \pi_{j,k} \leftarrow \frac{w_{j,k}}{\sum_{k=1}^{s_j} w_{j,k}};$$

    end for
  end for
  $\boldsymbol{\beta} \leftarrow \arg\max p\ell(\boldsymbol{\beta}|\boldsymbol{\theta})$ ▷ Update $\hat{\boldsymbol{\beta}}$
until Convergence

where $\mathcal{C} = \{t_{j,s_j} \mid \delta_{j,s_j} = 0, j \in \{1, \ldots, n\}\}$ is the collection of all observed censoring times. Maximizing it with respect to $h_c(t)$ gives

$$\hat{h}_c(t) = \frac{\mathrm{d}C(t;\boldsymbol{\theta}^{(i)})}{\sum_{j=1}^{n}\sum_{k=1}^{s_j} p_{j,k}(\boldsymbol{\theta}^{(i)})Y_{j,k}(t)}.$$

Therefore, for every record time $t_{j,k}$, we have

$$\hat{\overline{G}}(t_{j,k}) = \exp\Big\{-\sum_{t\le t_{j,k}}\hat{h}_c(t)\Big\} = \exp\Big\{-\sum_{t\le t_{j,k}}\frac{\mathrm{d}C(t;\boldsymbol{\theta}^{(i)})}{\sum_{l=1}^{n}\sum_{m=1}^{s_l} p_{l,m}(\boldsymbol{\theta}^{(i)})Y_{l,m}(t)}\Big\}$$

and $\hat{g}(t_{j,k}) = \hat{h}_c(t)\hat{\overline{G}}(t_{j,k})$.

We summarize the ECM estimation procedure in Algorithm 1. In our numerical studies we stop the algorithm if $\|\boldsymbol{\beta}^{(i)} - \boldsymbol{\beta}^{(i-1)}\|/\|\boldsymbol{\beta}^{(i)} + \boldsymbol{\beta}^{(i-1)}\| < 10^{-6}$ and $\|\boldsymbol{\pi}^{(i)} - \boldsymbol{\pi}^{(i-1)}\|/\|\boldsymbol{\pi}^{(i)} + \boldsymbol{\pi}^{(i-1)}\| < 10^{-8}$.

4.2. *Initialization*. Since the maximum likelihood estimation problem here is nonconvex, it may admit multiple local maxima. Therefore, we recommend setting multiple initial

values of $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ to help identify a good solution, as allowed by the available computational resources. In particular, we propose two simple but pragmatic initialization procedures that work well even with limited resources.

The first procedure is as follows:

(i) Fit a regular Cox model on all the certain records (Cases 1–2) and use the estimated coefficients to initialize $\boldsymbol{\beta}$; initialize $\hat{S}_{j,k}$ with the fitted survival function evaluated at $t_{j,k}$; initialize $\hat{h}_{j,k}$ with a nearest left neighbor interpolation of the fitted hazard function (if no left neighbor, use nearest right neighbor).

(ii) Switching event and censoring for all the certain records (Cases 1–2), estimate the hazard function for censoring by the Nelson–Aalen estimator (without covariates) and obtain the corresponding survival function estimate; initialize $\hat{\overline{G}}_{j,k}$ with the fitted survival function evaluated at $t_{j,k}$; initialize $\hat{h}_c(t_{j,k})$ with a nearest left neighbor interpolation of the fitted hazard function (if no left neighbor, use nearest right neighbor).

(iii) Plug $\hat{w}_{j,k} = h^*_{j,k}\hat{S}_{j,k}\hat{\overline{G}}_{j,k}$, where $h^*_{j,k} = \delta_{j,k}\hat{h}_{j,k} + (1-\delta_{j,k})\hat{h}_c(t_{j,k})$, into (7) as $w_{j,k}$ and initialize $\pi_{j,k}$ as the resulting $p_{j,k}$.

In the above procedure letting $\hat{h}^*_{j,k} = 1$ in step (iii) leads to a simpler alternative, which puts more weights to the uncertain event times before the censoring time, and thus may work better when Case 3a is estimated to have a larger size than Case 3b. This gives a second initialization procedure.

The two initialization procedures were applied in the simulation studies presented in Section 5, and the results were satisfactory in most scenarios.

4.3. *Inference*. In an EM or ECM algorithm, generally, standard error (SE) estimates for the parameter estimates cannot be easily produced along with the estimation procedure. A few approaches have been proposed for estimating the asymptotic covariance matrix for parameters of interest, including the supplemented EM (SEM) algorithm (Meng and Rubin (1991)), the profile likelihood approach (Murphy and van der Vaart (2000)), numerical differentiation methods based on forward difference and Richardson extrapolation (Jamshidian and Jennrich (2000)) and their variants with profiling (Xu, Baines and Wang (2014)). Unfortunately, none of these methods is readily applicable to our case. In our work we use the bootstrap (Efron (1979, 1981)) method that performs resampling at the subject level for survival data for making inference. Efron (1981) proposed the SE be estimated as sample standard deviation of bootstrap estimates, or based on inter-quantile range and normal approximation. The p-values from the Wald test for testing the significance of each regression coefficient can then be computed.

4.4. *Connection with trimmed likelihood*. We show that the proposed method is closely connected to a trimmed likelihood approach which offers an intuitive understanding of our method from the robust estimation perspective. The trimmed likelihood (Hadi and Luceño (1997), Rousseeuw (1984), Neykov et al. (2007)) is a general approach for conducting robust maximum likelihood estimation in the presence of outliers in which the observations are trimmed according to their contributions to the likelihood function. Our probabilistic modeling approach via ECM provides an efficient way for targeting the computationally infeasible trimmed likelihood estimator.

Recall the observed-data likelihood formulation given in (3). Denote

$$r_{j,k}(\boldsymbol{\beta}) = \big[f_j(t_{j,k})\overline{G}(t_{j,k})\big]^{\delta_{j,k}}\big[g(t_{j,k})S_j(t_{j,k})\big]^{1-\delta_{j,k}},$$

where $j \in \{1,\dots,n\}$, $k \in \{1,\dots,s_j\}$. For ease of notation, here we do not explicitly write out the dependency of $r_{j,k}(\boldsymbol{\beta})$ on the observed data and assume other unknown quantities

$h_0(\cdot)$ and $g(\cdot)$ have been profiled out. (In fact, the above can be regarded as a general survival modeling formulation in this section.) Then, the proposed maximum likelihood estimator can be expressed as

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}) \in \arg\max_{\boldsymbol{\beta},\boldsymbol{\pi}} \prod_{j=1}^{n} \left(\sum_{k=1}^{s_j} \pi_{j,k} r_{j,k}(\boldsymbol{\beta}) \right). \tag{12}$$

Here, each $\boldsymbol{\pi}_j$ is a probability vector, and there is no additional structural constraint on $\boldsymbol{\pi}$. Now, for each j, define $r_{j,(s_j)}(\boldsymbol{\beta})$ as the largest order statistic of $r_{j,k}(\boldsymbol{\beta})$, $k = 1, \ldots, s_j$. Then, a trimmed likelihood estimator can be constructed as

$$\tilde{\boldsymbol{\beta}} \in \arg\max_{\boldsymbol{\beta}} \prod_{j=1}^{n} r_{j,(s_j)}(\boldsymbol{\beta}). \tag{13}$$

Intuitively, (13) shows that the optimal $\boldsymbol{\beta}$ is reached when for each patient with uncertain records only the most plausible record (as judged by having the largest log-likelihood value among all the records) contributes to the overall log-likelihood function and the rest all get trimmed. Interestingly, it can be verified that the two methods in (12) and (13) share the same set of global solutions.

LEMMA 4.1. *The $\hat{\boldsymbol{\beta}}$ from solving* (12) *is a solution of* (13) *and vice versa.*

For each $\boldsymbol{\pi}_j = (\pi_{j,1}, \ldots, \pi_{j,s_j})$, we have $\hat{\boldsymbol{\pi}}_j = \arg\max_{\boldsymbol{\pi}_j} \sum_{k=1}^{s_j} \pi_{j,k} r_{j,k}(\hat{\boldsymbol{\beta}})$, because given $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ the problem in (12) is separable in each set of $\boldsymbol{\pi}_j$. Then, the maximum can be attained at $\hat{\pi}_{j,k_j^0} = 1$ and $\hat{\pi}_{j,k} = 0$ for $k \neq k_j^0$ where $k_j^0 \in \arg\max_k r_{j,k}(\hat{\boldsymbol{\beta}})$. It follows that the maximum value of the objective function in (12) can be written as $\prod_{j=1}^{n} r_{j,(s_j)}(\hat{\boldsymbol{\beta}})$ which clearly reveals that $\hat{\boldsymbol{\beta}}$ is a maximizer of the trimmed likelihood problem in (13). On the other hand, let $\tilde{\boldsymbol{\pi}}_j$ be that $\tilde{\pi}_{j,k_j^0} = 1$ and $\tilde{\pi}_{j,k} = 0$ for $k \neq k_j^0$, where $k_j^0 \in \arg\max_k r_{j,k}(\tilde{\boldsymbol{\beta}})$ with some abuse of notation. Then it can be seen that $\{\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\pi}}\}$ is necessarily a solution of (12).

In practice, however, finding the global solution of the trimmed likelihood problem is infeasible via a naive exhaustive search approach. For example, in our suicide risk study an exhaustive search amounts to fit $2^{584} \times 3^{39} \times 5^2$ many Cox models. In contrast, our probabilistic modeling approach can be regarded as an efficient way for targeting the trimmed likelihood estimator via the ECM algorithm with carefully constructed initial values.

5. Simulation study.

5.1. *Simulation settings.* Our simulation settings were designed to mimic the data integration process in the survival analysis of patients admitted to hospital due to unsuccessful suicide attempts in Connecticut. As shown in Figure 1, n_1 is the number of subjects with events observed for certain from Dataset I (Case 1); n_2 is the number of subjects whose true event time is included in the matched event times (Case 3a); n_3 is the number of subjects whose true event time is censored but for whom some false event times are matched (Case 3b); n_4 is the number of subjects whose event times are censored for certain since no match is found from Dataset II (Case 2). As such, $n = \sum_{i=1}^{4} n_i$ is the total sample size, and $n_2 + n_3 + n_4$ is the number of subjects that are censored before data matching.

We define a few quantities for designing the experiment: censoring rate of dataset I before matching (CR1) is $\text{CR1} = 1 - n_1/n$; matching rate (MR) is $\text{MR} = (n_2 + n_3)/(n_2 + n_3 + n_4)$; correct matching rate (CMR) $\text{CMR} = n_2/(n_2 + n_3)$. MR is the proportion of subjects having

TABLE 1
Summary of different simulation settings. The number of subjects in Group 1 *is fixed at* $n_1 + n_2 = 200$

				Group 1		Group 2			
Scenario #	CR1	MR	CMR	n_1 (Case 1)	n_2 (Case 3a)	n_3 (Case 3b)	n_4 (Case 2)	n	OCR
1	30	70	20	189	11	46	24	270	26
2	30	70	80	161	39	9	21	230	13
3	60	40	20	178	22	84	160	444	55
4	60	40	80	136	64	17	122	339	41
5	90	10	20	167	33	117	1350	1667	88
6	90	10	80	118	82	24	953	1177	83

CR1: Censoring rate before matching (%); MR: Matching rate (%); CMR: Correct matching rate (%); OCR: Oracle censoring rate (%).

matched records among subjects whose event times are censored from Dataset I; CMR is the proportion of the subjects whose true event time is contained in the matched event times. In all the settings we set MR $= 1 -$ CR1, assuming that the lower the CR1, the more likely that Dataset I misses true events among the censored records. The number of subjects who actually had events was fixed at $n_1 + n_2 = 200$ to keep an approximately same benchmark performance from oracle models under different settings.

Three levels of CR1 were considered, that is, CR1 $\in \{30\%, 60\%, 90\%\}$, corresponding to moderate, heavy and severe censoring, respectively. Two levels of CMR were considered, that is, CMR $\in \{20\%, 80\%\}$; the larger the CMR, the more valuable information can be potentially recovered from Dataset II. Given (CR1, MR, CMR) and with the condition $n_1 + n_2 = 200$, the values of n_i's, $i = 1, \ldots, 4$ were then completely determined. Table 1 summarizes the sample size and its decomposition into the four cases and for each of the six simulation scenarios determined by the combinations of CR1 and CMR.

For ease of data generation, we divide the subjects into two groups: Group 1 contains those whose true event times are included in the observed data, not necessarily certain though (Case 1 and Case 3a); Group 2 contains those whose true event times are not in the observed data (Case 2 and Case 3b). Define oracle censoring rate (OCR), OCR $= (n_3 + n_4)/n$, the proportion of Group 2 in the sample which is unobserved but completely determined for each setting after the values of n_i's are determined. Our strategy was to generate true event time and censoring time for all subjects for a given OCR. First, identify subjects in Case 3a from Group 1, identify subjects in Case 3b from Group 2 and then generate fake event times for those in Case 3a and Case 3b, respectively.

The true event times were generated from Cox model (1) with a Weibull baseline hazard function. Four independent covariates were included in the model; the first three were from the standard normal distribution and the fourth was from the Bernoulli distribution with rate 0.5. All four true regression coefficients were set to be 1. The censoring time was generated from the uniform distribution over (0.5, 12.5). The Weibull-shape parameter was set to be 2, 1 and 0.7 for the moderate, heavy and severe censoring scenarios in terms of CR1, respectively. The Weibull-scale parameter was tuned in each setting so that the OCR determined in that setting is attained on average.

To identify Case 3a subjects from Group 1 and Case 3b subjects from Group 2, we treated the data uncertainty as a missing-label problem. The labels are observed for the $n_1 + n_4$ subjects in Cases 1–2 but are missing for the $n_2 + n_3$ subjects in Case 3. Two missing mechanisms were considered for the labels, missing completely at random (MCAR) and missing not at random (MNAR). In the MCAR mechanism the probability of a label being missing

was completely random, regardless of the underlying true event time. In the MNAR mechanism the probability of a label being missing was proportional to the true event time; the longer the true event time, the more likely a subject was identified as Case 3a from Group 1 or Case 3b from Group 2. Such decomposition ensures that the sample size decomposition of each simulated data closely matches its corresponding setting in Table 1.

The last step was to generate fake event times for subjects in Case 3. For subjects in Case 3a, their censoring times were observed and true event times were included in the matches. The number of additional fake event times was set to be zero or one with probability 0.9 and 0.1, respectively. In other words, the possible records for each of them consisted of one observed censoring time, one true event time and one additional fake event time with probability 0.1. For subjects in Case 3b, their true event times were censored, and the number of fake event times was set to be one or two with probability 0.9 and 0.1, respectively. In other words, each of them had one observed censoring time, one or two fake event times with probability 0.9 or 0.1, respectively. Each fake event time was generated from Cox model (1) with one extra covariate in addition to the existing four covariates, conditional on that the fake event time was less than the censoring time (Nadarajah and Kotz (2006)). This extra covariate took value -1 or 1 with equal probability, and its coefficient was set to be 3.

5.2. *Competing methods and evaluation metrics*. Three competing methods were considered, multiple imputation (MI) and two naive approaches. MI was originally introduced by Rubin (1987) for nonresponse in surveys which imputes every missing value multiple times with draws from certain distribution and summarized the results from the multiple versions of the complete data. In our setup the missing values are the truth indicators. Given a simulated dataset, we imputed 200 times the truth indicators for the subjects in Case 3 and took the average of the coefficient estimates from fitting the regular Cox model with each imputed data as the final estimates. Specifically, in each imputation and for each subject the truth indicator vector was generated from a multinomial distribution, where the probability of censoring was set to be proportional to $n_4/(n_1 + n_4)$, and the remaining probability was equally split among the uncertain event records. The two naive approaches were based on the regular Cox model as well. The first (denoted by C.Cox) fits the regular Cox model to Dataset I, which treats all subjects in Case 3 as censored, completely ignoring integration with Dataset II. C.Cox may give biased estimator for not considering the events missed by Dataset I. The second approach (denoted by U.Cox) excludes those subjects with multiple event times after matching with Dataset II (Case 3) and fits the regular Cox model with the remaining subjects with unique records (Case 1 and Case 2). The data used by U.Cox is a subset of that used by C.Cox. By removing subjects in dataset I whose event times were not uniquely recorded, U.Cox may give less efficient but unbiased estimation under MCAR.

The proposed integrative Cox model is denoted by I.Cox. We also included two oracle procedures where the true event indicators are known a priori, the oracle Cox model (O.Cox) and the oracle Weibull model (O.Weibull). They give the best achievable performances, infeasible in practice but can be used as references in comparison.

We measured the estimation performance by the ℓ_2-norm of $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, that is, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = [(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)]^{1/2}$, where $\boldsymbol{\beta}_0$ is the underlying true coefficient vector and $\hat{\boldsymbol{\beta}}$ is its estimator. In addition, we estimated the baseline survival functions from the purposed I.Cox model and two naive Cox methods, and compared them with the true parametric curve over a tense time grid from 0 to 12 with step size of 0.1. For each subject with multiple records, the estimated probabilities $\hat{\boldsymbol{\pi}}_j$ from the proposed I.Cox model can be used to identify the true record. We used the Bayes' rule to select the record with the largest estimated probability; by comparing to the underlying truth, we computed the correct identification rate of the true records among the subjects having uncertain records. The experiment was replicated 1000 time under each setting and the results were then averaged.

TABLE 2
Comparison on parameter estimation performance through mean of $100 \times \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|$ (*with the standard deviation given in parenthesis*)

#	O.Weibull	O.Cox	I.Cox	U.Cox	C.Cox	MI
			MCAR			
1	18.0 (7.3)	20.7 (8.6)	22.4 (9.7)	24.9 (9.9)	81.1 (22.0)	81.0 (9.9)
2	17.5 (7.7)	20.8 (8.9)	22.1 (9.6)	23.4 (10.0)	139.7 (14.7)	80.8 (11.0)
3	18.5 (7.9)	19.7 (8.7)	23.6 (10.2)	22.3 (9.2)	55.9 (17.2)	81.2 (9.9)
4	18.6 (7.6)	20.3 (8.7)	22.8 (10.1)	26.0 (11.4)	107.7 (15.9)	94.2 (12.0)
5	18.0 (8.0)	18.2 (8.1)	20.6 (9.0)	20.2 (9.0)	30.5 (12.3)	39.5 (11.7)
6	18.4 (8.2)	18.8 (8.3)	22.2 (9.8)	30.1 (12.6)	51.8 (12.3)	51.4 (11.3)
			MNAR			
1	18.0 (7.3)	20.7 (8.6)	22.4 (9.4)	27.6 (10.6)	48.4 (21.0)	81.5 (10.0)
2	17.5 (7.7)	20.8 (8.9)	22.4 (9.7)	24.4 (10.4)	79.9 (20.9)	55.2 (11.5)
3	18.5 (7.9)	19.7 (8.7)	22.5 (10.2)	22.9 (9.7)	27.1 (11.8)	86.4 (8.9)
4	18.6 (7.6)	20.3 (8.7)	23.2 (10.2)	24.2 (10.7)	38.3 (15.0)	51.6 (12.0)
5	18.0 (8.0)	18.2 (8.1)	21.1 (9.2)	20.7 (9.3)	21.1 (9.0)	40.2 (9.3)
6	18.4 (8.2)	18.8 (8.3)	23.5 (10.3)	30.2 (13.0)	27.3 (11.3)	30.1 (10.8)

MCAR: Missing completely at random; MNAR: Missing not at random.

5.3. *Simulation results*. Table 2 summarizes the simulation results on parameter estimation. As expected, the two practically-infeasible oracle approaches perform the best which provide benchmarks for comparison. The I.Cox method and the U.Cox method appear to have a clear advantage over the C.Cox method and the MI method under most settings. The disadvantage of the C.Cox method is expected; subjects in Case 3a are mistakenly treated as censored which increases the variance in estimation due to less events and introduces bias due to the mistakenly treated censoring. Its performance is even worse in MCAR settings because in the MNAR setting longer survival time is more likely to be uncertain, such that the true event time is more likely to be close to the censoring time than under MCAR. The MI method performs worse than the C.Cox method in the setting with lower CMR under MNAR, unlike in other settings where they are less different because the imputation does not account for the informative missingness and lower CMR means higher noise in data integration.

Between I.Cox and U.Cox, it appears that the I.Cox method either substantially outperforms U.Cox, or has comparable performance as compared to U.Cox. Specifically, when CR1 is moderate (30%) and MR is high (70%), I.Cox outperforms U.Cox with more advantage in the MNAR case than in the MCAR case. When CR1 is heavy (60%) with 40% MR, I.Cox outperforms U.Cox in the cases where CMR is 80% and in the MNAR case with 20% CMR; otherwise, it has a close but slightly worse performance than U.Cox. Lastly, under severe CR1 (90%) and low MR (10%), I.Cox outperforms U.Cox when CMR is 80% and has a slightly worse performance when CMR decreased to 20%. It is not surprising that I.Cox does not always outperform U.Cox, because the potential gain from data integration depends on the quality of both the original data (Dataset I) and the matching data (Dataset II). Indeed, I.Cox did not outperform U.Cox in Scenario 3 and Scenario 5 when CR1 is high and CMR is very low. In general, data integration is beneficial when the original data misses a substantial amount of true event records and thus may have inadequate or biased information for model estimation, and/or when the correct information that can be recovered by the matching data "exceeds" the accompanying noise/false information.

Figure 2 presents a visual comparison on the estimation of the baseline survival function from I.Cox and two naive Cox methods in different settings under MCAR. The true baseline survival curves are included. The I.Cox clearly performs the best overall, and in most cases

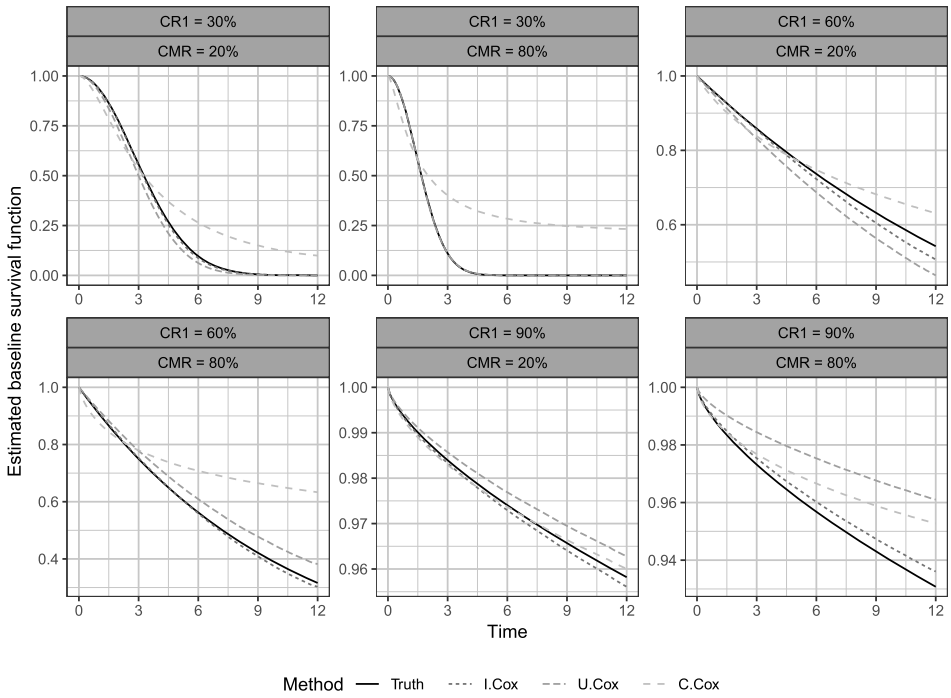

FIG. 2. *Mean of the estimated baseline survival function in various simulation settings when true record labels are missing at random* (*MCAR*).

the mean of its baseline survival function estimates over 1000 replications is close to the corresponding true curve. In contrast, both U.Cox and C.Cox, especially the latter, may lead to substantial overestimation of the survival probabilities. We have also checked the variation of the estimated survival curves from these methods and I.Cox performs satisfactorily. See Section 2 of the Supplementary Materials (Wang et al. (2020)) for an example plot of mean survival curves with pointwise 95% empirical confidence intervals and similar results under MNAR.

Table 3 reports the mean correct identification rate for all the subjects with uncertain records in Case 3 from the survival analysis with the I.Cox method. The rate ranges from 80.0% to 94.2% which means that the true records can be correctly identifies by the I.Cox model for at least 80% of subjects having uncertain records in all cases. We remark that in practice the main focus of such integrative analysis is still on the estimation and inference of $\boldsymbol{\beta}$; one should be cautious on using the estimated probabilities to identify the true records as the empirical evidence from our simulation study is certainly limited.

To check the performance of I.Cox in making inferences about the unknown covariate coefficients in comparison with U.Cox and O.Cox, we used bootstrap with 1000 bootstrap samples. The confidence intervals based on sample standard deviation and interquantile pro-

TABLE 3
Mean correct identification rates in percentage for subjects in Case 3 *under different simulation settings*

	Scenario					
	1	2	3	4	5	6
MCAR	85.7	89.5	83.1	83.8	80.4	80.0
MNAR	87.5	88.6	90.5	85.5	94.2	86.0

TABLE 4
Summaries of point estimate, standard error, and empirical coverage of 95% confidence intervals for two covariate coefficients

#	Method	$\hat{\beta}_1$	SE($\hat{\beta}_1$)	ESE($\hat{\beta}_1$)	CP($\hat{\beta}_1$)	$\hat{\beta}_4$	SE($\hat{\beta}_4$)	ESE($\hat{\beta}_4$)	CP($\hat{\beta}_4$)
					MCAR				
1	I.Cox	1.02	0.088	0.089	94.8	1.02	0.171	0.170	94.9
	U.Cox	0.94	0.086	0.088	87.0	0.93	0.162	0.168	92.2
	O.Cox	1.01	0.082	0.085	94.6	1.01	0.156	0.155	94.8
2	I.Cox	1.02	0.088	0.087	94.9	1.02	0.166	0.170	94.3
	U.Cox	1.01	0.095	0.094	95.7	1.01	0.176	0.184	94.0
	O.Cox	1.01	0.084	0.084	95.9	1.01	0.157	0.161	94.7
3	I.Cox	1.03	0.089	0.092	93.0	1.04	0.177	0.180	93.2
	U.Cox	0.96	0.086	0.087	91.4	0.96	0.164	0.163	94.1
	O.Cox	1.01	0.081	0.085	93.7	1.01	0.154	0.154	95.2
4	I.Cox	1.03	0.086	0.088	94.1	1.02	0.170	0.172	95.1
	U.Cox	1.05	0.099	0.096	94.5	1.04	0.192	0.189	95.2
	O.Cox	1.01	0.081	0.081	95.1	1.00	0.157	0.157	95.5
5	I.Cox	1.03	0.082	0.083	92.6	1.01	0.167	0.167	94.7
	U.Cox	1.02	0.084	0.085	94.3	1.03	0.164	0.161	95.3
	O.Cox	1.01	0.077	0.078	94.9	1.01	0.150	0.149	95.4
6	I.Cox	1.04	0.085	0.088	91.7	1.04	0.163	0.170	93.4
	U.Cox	1.09	0.106	0.105	87.7	1.10	0.199	0.196	93.0
	O.Cox	1.00	0.080	0.078	95.5	1.01	0.152	0.154	94.9
					MNAR				
1	I.Cox	1.02	0.088	0.088	95.6	1.02	0.172	0.168	95.6
	U.Cox	0.91	0.086	0.086	80.5	0.91	0.162	0.165	90.3
	O.Cox	1.01	0.082	0.085	94.6	1.01	0.156	0.155	94.8
2	I.Cox	1.02	0.088	0.089	94.1	1.02	0.167	0.173	94.7
	U.Cox	0.96	0.092	0.095	92.3	0.95	0.175	0.183	92.7
	O.Cox	1.01	0.084	0.084	95.9	1.01	0.157	0.161	94.7
3	I.Cox	1.04	0.088	0.090	92.4	1.04	0.170	0.169	94.2
	U.Cox	0.95	0.087	0.090	89.2	0.95	0.164	0.163	93.3
	O.Cox	1.01	0.081	0.085	93.7	1.01	0.154	0.154	95.2
4	I.Cox	1.03	0.087	0.089	94.1	1.02	0.170	0.173	95.0
	U.Cox	1.00	0.094	0.095	95.6	1.00	0.192	0.188	95.3
	O.Cox	1.01	0.081	0.081	95.1	1.00	0.157	0.157	95.5
5	I.Cox	1.04	0.082	0.083	91.8	1.04	0.161	0.162	95.1
	U.Cox	1.03	0.084	0.085	93.3	1.04	0.165	0.165	94.8
	O.Cox	1.01	0.077	0.078	94.9	1.01	0.150	0.149	95.4
6	I.Cox	1.06	0.086	0.087	89.9	1.06	0.164	0.175	93.3
	U.Cox	1.08	0.103	0.104	88.2	1.11	0.202	0.209	92.1
	O.Cox	1.00	0.080	0.078	95.5	1.01	0.152	0.154	94.9

SE: Standard error estimate; ESE: Empirical standard error from point estimates; CP: Coverage probability (%) of 95% confidence intervals.

duced estimates in good agreement, and we report those based on sample standard deviation. The results of point estimate, SE, and empirical coverage percentage for the coefficient of one continuous covariate and the binary covariate are summarized in Table 4. The bootstrap SE estimates appear to be close to the empirical SEs of the coefficient estimates in most of the settings. The coverage rate of 95% confidence intervals constructed from the SE estimates

and normal approximation is close to the nominal level in most cases. The worst cases are for β_1 under MNAR when the censoring of Dataset I is severe.

We have explored the asymptotic behaviors of the I.Cox estimator empirically. Following the original sample size decomposition given in Table 1, we increase the total sample size to two, four, eight and 16 times under each original setting. The results show that the mean of $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|$ decreases as the sample size increases, and the rate of convergence is approximately the square root of the sample size. We have also done simulation studies with fixed total sample size, and the results are similar to what we have presented. More details are available in Sections 3 and 4 of the Supplementary Materials (Wang et al. (2020)).

6. Survival analysis of the Connecticut data. We conducted a marginal screening analysis using I.Cox over the aforementioned 58 indicators of ICD-9 categories with three demographic variables, age, male, (vs. female) and White (vs. nonWhite) always included in the model. That is, each ICD-9 indicator was included as the fourth variable in the screening process. The inference results were obtained based on 1000 bootstrap samples, following the procedure detailed in Section 4.3. After the p-values of all the ICD-9 indicators were gathered from the marginal models, the Benjamini–Hochberg procedure (Benjamini and Hochberg (1995)) was applied to control the false discovery rate (FDR) at 5%. For comparison we repeated the same analysis using C.Cox, which ignored matching, and U.Cox which discarded all the uncertain events from matching.

The coefficient estimates for male and White from all the marginal models were significant at 5% level. Males were at significantly higher risk of death than females, and whites were at significantly higher risk than nonwhites. These findings of disparity in gender and race agree well with existing studies (e.g., Kung, Pearson and Wei (2005), Pena et al. (2012)). The age effect was less significant compared with gender and race. Most estimates for the coefficient of age from the marginal models were significantly greater than zero at 10% level, providing mild evidence that the survival time after suicide attempt tends to decrease with age for the patients in the study (age 15–30).

The screening analysis of ICD-9 codes revealed interesting and insightful results. By controlling the FDR at 5% for the results from each method, neither C.Cox nor U.Cox identified any significant ICD-9 category; in contrast, I.Cox identified four ICD-9 categories to be significantly associated with the risk of death after unsuccessful suicide attempt. The p-values for coefficient estimates of the four ICD-9 indicators are reported in the upper part of Table 5. The coefficient for ICD-9 code 292 was significantly positive, indicating that patients with drug-induced mental disorder had significantly higher risk than others after controlling for age, gender and race. Patients with borderline personality disorders (ICD-9 code 301) were also found to have a significantly higher risk of death. These results are supported by several studies, for example, Harris and Barraclough (1997), Lieb et al. (2004) and McGirr et al. (2007), among others. The I.Cox model also suggests that patients with dyspnea respiratory abnormalities and chest pain (ICD-9 code 786) had significantly higher risk. In the literature chest pain was reported to have positive association between psychiatric illness and panic disorder by Katon et al. (1988) and Fleet et al. (1996), respectively, which provided a possible explanation. Patients having postsurgical acquired absence of organ and other postprocedural status (ICD-9 code V45) were also under higher risk of death which may or may not be directly related to suicide.

We also checked the screening results without FDR control. The additional ICD-9 codes with unadjusted p-values under 5% are reported in the lower part of Table 5. For example, the effect of disorders of lipoid metabolism indicated by ICD-9 code 272 was identified by I.Cox. The positive association between suicidal behavior and lipid metabolism in depressive disorders was reported by Koponen et al. (2015). Overall, various mental disorders, psychological issues and drug dependence and abuse appear to be associated with shortened survival

TABLE 5

Selected ICD-9 categories by I.Cox and their brief descriptions. Columns 2–4 reports p-values (unadjusted) of coefficient estimates from I.Cox, C.Cox and U.Cox methods, respectively, where the significance is indicated by asterisk and the sign of estimates is given in subscripts

ICD-9	I.Cox	C.Cox	U.Cox	Description
		Significant ICD-9 codes under 5% FDR control		
786	0.000^{*}_{+}	0.004_{+}	0.002_{+}	Dyspnea, respiratory abnormalities and chest pain
V45	0.000^{*}_{+}	0.088_{+}	0.045_{+}	Postsurgical acquired absence of organ and other postprocedural status
292	0.001^{*}_{+}	0.007_{+}	0.007_{+}	Drug-induced mental disorders
301	0.002^{*}_{+}	0.069_{+}	0.066_{+}	Borderline personality disorder
		Additional ICD-9 codes with individual p-value under 5%		
780	0.010^{*}_{+}	0.178_{+}	0.169_{+}	Alteration of consciousness, convulsions and sleep disturbances
299	0.019^{*}_{+}	0.050^{*}_{+}	0.035^{*}_{+}	Pervasive developmental disorders
298	0.036^{*}_{+}	0.075_{+}	0.044^{*}_{+}	Other nonorganic psychoses
304	0.041^{*}_{+}	0.014^{*}_{+}	0.011^{*}_{+}	Drug dependence (such as opioid type, cocaine or cannabis)
966	0.041^{*}_{+}	0.140_{+}	0.129_{+}	Poisoning by anticonvulsants drugs
E98	0.043^{*}_{-}	0.046^{*}_{-}	0.065_{-}	Poisoning by analgesics, tranquilizers with undetermined reason
272	0.046^{*}_{+}	0.139_{+}	0.094_{+}	Disorders of lipoid metabolism
070	0.053_{+}	0.008^{*}_{+}	0.008^{*}_{+}	Chronic viral hepatitis C
V65	0.143_{+}	0.027^{*}_{+}	0.047^{*}_{+}	Counseling on substance use and abuse
874	0.338_{+}	0.029^{*}_{+}	0.063_{+}	Open wound of neck without mention of complication
969	0.421_{-}	0.027^{*}_{-}	0.024^{*}_{-}	Poisoning by antidepressants, antipsychotics and neuroleptics

I.Cox: Integrative Cox model; C.Cox: Regular Cox model fitted to dataset I before matching; U.Cox: Regular Cox model fitted to data with matched records removed.

time after unsuccessful suicide attempts. Therefore, by taking the data uncertainty into consideration and utilizing information from the second data source, the proposed I.Cox method reveals much more insightful results than the naive approaches.

We then turned our attention to joint modeling to check the estimation and predictive power of the joint model with all the identified ICD-9 categories. Table 6 summarizes the refitted I.Cox model with the three demographic variables (age, gender and race) and the four significant ICD-9 indicators identified from marginal screening. The coefficient estimates of male,

TABLE 6

Coefficient estimates from joint model including significant ICD-9 categories from marginal screening by I.Cox with FDR controlled at 5%

Predictor	$\hat{\beta}$	$\exp(\hat{\beta})$	$\mathrm{SE}(\hat{\beta})$	z	$\Pr(>\lvert z\rvert)$
Demographics					
Age	0.12	1.22	0.11	1.11	0.269
Male	1.81	6.11	0.32	5.63	0.000
White	2.18	8.86	0.38	5.78	0.000
ICD-9 Code					
786	1.54	4.67	0.36	4.23	0.000
V45	1.68	5.34	0.57	2.95	0.003
292	0.69	1.98	0.31	2.21	0.027
301	0.60	1.82	0.24	2.48	0.013

White and four ICD-9 indicators were all significantly positive at 5% level, consistent with the results from screening while coefficient estimate of age was insignificant. Because neither of the naive Cox methods suggested any significant ICD-9 category with FDR controlled at 5% from marginal screening, their joint models only included the three demographic variables. We checked that the coefficient estimates were all significant at 5%.

For the three joint models resulting from I.Cox and two naive methods, we performed an out-of-sample comparison analysis on their prediction performance. (We excluded age in the joint model of I.Cox since it was insignificant.) Specifically, we randomly split the patients into a training set and a test set. Patients having events and patients having censoring times were put in different strata so that the training set and the testing set had about the same censoring rate. For U.Cox, patients in Case 1 and Case 2 were randomly selected into training set with probability 0.8; for C.Cox, patients having certain event times (Case 1) and the remaining patients having censoring times in Dataset I were randomly selected into the training set, separately, with probability 0.8; for I.Cox, patients in Case 1, Case 2 and Case 3 were randomly selected into the training set, respectively, with probability 0.8, 0.8 and 1. As such, for each method the testing set only consisted of patients whose records are certain (Case 1 and Case 2) which makes an objective evaluation of fitted models possible. In each split a fitted model using the training set was used to predict the survival outcomes of patients in the testing set and classify them to a high risk group and a low risk group based on their risk scores. By comparing the group classification to the actual outcomes, we computed the receiver-operator characteristics (ROC) curve of the survival outcomes. The random split procedure was repeated 1000 times and the results were then averaged.

Figure 3 presents the ROC curves (on the left panel) and the curves (on the right panel) showing the relationship between the size of the high risk group and the proportion of subjects having observed suicide death that were captured in the high risk group. Here, the ROC curves are based on binary classification using the predicted risk scores; this is motivated by the clinical setting of a suicide prevention program where a group of patients with high risk of suicidal death is identified and subsequently monitored for suicide prevention. We remark that one may also use a time-dependent ROC analysis (Heagerty and Zheng (2005)) to quantify the prediction performance of a survival model. On average, the area under curve (AUC) was 0.825 for I.Cox, 0.761 for C.Cox and 0.757 for U.Cox. Therefore, the I.Cox model provided a better prediction on survival outcomes than both of the naive methods overall. The results on the right panel converted the ROC curves based on the censoring rate and showed that in order

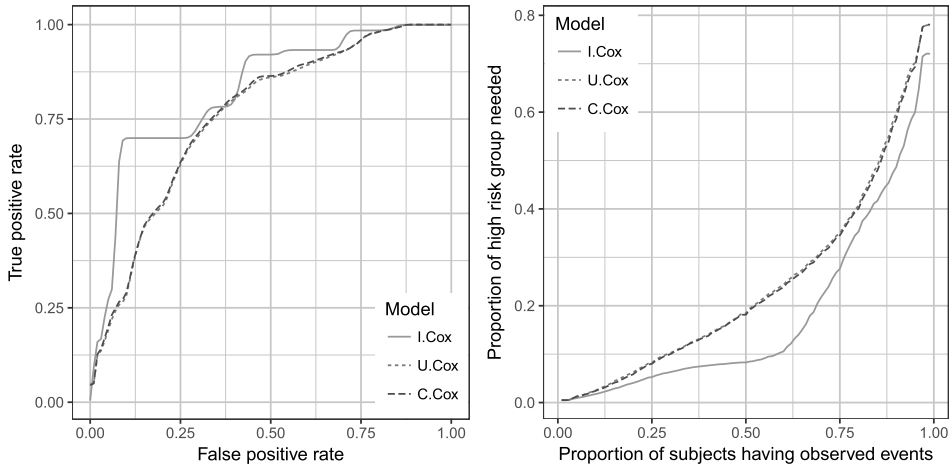

FIG. 3. *Out-of-sample comparison of the prediction performance on survival outcomes of I.Cox and the naive methods using random splitting.*

to capture 60% of the patients having observed events, the size of the high risk group needed was 10.6% on average for I.Cox, much less than the sizes, 23.8% and 24.3%, for C.Cox and U.Cox, respectively. Translating to the real clinical setting, this means that in order to capture 60% of the patients that would die, using I.Cox allows us to achieve this by monitoring only 10.6% of all the patients while using the native Cox methods will require 25%, a much larger population.

7. Discussion. We studied a general survival modeling setup with integrated data in which the survival outcome, that is, the time to certain event of interest, needed to be captured from multiple datasets through record linkage. Such problems are especially prevalent in medical research and healthcare analytics. Some commonly encountered events of interest include occurrence of disease, hospital readmission after discharge and death following certain diagnostics or treatment. However, patients' medical records are often scattered among many healthcare providers and government agencies. These datasets are generally deidentified to protect patient privacy, but, due to limitations in the current healthcare system, the deidentification of each dataset is often done separately before data integration, causing the aforementioned record linkage issues. To the best of our knowledge, building a healthcare information exchange system to connect healthcare providers is still largely an ongoing effort. Moreover, analyzing uncertain survival or time-to-event data is challenging due to censoring. When the censoring rate is high, for example, the event is rare, the information on event times can be quite limited and the results could become sensitive to inaccuracies and anomalies in event times. Therefore, properly handling the uncertainty in event times holds the key to ensure the validity of statistical inference.

Data integration with partial identifier is a double-edged sword in integrative statistical analysis. As a powerful tool to combine information from multiple sources, integrative analysis with probabilistic uncertainty modeling needs to be applied with care depending on the degree of imperfectness or noise. Imperfect data integration introduces noise and sometimes errors into the integrated data, the consequence of which could outweigh the potential gain in integrative data analysis. Although it is difficult to provide a specific guideline on when to use integrative analysis, we suggest that practitioner always perform out-of-sample analysis to evaluate and compare different methods whenever possible. To ensure the evaluation is objective, only the data without uncertainty should be used in testing.

Our case study has an additional distinguishing feature in that it is the outcome variable (survival time) that is obtained from data integration. This is, in contrast to other integrative data analysis settings, where usually predictors or features are obtained from multiple data sources. In our application we obtained insightful results on potential risk factors associated with death following suicide attempt which otherwise would have been missed by the naive approaches. Compared with the method of Snapinn (1998), our method is more attractive in that it does not require additional diagnostic variables or prior knowledge on the characterization of the truth indicators.

Several directions are worth pursuing for future research. The standard errors of the estimates cannot be easily produced along with the proposed estimation procedure. Although bootstrap is shown to perform well, the method would be more attractive in practice if a less computationally intensive inference approach were available. Under realistic settings of imperfect data linkage, the proposed method is shown to outperform several naive approaches. A natural theoretical question of interest is to quantify how the potential gain from data integration is associated with the quality of the original data and the match data. Our model framework is flexible and can be further extended to other survival models such as parametric survival models and competing risk models. Other extensions include the modeling of censoring times with covariates and the incorporation of certain known missing mechanism

of the label of true endpoint. In our application we adopted a marginal screening approach to identify important predictors; it would be interesting to extend the proposed method to conduct variable selection with high-dimensional predictors through regularized estimation. The rareness of suicide attempt brings many challenges in its modeling and prediction, including the occurrence of quasicomplete separation; these issues will need to be carefully studied in the future.

It is promising to further explore the trimmed likelihood formulation to better understand the robustness of the proposed approach and design better algorithm to target its global optimal solution. This formulation also sheds light on the consistency of the resulting estimator of the proposed method through the perspective of robust estimation and outlier detection. It shows that at least two conditions, regarding the proportion and magnitude of the "outliers"—fake records—are required. First, the proportion of patients with uncertain records should be under control, for example, $(n_2 + n_3)/n \to c$ for some $0 \le c < 1$ as $n \to \infty$. Second, the fake records have to be distinguishable from the true one; for example, for patient j we need $k^* = \arg\max_k r_{j,k}(\boldsymbol{\beta}^*)$ for n sufficiently large, where the k^*th record is the truth and $\boldsymbol{\beta}^*$ denotes the true coefficient vector. A thorough investigation of the theoretical properties of the proposed method along this direction is of interest.

Acknowledgments. Chen's research was supported in part by NSF Grants DMS-1613295 and IIS-1718798, and NIH Grant R01-MH112148. Aseltine's research was supported in part NIH Grant R01-MH112148.

SUPPLEMENTARY MATERIAL

Supplementary materials "Integrative survival analysis with uncertain event times in application to a suicide risk study" (DOI: 10.1214/19-AOAS1287 SUPP; .pdf). We provide detailed derivations of the likelihood formulation, additional supporting tables/figures from simulation studies, and discussions on the properties of the proposed method.

REFERENCES

ALBERT, A. and ANDERSON, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10. MR0738319 https://doi.org/10.1093/biomet/71.1.1

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BOHENSKY, M. A., JOLLEY, D., SUNDARARAJAN, V., EVANS, S., PILCHER, D. V., SCOTT, I. and BRAND, C. A. (2010). Data linkage: A powerful research tool with potential problems. *BMC Health Serv. Res.* **10** 346. https://doi.org/10.1186/1472-6963-10-346.

BOSTWICK, M. J., PABBATI, C., GESKE, J. R. and MCKEAN, A. J. (2015). Suicide attempt as a risk factor for completed suicide: Even more lethal than we knew. *Am. J. Psychiatr.* **173** 1094–1100.

BRESLOW, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30** 89–99.

CHEN, K. and ASELTINE, R. (2017). Using hospitalization and mortality data to target suicide prevention activities: A demonstration from Connecticut. *J. Adolesc. Health* **61** 192–197.

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276. MR0400509 https://doi.org/10.1093/biomet/62.2.269

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

EFRON, B. (1981). Censored data and the bootstrap. *J. Amer. Statist. Assoc.* **76** 312–319. MR0624333

FLEET, R. P., DUPUIS, G., MARCHAND, A., BURELLE, D., ARSENAULT, A. and BEITMAN, B. D. (1996). Panic disorder in emergency department chest pain patients: Prevalence, comorbidity, suicidal ideation, and physician recognition. *Am. J. Med.* **101** 371–380.

HADI, A. S. and LUCEÑO, A. (1997). Maximum trimmed likelihood estimators: A unified approach, examples, and algorithms. *Comput. Statist. Data Anal.* **25** 251–272. MR1478539 https://doi.org/10.1016/S0167-9473(97)00011-X

HARRIS, E. C. and BARRACLOUGH, B. (1997). Suicide as an outcome for mental disorders. A meta-analysis. *Br. J. Psychiatry* **170** 205–228.

HARRON, K., GOLDSTEIN, H. and DIBBEN, C. (2015). *Methodological Developments in Data Linkage*. Wiley.

HEAGERTY, P. J. and ZHENG, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61** 92–105. MR2135849 https://doi.org/10.1111/j.0006-341X.2005.030814.x

HOF, M. H. P. and ZWINDERMAN, A. H. (2012). Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Stat. Med.* **31** 4231–4242. MR3040077 https://doi.org/10.1002/sim.5498

HOF, M. H. P. and ZWINDERMAN, A. H. (2015). A mixture model for the analysis of data derived from record linkage. *Stat. Med.* **34** 74–92. MR3286240 https://doi.org/10.1002/sim.6315

JAMSHIDIAN, M. and JENNRICH, R. I. (2000). Standard errors for EM estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 257–270. MR1749538 https://doi.org/10.1111/1467-9868.00230

KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1924807 https://doi.org/10.1002/9781118032985

KATON, W., HALL, M. L., RUSSO, J., CORMIER, L., HOLLIFIELD, M., VITALIANO, P. P. and BEITMAN, B. D. (1988). Chest pain: Relationship of psychiatric illness to coronary arteriographic results. *Am. J. Med.* **84** 1–9. https://doi.org/10.1016/0002-9343(88)90001-0

KOPONEN, H., KAUTIAINEN, H., LEPPÄNEN, E., MÄNTYSELKÄ, P. and VANHALA, M. (2015). Association between suicidal behaviour and impaired glucose metabolism in depressive disorders. *BMC Psychiatry* **15** 163. https://doi.org/10.1186/s12888-015-0567-x

KUNG, H.-C., PEARSON, J. L. and WEI, R. (2005). Substance use, firearm availability, depressive symptoms, and mental health service utilization among White and African American suicide decedents aged 15 to 64 years. *Ann. Epidemiol.* **15** 614–621.

LIEB, K., ZANARINI, M. C., SCHMAHL, C., LINEHAN, M. M. and BOHUS, M. (2004). Borderline personality disorder. *Lancet* **364** 453–461.

MCGIRR, A., PARIS, J., LESAGE, A., RENAUD, J. and TURECKI, G. (2007). Risk factors for suicide completion in borderline personality disorder: A case-control study of cluster B comorbidity and impulsive aggression. *J. Clin. Psychiatry* **68** 721–729.

MEIER, A. S., RICHARDSON, B. A. and HUGHES, J. P. (2003). Discrete proportional hazards models for mismeasured outcomes. *Biometrics* **59** 947–954. MR2025118 https://doi.org/10.1111/j.0006-341X.2003.00109.x

MENG, X.-L. and RUBIN, D. B. (1991). Using EM to obtain asymptotic variance–covariance matrices: The SEM algorithm. *J. Amer. Statist. Assoc.* **86** 899–909.

MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503 https://doi.org/10.1093/biomet/80.2.267

MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. MR1803168 https://doi.org/10.2307/2669386

NADARAJAH, S. and KOTZ, S. (2006). R programs for truncated distributions. *J. Stat. Softw.* **16** 1–8.

NEYKOV, N., FILZMOSER, P., DIMOVA, R. and NEYTCHEV, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Comput. Statist. Data Anal.* **52** 299–308. MR2409983 https://doi.org/10.1016/j.csda.2006.12.024

PATRICK, A. R., MILLER, M., BARBER, C. W., WANG, P. S., CANNING, C. F. and SCHNEEWEISS, S. (2010). Identification of hospitalizations for intentional self-harm when E-codes are incompletely recorded. *Pharmacoepidemiol. Drug Saf.* **19** 1263–1275. https://doi.org/10.1002/pds.2037.

PENA, J. B., MATTHIEU, M. M., ZAYAS, L. H., MASYN, K. E. and CAINE, E. D. (2012). Co-occurring risk behaviors among White, Black, and Hispanic US high school adolescents with suicide attempts requiring medical attention, 1999–2007: Implications for future prevention initiatives. *Soc. Psychiatry Psychiatr. Epidemiol.* **47** 29–42. https://doi.org/10.1007/s00127-010-0322-z.

PRITCHARD, C. and HANSEN, L. (2015). Examining undetermined and accidental deaths as source of 'under-reported-suicide' by age and sex in twenty western countries. *Community Ment. Health J.* **51** 365–376. https://doi.org/10.1007/s10597-014-9810-z.

R DEVELOPMENT CORE TEAM (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

RICHARDSON, B. A. and HUGHES, J. P. (2000). Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. *Biostatistics* **1** 341–354.

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880. MR0770281

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519 https://doi.org/10.1002/9780470316696

SNAPINN, S. M. (1998). Survival analysis with uncertain endpoints. *Biometrics* **54** 209–218.

SUOMINEN, K., ISOMETSÄ, E., SUOKAS, J., HAUKKA, J., ACHTE, K. and LÖNNQVIST, J. (2004). Completed suicide after a suicide attempt: A 37-year follow-up study. *Am. J. Psychiatr.* **161** 562–563.

TANCREDI, A. and LISEO, B. (2015). Regression analysis with linked data: Problems and possible solutions. *Statistica* **75** 19–35.

TØLLEFSEN, I. M., THIBLIN, I., HELWEG-LARSEN, K., HEM, E., KASTRUP, M., NYBERG, U., ROGDE, S., ZAHL, P.-H., ØSTEVOLD, G. et al. (2016). Accidents and undetermined deaths: Re-evaluation of nationwide samples from the Scandinavian countries. *BMC Public Health* **16** 449. https://doi.org/10.1186/s12889-016-3135-5.

WANG, W., ASELTINE, R., CHEN, K. and YAN, J. (2020). Supplement to "Integrative survival analysis with uncertain event times in application to a suicide risk study." https://doi.org/10.1214/19-AOAS1287SUPP.

WINGLEE, M., VALLIANT, R. and SCHEUREN, F. (2005). A case study in record linkage. *Surv. Methodol*. **31** 3–11.

XU, C., BAINES, P. D. and WANG, J.-L. (2014). Standard error estimation using the EM algorithm for the joint modeling of survival and longitudinal data. *Biostatistics* **15** 731–744. https://doi.org/10.1093/biostatistics/kxu015.

ZHAO, Q., SHI, X., XIE, Y., HUANG, J., SHIA, B. and MA, S. (2015). Combining multidimensional genomic measurements for predicting cancer prognosis: Observations from TCGA. *Brief. Bioinform*. **16** 291–303. https://doi.org/10.1093/bib/bbu003.