Survival Analysis with Multiple Noisy Labels

Donna Tjandra
Computer Science and Engineering
University of Michigan
Ann Arbor, Michigan
dotjandr@umich.edu

Jenna Wiens
Computer Science and Engineering
University of Michigan
Ann Arbor, Michigan
wiensj@umich.edu

***Abstract*—In many applications, collecting ground truth labels is labor intensive and costly. Thus, researchers often turn to pragmatic labeling tools based on heuristics, at the potential cost of introducing noise. When multiple different labeling tools are used, we find ourselves in the setting of *multiple noisy labels*. Previous work studying supervised learning with multiple noisy labels focuses on classification and proposes different strategies to aggregate labels. Here, we move beyond classification and study multiple noisy labels in the context of time-to-event prediction (i.e., survival analysis). As we show, survival analysis presents additional challenges when learning from multiple noisy labels since outcomes may be censored. We formalize the problem of multiple noisy labels in survival analysis and propose a novel approach. Our approach leverages a reference set with both noisy and ground truth labels to model the noisy time-to-event distribution and their associated errors and then uses these distributions to predict the ground truth time-to-event distribution. When predicting sepsis onset in the MIMIC-III dataset, our approach more accurately estimates time-to-events compared to the next best baseline (median time-to-event error across 10 replications: 14.5 hours [interquartile range 13.25-15.75] vs. 17.50 hours [interquartile range 16.25-18.00]). CODE**

***Index Terms*—Survival Analysis, Time-to-Event Prediction, Noisy Labels, Multiple Labelers, Health Application**

I. Introduction

Motivation. In survival analysis, one aims to estimate the probability of an event (e.g., death) occurring over time. Training survival analysis models requires accurately labeled time-to-events (TTEs). While obtaining accurate TTEs can be difficult across many domains, we focus on TTE prediction in healthcare as our motivating example. In healthcare, the TTE sometimes corresponds to the time of disease onset. However, for some diseases, TTEs can be difficult to identify without manual chart review by a clinical expert (e.g., sepsis [1]–[3] and Alzheimer's disease [4]), making it challenging to efficiently label large electronic health record datasets for model training. As a result, automated pragmatic labeling tools based on the structured components of the electronic health record are often used instead [5], [6]. For example, one could label sepsis onset as 1) when the CDC (Center for Disease Control and Prevention) definition is met [7] or 2) when the Sepsis3 definition is met [8]. However, this introduces two sources of error. First, both labeling approaches can incorrectly identify whether or not sepsis onset occurs. Second, even among patients with sepsis who are correctly identified as having sepsis, the time of disease onset can still be incorrect. *In the absence of ground truth TTE labels for the majority of patients with respect to some outcome, one can potentially learn a more accurate survival model by combining noisy proxies from multiple different labelers/annotators.* We refer to this setting as **survival analysis with multiple noisy labels**.

Current Gaps. Work studying multiple noisy labels focuses almost entirely on classification, where approaches generally aggregate the noisy labels from a dataset at 1) preprocessing time [9], [10], or 2) inference time [11]–[14]. Survival analysis is a more complex setting compared to standard classification since individual examples may have their outcomes censored (i.e., we only observe an individual up to a certain point and it is unclear whether or when the event occurred thereafter). Generally, rather than discarding these examples during training, survival analysis specifically leverages the information contained in the fact that we know the event did *not* occur up to a certain point. Work addressing noisy labels from the survival analysis literature is largely limited to the single labeler case [15]. In survival analysis with multiple noisy labels, where noisy labels correspond to noisy TTEs from different labelers, one could naïvely aggregate during preprocessing by taking the average of the noisy TTEs and then use this aggregate as ground truth during training. **However, this requires assumptions on the relationship between the ground truth and noisy TTEs. Moreover, it is not immediately obvious how to aggregate censored outcomes.** Instead, one could aggregate at inference time by first using standard techniques to learn to model each labeler separately and then aggregating (e.g., averaging) the labeler-specific predictions. Doing so addresses issues around aggregating censored labels since we can still generate predictions for labelers with censored outcomes. However, this method still requires assumptions on the relationship between the noisy and ground truth TTEs (e.g., the ground truth TTE is the average of the noisy ones).

Our Idea. To address the limitations of past work, we introduce a novel approach for survival analysis with multiple noisy labels. Applied to a variety of experimental settings involving both synthetic and real data, our approach is more robust than adaptations of approaches from classification with respect to the rate of censorship in the data at training, and it does not require assumptions on how the noisy and ground truth TTEs are related. Our approach leverages a small reference set, i.e., a small subset of data for which we have expert-labeled TTEs that serve as ground truth. Reference sets can

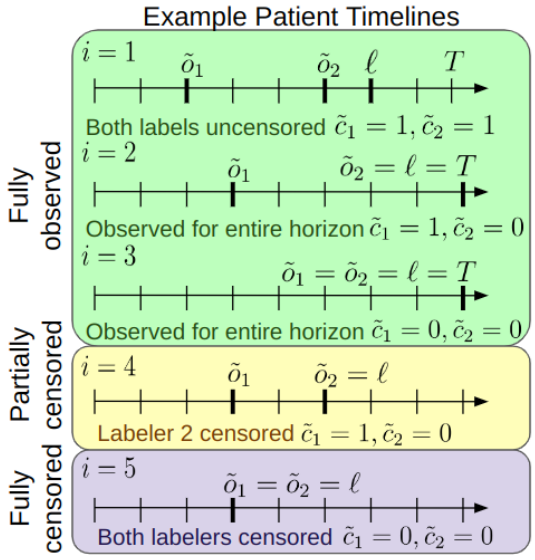

Fig. 1: Demonstration with example timelines. Instead of observing ground truth time-to-events (TTEs) in our dataset, we observe $m > 1$ noisy proxies. In this figure we have $m = 2$ labelers for five example individuals. Notation) ℓ: last time observed, $\tilde{c}_j$: indicator for whether an event was recorded for labeler j $\tilde{o}_j$: observed time for labeler j, which is ℓ if $\tilde{c}_j == 0$ or noisy TTE $\tilde{e}_j$ otherwise; T: length of horizon.

be constructed by randomly selecting a subset of individuals in the training data and then having a subject expert assign ground truth labels to these individuals via manual review. While this is still associated with a cost, it is significantly less costly than labeling the entire dataset, in many settings. Overall, our contributions are as follows.

- We formalize the multiple noisy labels problem in the context of survival analysis
- We adapt existing approaches from the multiple noisy labels literature in classification and identify their shortcomings in the survival analysis setting
- We propose a novel approach for survival analysis with multiple noisy labels and show that our approach is more robust than the baselines across a variety of settings

II. Methods

We formalize the multiple noisy labels problem in the context of survival analysis and describe our proposed approach.

A. Notation and Problem Setup

In survival analysis with clean/correct labels, our dataset is: $\mathcal{D} = \{\mathbf{x}^{(i)}, c^{(i)}, o^{(i)}\}_{i=1}^{n}$. Here, n is the number of individuals, $\mathbf{x} \in \mathbb{R}^d$ is a feature vector, d is the number of features, and c is a binary indicator for whether the event occurred (i.e., $c = 0$ if the event is censored and $c = 1$ otherwise). The observed time, o, corresponds to the TTE, denoted with e, if $c = 1$. If $c = 0$, o corresponds to the last time of observation, ℓ (i.e., we know that the event did not occur by time o, but we do not know what happened after time o).

We consider a setup in which we aim to predict survival within a fixed time horizon [16]. Given an event of interest, we aim to predict: 1) whether the event occurs within T time steps (i.e., $P(e \in \{1, 2, ..., T\}|\mathbf{x})$) and 2) the probability of the event occurring at each time point $1, 2, ..., T$ given that it occurred within T time steps (i.e., $P(e = t|\mathbf{x}, e \in \{1, 2, ..., T\})$ for $t = 1, 2, ..., T$). Using these predictions, the corresponding survival function is $\hat{S}(t|\mathbf{x}) = 1 - \sum_{u=1}^{t} \hat{P}(e = u|\mathbf{x}) = 1 - \sum_{u=1}^{t} \hat{P}(e = u|\mathbf{x}, e \in \{1, 2, ..., T\}) P(e \in \{1, 2, ..., T\}|\mathbf{x})$, and the predicted TTE is the median value of $\hat{S}(t|\mathbf{x})$ (i.e., $\hat{e} = argmin_t \hat{S}(t|\mathbf{x}) \leq 0.5$) [17], where $\hat{\cdot}$ denotes a prediction. Unless otherwise indicated, let superscripts in parentheses denote individual indices (e.g., $\mathbf{x}^{(i)}$) and subscripts denote indices into vectors (e.g., x_k). Where convenient, we drop the indexing superscripts. We assume that $e \sim D(f(\mathbf{x}))$, where f can be any function, and D is a distribution. We make no assumptions on D other than $f(\mathbf{x})$ is the median of D and that D can be approximated by the empirical survival distribution.

In the multiple noisy labels setting with m labelers, instead of observing o in the dataset, we observe $\tilde{\mathbf{o}} \in \{1, 2, ..., T\}^m$, a vector of observed times for each labeler. Each entry $\tilde{o}_j$ corresponds to the observed time for labeler $j \in \{1, 2, ..., m\}$, which can be a noisy TTE or ℓ. If we observe a noisy TTE for labeler j, denoted $\tilde{e}_j$, we assume that it can be written as $\tilde{e}_j = e + \tilde{g}_j(\mathbf{x})$ where $\tilde{g}_j(\mathbf{x})$ is a labeler-specific error value that can be instance-dependent. Taken together, $\tilde{\mathbf{o}} = (\tilde{o}_1, \tilde{o}_2, ..., \tilde{o}_m)$. Similarly, instead of observing c, we observe $\tilde{\mathbf{c}} \in \{0, 1\}^m$, where each $\tilde{c}_j$ represents an indicator for event occurrence for labeler j.

We define three ways that individuals can be considered based on censorship (**Figure 1**). Let $\mathcal{D}^{(i)} = (\mathbf{x}^{(i)}, \tilde{\mathbf{c}}^{(i)}, \tilde{\mathbf{o}}^{(i)})$. The first is **fully observed**: $\mathcal{U} = \{\mathcal{D}^{(i)} | (\tilde{c}_j^{(i)} > 0 \forall j) \vee (\max_j o_j^{(i)} == T)\}$, which includes those who have a recorded TTE for all labelers (i.e., uncensored) or were observed for the entire prediction horizon with potentially no event for some of the labelers (e.g., administratively censored). Note that, since we are only concerned with predicting event occurrence within some time horizon, administratively censored individuals have a fully observed outcome (i.e., no event for the labelers with no recorded event). This is illustrated by individuals $i = 2$ and $i = 3$ in **Figure 1**. Both are observed for the entire horizon but individual $i = 2$ does not have a TTE for labeler $j = 2$, and individual $i = 3$ does not have a TTE for both labelers. The second is **fully censored**: $\mathcal{C} = \{\mathcal{D}^{(i)} | (\tilde{c}_j^{(i)} == 0) \wedge (\max_j o_j^{(i)} < T)\}$, which includes those whose were not observed for the entire prediction horizon and have censored TTEs for all labelers. The third is **partially censored**: $\mathcal{P} = \{\mathcal{D}^{(i)} | ((max_j \tilde{c}_j^{(i)} - min_j \tilde{c}_j^{(i)}) > 0) \wedge (\max_j o_j^{(i)} < T)\}$, which includes those who were not observed for the entire horizon and whose noisy TTEs are censored for at least one, but not all labelers. We use the common assumption [18], that censorship is independent of $\mathbf{x}$ (i.e., $P(\mathcal{D}^{(i)} \in \mathcal{C}) = P(\mathcal{D}^{(i)} \in \mathcal{C}|\mathbf{x}^{(i)})$ and $P(\mathcal{D}^{(i)} \in \mathcal{P}) = P(\mathcal{D}^{(i)} \in \mathcal{P}|\mathbf{x}^{(i)})$). We also assume that ℓ is uncorrelated with e.

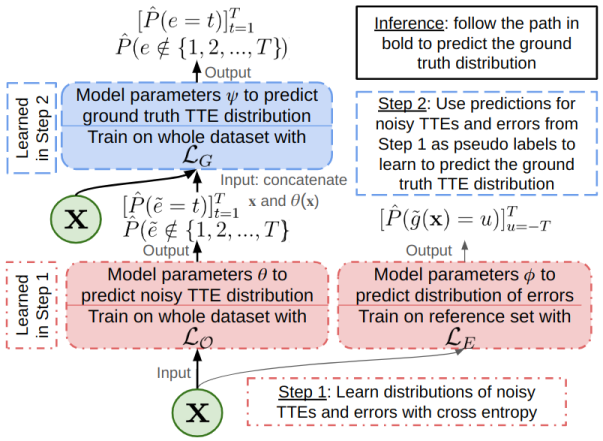

Fig. 2: Overview of proposed approach. Based on **x**, we first predict the noisy TTEs and their errors as intermediate outputs. We then concatenate the noisy TTE prediction with **x** and map this to a prediction of the ground truth TTE as the final output of the approach. The approach can be implemented using the three components shown in the shaded boxes that are trained as described in Step 1a, Step 1b, and Step 2.

B. Proposed Approach

Our approach (**Figure 2**, **Appendix A**) consists of two steps. In the first step, we train two models to predict, for a given individual, a) the noisy TTEs and b) the errors of the noisy TTEs. By predicting the noisy TTE errors, we assume that the same labelers are used across the dataset and are somewhat systematic (i.e., predictable) in their labeling behaviors. From these two predictions, we aim to recover the ground truth TTE, which can then be used as a pseudo-label in the second step. In the second step, we train a third model to map the input features and noisy TTE predictions from the first step to the ground truth TTE, which we use to predict the survival function. *Throughout, we use cross entropy loss to learn the respective distributions since it does not require any assumptions on the forms of the distributions other than that they can be approximated by the empirical distributions.*

To learn to predict the noisy TTE errors, we assume access to a small subset of randomly chosen individuals for whom we have both noisy and expert-assigned labels (i.e., an unbiased reference set). In practice, there are many settings for which reference sets are available. With health data, it is often possible to have a clinician review a small subset of randomly chosen patient charts to obtain ground truth.

We now provide more details for each step.

Step 1a: Noisy TTE Prediction. Here, we learn to predict the distribution of noisy TTEs as the output of the model parameterized by θ. Let $\tilde{e}$ denote a noisy TTE sampled from $\{\tilde{e}_1, \tilde{e}_2, ..., \tilde{e}_m\}$ with uniform probability given **x**. Note that the $\tilde{e}_j$'s themselves are not necessarily from the same distribution. For example, given **x**, $\tilde{e}_1$ may come from N(1, 2) while $\tilde{e}_2$ may be from N(3, 1). Predicting the distribution of noisy TTEs then becomes predicting $[P(\tilde{e}=1|\mathbf{x}), P(\tilde{e}=2|\mathbf{x}), ..., P(\tilde{e}=T|\mathbf{x}), P(\tilde{e} \notin \{1,2,...,T\}|\mathbf{x})] \in [0,1]^{T+1}$.

For labeler j, censored and uncensored labels are handled in ways similar to past work [19]–[22]. For uncensored labels, we can use the corresponding $\tilde{o}_j^{(i)}$ as supervision with cross entropy loss through $\mathcal{L}_{\mathcal{U}}$. For censored labels, we minimize the probability of event occurrence before the time of follow up using $\mathcal{L}_{\mathcal{C}}$ since we have no other information.

$$\mathcal{L}_{\mathcal{U}}(i,j) = -log\left(\hat{P}(\tilde{e}^{(i)} = \tilde{o}_j^{(i)}|\mathbf{x}^{(i)})\right)$$

$$\mathcal{L}_{\mathcal{C}}(i,j) = -log\left(1 - \sum_{t=1}^{\tilde{o}_j^{(i)}} \hat{P}\left(\tilde{e}^{(i)} = t|\mathbf{x}^{(i)}\right)\right)$$

We can write the objective function over all individuals and all labelers with $\mathcal{L}_{\mathcal{O}}$ (where $\mathbb{I}$ is the indicator function)

$$\mathcal{L}_{\mathcal{O}} = \sum_{i=0}^{n}\sum_{j=1}^{m} \mathbb{I}(\tilde{c}_j^{(i)} == 1)\mathcal{L}_{\mathcal{U}}(i,j) + \mathbb{I}(\tilde{c}_j^{(i)} == 0)\mathcal{L}_{\mathcal{C}}(i,j)$$

For individuals in $\mathcal{U}$, we use the first term of $\mathcal{L}_{\mathcal{O}}$. For individuals in $\mathcal{C}$, we use the second term of $\mathcal{L}_{\mathcal{O}}$. For individuals in $\mathcal{P}$, we use both terms of $\mathcal{L}_{\mathcal{O}}$.

Step 1b: Error Prediction. Here, we learn to predict the distribution of errors as the output of the model parameterized by ϕ. To learn the distribution of errors, we use a similar method to learning the distribution of noisy TTEs. However, instead of using the entire dataset, we only use the reference set, denoted with $\mathcal{A}$. For such individuals, we have $(\mathbf{x}, \tilde{\mathbf{c}}, \tilde{\mathbf{o}}, c, o)$ and assume that $\mathcal{A} \subset \mathcal{U}$. Using $\mathcal{A}$, we train ϕ with $\mathcal{L}_E$. Note that, since we assume $\mathcal{A} \subset \mathcal{U}$, individuals with $c == 0$ are those who do not experience the event by the end of the prediction horizon. $\mathcal{L}_E$ only considers those who experience the event when learning the distribution of errors. Those who do not experience the event are used in Step 2 to learn $P(e \notin \{1,2,...,T\})$. Similar to $\tilde{e}$, let $\tilde{g}(\mathbf{x})$ denote the error of a noisy TTE sampled from $\{\tilde{g}_1(\mathbf{x}), \tilde{g}_2(\mathbf{x}), ..., \tilde{g}_m(\mathbf{x})\}$ with uniform probability. Like $\tilde{e}_j$, the $\tilde{g}_j(\mathbf{x})$ values are not necessarily drawn from the same distribution. Predicting the TTE error distribution then becomes predicting $[P(\tilde{g}(\mathbf{x}^{(i)}) = -T|\mathbf{x}), P(\tilde{g}(\mathbf{x}^{(i)}) = -T+1|\mathbf{x}), ..., P(\tilde{g}(\mathbf{x}^{(i)}) = T|\mathbf{x})] \in [0,1]^{2T+1}$, where $\Delta o_j^{(i)} = o^{(i)} - \tilde{o}_j^{(i)}$.

$$\mathcal{L}_E = -\sum_{i \in \mathcal{A}}\sum_{j=1}^{m} \mathbb{I}(c^{(i)} == 1) log\hat{P}\left(\tilde{g}(\mathbf{x}^{(i)}) = \Delta o_j^{(i)}|\mathbf{x}^{(i)}\right)$$

Step 2: Ground Truth TTE Prediction. In Step 2, we aim to predict the ground truth TTE distribution as the output of the model parameterized by ψ, where we map the features and our prediction for the distribution of noisy TTEs to a prediction of the ground truth TTE distribution (i.e., $[P(e=1|\mathbf{x}), P(e=2|\mathbf{x}), ..., P(e=T|\mathbf{x}), P(e \notin \{1,2,...,T\}|\mathbf{x})] \in [0,1]^{T+1}$). Since we lack reference TTEs for the majority of the dataset, we cannot apply cross entropy loss as we did when learning the distribution of noisy TTEs. However, since we learned to predict the distributions of the noisy TTEs and errors, we can use these predictions to estimate the ground truth TTE, which can then be used as pseudo-labels for this

step. Given the pseudo-labels, we use a cross-entropy loss to learn the ground truth survival distribution. This is represented in the first term of $\mathcal{L}_G$ below.

$$\mathcal{L}_G = \frac{1}{n}\sum_{i=1}^{n} -log\left(\hat{P}(e^{(i)} = \hat{e}^{(i)}_{avg} | e^{(i)} \leq T, \mathbf{x}^{(i)})\right)$$
$$- \mathbb{I}(\mathcal{D}^{(i)} \in \mathcal{A} \wedge c^{(i)} == 0) log\left(\hat{P}(e \notin \{1, 2, ..., T\} | \mathbf{x}^{(i)})\right)$$

Here, $\hat{e}_{avg} = \lfloor \sum_t t\hat{P}(\tilde{e} = t | \tilde{e} \leq T, \mathbf{x}) + \sum_\epsilon \epsilon \hat{P}(\tilde{g}(\mathbf{x}) = \epsilon | \mathbf{x}) \rfloor$ is the mean predicted TTE [23], offset by the error prediction, whose value is clipped to 1 or T if needed. The term $e \leq T$ is shorthand for $e \in \{1, 2, ..., T\}$, which describes the event occurring within the prediction horizon with $P(e \leq T) = 1 - P(e \notin \{1, 2, ..., T\})$. The term $\hat{P}(e = t | e \leq T) = \hat{P}(e = t)/\hat{P}(e \leq T)$ is the conditional probability of event occurrence. The second term of $\mathcal{L}_G$ provides supervision over the probability of the event occurring outside of the horizon, using individuals from $\mathcal{A}$. Note that we use e instead of $\hat{e}_{avg}$ for individuals in the reference set in $\mathcal{L}_G$ and that θ, ϕ, and ψ can be implemented with any architecture (e.g., feed forward network) that outputs a probability distribution. Thus, we expect that the time and space complexity of our approach with respect to training, storage and inference will be the same as any standard implementation of these architectures.

While we could rely on only the pseudo-labels from Step 1 to construct the survival curve, this will likely lead to miscalibrated predictions. This is because the distributions from which the noisy TTEs are drawn do not necessarily match that of the ground truth TTE. For example, in healthcare, the time when a patient is billed for a diagnostic code may not directly match the natural progression of the respective condition [24]. Thus, Step 2 serves as a re-calibration step.

In contrast to adaptations from classification approaches, our approach incorporates censored individuals at each stage and learns how to aggregate the noisy TTEs. We hypothesize that, because of this, we will outperform existing approaches designed for classification. In the next section, we empirically explore this hypothesis. But first, we provide some practical guidelines based on a theoretical analysis of our approach.

C. Practical Guidelines on Expected Error

Here, we explore in theory how the approach could be affected by different levels of label noise with respect to event occurrence. Based on our findings, we propose practical guidelines for the use of the approach. Empirically, we also expect the approach to depend on the size of the reference set, whether the reference set is biased (i.e., certain types of individuals are more likely to be included), and how difficult the outcome is to learn.

1) TTE Prediction Error in Positive Individuals: In positive individuals (i.e., those who experience the event within the prediction horizon), we aim to minimize $|e - \hat{e}|$. If we accurately learn the distributions of noisy TTEs and noisy TTE errors for true positive individuals, then $|e - \hat{e}|$ will mainly depend on false negatives (i.e., positive individuals who are identified as negative by some labelers). We can express $P(|e - \hat{e}| \geq \varepsilon)$ for positive individuals with TTE e and error tolerance value ε as

$$P(|e - \hat{e}| \geq \varepsilon) \text{ (when considering positive individuals)}$$
$$\leq \frac{1}{\varepsilon}\alpha T \text{ (Justification in } \textbf{Appendix B}\text{)}$$

where $\alpha = \prod_{j=1}^{m} \alpha_j$ is the overall false negative rate, and α_j is the false negative rate for labeler j. In other words, *the probability that the absolute TTE error is above some value depends on the false negative rate*, where a lower false negative rate will lower this probability. As a result, if we aim to train a model with target values ε and $0 < p < 1$ such that $P(|e - \hat{e}| \geq \varepsilon) \leq p$, then checking whether $\alpha \leq \frac{\varepsilon p}{T}$ could serve as a guideline for whether ε and p are realistic.

2) Prediction Error for Event Occurrence: Similarly, for all individuals, we aim to minimize $|P(e \notin \{1, 2, ..., T\}) - \hat{P}(e \notin \{1, 2, ..., T\})|$. We learn this probability implicitly during Step 1a and then fine tune it on the reference set in Step 2. As a result, we expect that the error will mainly depend on false negative and false positive individuals. Therefore, we can express $P(|P(e \notin \{1, 2, ..., T\}) - \hat{P}(e \notin \{1, 2, ..., T\})| \geq \delta)$ for error tolerance value δ as

$$P(|P(e \notin \{1, 2, ..., T\}) - \hat{P}(e \notin \{1, 2, ..., T\})| \geq \delta)$$
$$\leq \frac{1}{\delta}(\rho\alpha + (1 - \rho)\beta) \text{ (Justification in } \textbf{Appendix B}\text{)}$$

where β is the false positive rate analog of α, and ρ is the positive rate. In other words, $P(|P(e \notin \{1, 2, ..., T\}) - \hat{P}(e \notin \{1, 2, ..., T\})| \geq \delta)$ *depends on a combination of the false negative rate, false positive rate and positive rate*. As a result, if we aim to train with target values δ and $0 < q < 1$ such that $P(|P(e \notin \{1, 2, ..., T\}) - \hat{P}(e \notin \{1, 2, ..., T\})| \geq \delta) \leq q$, then checking whether $\rho\alpha + (1 - \rho)\beta \leq \delta q$ could serve as a guideline for whether δ and q are realistic.

III. Experimental Setup

We empirically explored how our proposed approach compares with approaches for learning with multiple noisy labels adapted from classification in a variety of tasks using both synthetic and real datasets. Here, we describe the datasets, baselines, and evaluation. More details on the overall implementation are with our code and **Appendices C and D**.

A. Datasets

Synthetic Dataset. Our synthetic dataset is generated as follows. We consider a horizon length of 200 (i.e., $T = 200$) and generate 5,000 individuals with 100 features per individual. Individuals are labeled as positive or negative based on y (e.g., if y is above some threshold then the individual is positive). After, we obtain e for positive individuals by drawing from a normal distribution, with mean $f(\mathbf{x})$. We draw e from a normal distribution here to provide contrast with the skewed distribution we use to generate the noisy TTEs described below. This allows us to emphasize the importance of Step 2

in our experiments. I is the identity matrix; $\sigma()$ is a sigmoid. Data generation is summarized below.

$$\mathbf{x}^{(i)} \sim N(\mathbf{0}, I)^{100} \text{ (feature vector)}$$
$$\mathbf{w}_y, \mathbf{w}_e \sim N(\mathbf{0}, I)^{100} \text{ (coefficients)}$$
$$y^{(i)} = \sigma(\mathbf{w}_y \cdot \mathbf{x}^{(i)}) \text{ (risk of event)}$$
$$z^{(i)} = \mathbf{w}_e \cdot \mathbf{x}^{(i)} \text{ (unscaled expected TTE)}$$
$$f(\mathbf{x}^{(i)}) = T\left[\frac{z^{(i)} - min_k z^{(k)}}{max_k z^{(k)} - min_k z^{(k)}}\right] \text{ (expected TTE)}$$

Given the above components, our noise generating procedure is described using the steps below for each labeler $j = 1, 2, 3$ (i.e., $m = 3$). Whether labeler j mislabels event occurrence is determined through y_n^j (e.g., individual is mislabeled if y_n^j is above a threshold). Noisy TTEs are generated according to the offset values in $\boldsymbol{\epsilon}$ so that the distribution of noisy labels does not match the ground truth TTE distribution.

$$\mathbf{w}_n^j \sim N(\mathbf{0}, I)^{100} \text{ (coefficients)}$$
$$y_n^{j(i)} = \sigma(\mathbf{w}_n^j \cdot \mathbf{x}^{(i)}) \text{(risk of mislabeling event occurrence)}$$
$$\boldsymbol{\epsilon} = [-20, -20, 40] \text{ (labeler errors for TTE)}$$
$$\tilde{g}_j(\mathbf{x}) = \begin{cases} \epsilon_j, & x_1 < 0 \\ -\epsilon_j, & x_1 \geq 0 \end{cases}$$

By this step, before adding censorship, the values of $\tilde{\mathbf{o}}$ and $\tilde{\mathbf{c}}$ for each individual, for labeler j are as follows.

- True positives: $\tilde{o}_j = e + \tilde{g}_j(\mathbf{x})$, $\tilde{c}_j = 1$
- False positives: $\tilde{o}_j = f(\mathbf{x}) + \tilde{g}_j(\mathbf{x})$, $\tilde{c}_j = 1$
- True/false negatives: $\tilde{o}_j = T$, $\tilde{c}_j = 0$

To add censorship, individuals were chosen at random to be in $\mathcal{C}$ and $\mathcal{P}$ and the values of $\tilde{\mathbf{o}}$ such that $\tilde{o}_j < \ell$ are set to ℓ and the corresponding $\tilde{c}_j$ is changed if needed. For individuals in $\mathcal{C}$, $\ell \sim U(1, \min_j \tilde{e}_j - 1)$. For individuals in $\mathcal{P}$, $\ell \sim U(\min_j \tilde{e}_j, \max_j \tilde{e}_j - 1)$. $U(a, b)$ is a uniform distribution.

Real Dataset. Within the healthcare domain, we leveraged MIMIC-III, a publicly available dataset of electronic health record data [25] and considered the time of sepsis onset as our prediction task. Experiments using this dataset are not regulated as human subjects research since the data are de-identified and publicly available. MIMIC-III includes data on vital signs, medications, diagnostic and procedure codes, and laboratory measurements for admissions to the intensive care unit. We chose to predict the time to sepsis onset as our task since multiple definitions of sepsis have been proposed and, to date, there remains a lack of consensus [2].

Like past work [26], we considered features relating to patient demographics and vital signs, data collected during the first seven hours of the admission, and admissions taking place between 2008-2012. We also considered a time horizon of 24 hours starting from the seventh hour of admission. After preprocessing (more detail in **Appendix C**), our cohort consisted of 19,866 admissions, and each admission had 345 features. To approximate ground truth sepsis onset, we used a composite definition based on 1) the clinical surveillance definition created by the Center for Disease Control and Prevention (CDC) and 2) the Centers for Medicare and Medicaid Services (CMS) definition [7], [27]–[29]. The composite time to sepsis onset was defined like past work [30], using the CMS time if available, the CDC definition if the CMS time was not available, or no time was recorded if neither were available. As our noisy labels, we considered the Sepsis1 definition [31], [32], Sepsis3 definition [8], [26], and a composite definition made based on Sepsis3 and Sepsis1 where the Sepsis3 time was used if available and the Sepsis1 time was used if the Sepsis3 time was unavailable. All sepsis definitions were implemented as defined by Sean et al. [33]. Individual components, such as SIRS and SOFA, were identified based on past work[1][2]. Our inclusion/exclusion criteria were the same as Moor et al. [26], and also excluded admissions with TTEs within the first seven hours.

TABLE I: MIMIC-III Noisy Sepsis Definitions. Most negative patients were correctly identified, so we show the true positive (TP) and false negative (FN) rates. We also show the noise means among the positive patients as the median and IQR (interquartile range). Abbreviations) Sep1: Sepsis1, Sep3: Sepsis3, Sep3/1: Sepsis3/1 composite

Definition	TP rate	FN Rate	Noise Mean (hours)
Sep1	72 (13.11%)	477 (86.89%)	17 [9-20]
Sep3	98 (17.85%)	451 (82.15%)	17 [6-20]
Sep3/1	106 (19.31%)	443 (80.69%)	17 [4-20]

Within the cohort, 486 (2.45%) admissions met the CMS definition, 122 (0.61%) met the CDC definition, 549 (2.76%) met the CDC/CMS composite definition, 134 (0.67%) met the Sepsis1 definition, 170 (0.86%) met the Sepsis3 definition, and 213 (1.07%) met the Sepsis3/1 composite definition. Since predictions were made over the course of a single hospital admission, no patients were lost to follow-up (i.e., observed for < 31 hours). More information summarizing the noisy labels is in **Table I**. For patients with noisy TTEs, 62.44% were administratively censored for at least labeler.

B. Baseline Approaches

We adapted approaches from classification with multiple noisy labels as baselines for comparison. Generally, these approaches aggregate the noisy labels at 1) preprocessing time [9], [10], or 2) inference time [11]–[14]. The first two aggregate at preprocessing. The last aggregates at inference.

Naïve Average. Noisy TTEs are averaged into a single label during preprocessing and are used to train a model with $\mathcal{L}_O$ [9]. That is, for individuals in $\mathcal{U}$ and $\mathcal{P}$, we aggregated $\tilde{\mathbf{c}}$ into $\mathbb{I}(\exists j\ \tilde{c}_j == 1)$, and $\tilde{\mathbf{o}}$ into the average of all uncensored TTEs (i.e., $\frac{1}{|\{j|\tilde{c}_j==1\}|}\sum_{j=1}^{m}\mathbb{I}(\tilde{c}_j == 1)\tilde{o}_j$). For individuals where we observe all of the labels, we expect to recover the expected ground truth TTE if the ground truth TTE is the average of the noisy TTEs. Otherwise, our estimate of the ground truth is biased since we aggregate over an incomplete label set.

[1]https://github.com/MIT-LCP/mimic-code/tree/main
[2]https://github.com/BorgwardtLab/mgp-tcn

Voting Average. Noisy TTEs are aggregated via an averaging strategy that builds on *Naïve Average* [9], [10]. Unlike the *Naïve Average*, we only took the average over the noisy TTEs if there was a noisy TTE for at least half of the labelers. Otherwise, we considered the individual as fully censored and last seen at ℓ. That is, if $\frac{1}{m}|\{j|\tilde{c}_j == 1\}| \geq 0.5$, then we trained with the average of the noisy TTEs, otherwise we ignored any available noisy TTEs and train with $c = 0, o = \ell$ after aggregation. While this produces fewer incorrect estimates than *Naïve Average*, it ignores some noisy TTEs.

Independent. We trained a multitask model to focus on each labeler separately [12], [14] and then aggregated labeler-specific predictions at inference time via averaging. That is, we learned to predict $P(\tilde{e}_j = t)$ for all labelers $j = 1, 2, ..., m$ and time points $t = 1, 2, ..., T$ and then aggregated with $P(e = t) = \frac{1}{m}\sum_{j=1}^{m} P(\tilde{e}_j = t)$. Given our multitask structure, we trained each head to focus on a single labeler with $\mathcal{L}_O$. This baseline addresses the limitations around including censored labels, but it assumes a known aggregation function and that the noisy and ground truth TTE distributions match.

C. Evaluation

We evaluated all approaches with three metrics [17]: 1) For TTE prediction accuracy, we measured the signed difference between the predicted TTE and ground truth TTE (i.e., $\hat{e} - e$). The predicted TTE was taken as the median of the model's prediction. In our results, we calculated the median difference between the predicted and ground truth TTEs and then reported the median over 10 replications changing the data split with error bars representing the interquartile range (IQR). On the synthetic data, we also show the distribution of values across all runs. 2) We measured discriminative performance with the time-dependent C-index [34]. For the remainder of this paper, we refer to the time-dependent C-index as the C-index for brevity. 3) We measured calibration error on the synthetic dataset with the integrated mean squared error (IMSE) between the predicted and ground truth survival curves. For the MIMIC-III dataset, we measured calibration error using the distributional divergence for calibration (DDC) [21] since we did not have access to the ground truth survival curves. For discriminative performance and calibration error, we reported the median and IQR of values over 10 replications and show the distribution of values for the synthetic data. For TTE accuracy, values closer to 0 are better. For discriminative performance, 0.5 shows random performance, and higher is better. For calibration error, lower is better.

IV. Results and Discussion

We examined the approaches through sensitivity analyses and ablation studies on synthetic data where we aimed to test the following hypotheses

- Our approach can learn in the presence of partial censorship more effectively than the baselines
- Our approach is effective when the assumed aggregation function is mis-specified through Steps 1b and 2
- Our approach is effect in the presence of label noise with respect to whether the event occurs

In our experiments, we measured prediction accuracy with the three metrics described above in settings where we varied 1) the amount of partial censorship in the dataset, 2) the amount by which the assumed and ground truth aggregation functions differed, and 3) the noise rate with respect to whether the event occurred. Then, we evaluated performance of the approaches on sepsis onset prediction in MIMIC-III.

A. Experiments on Synthetic Data

For robustness to censorship, we compared the proposed approach to the baselines, since we expect that some of the limitations of the baselines stem from a lack of robustness to partial censorship. For robustness to deviations from the assumed aggregation function, we compared to ablations of our proposed approach since we expect that Steps 1b and 2 make our approach effective in this setting. For robustness with respect to occurrence level noise, we compared to an ablation since we expect that the second term of $\mathcal{L}_G$ makes our approach effective in this setting. For all experiments on the synthetic data, ground truth labels in the test set were uncensored for evaluation.

Robustness to Censorship. We varied the rate of partial censorship in the synthetic dataset during training from 0% to 80% while fixing the rate of full censorship during training at 10%. The ground truth TTE was the average of the noisy TTEs so we did not use the error prediction component of the proposed approach (i.e., we trained θ and ψ with Steps 1a and 2). *Proposed* was robust across all three metrics, consistently showing the best performance as the rate of partial censorship increased (**Figure 3**). Both of the averaging baselines performed well in low censorship settings (i.e., partial censorship rate = 0% at training) but degraded as the amount increased. *Naïve Average* underestimated the TTE at a partial censorship rate of 80% with a TTE error -13 [IQR -19.6,-11.5]. In addition, at high rates of partial censorship, the distribution of results across replications became skewed. This is because, during training, there was sometimes a trade-off between TTE prediction accuracy and discriminative performance. As a result, for a few splits at high rates of partial censorship, the approach had high discriminative performance at the cost of an extremely biased TTE estimate. *Voting Average*, tended to overestimate the TTE as the rate of partial censorship increased. For example, at a partial censorship rate of 0.8, the TTE error was 3.5 [IQR -0.75,14]. We hypothesize that this was because it ignored the noisy TTEs in individuals where more than half of the labels were censored. We also noticed that the results had more variability than *Naïve Average*, and we hypothesize that this is due to *Voting Average* having a more complicated aggregation scheme.

Independent had TTE prediction errors that were more consistent and did not degrade to the extent of the averaging baselines. However, the discriminative performance and calibration error of this approach was noticeably worse than *Proposed*, even at low rates of censorship. This is in line

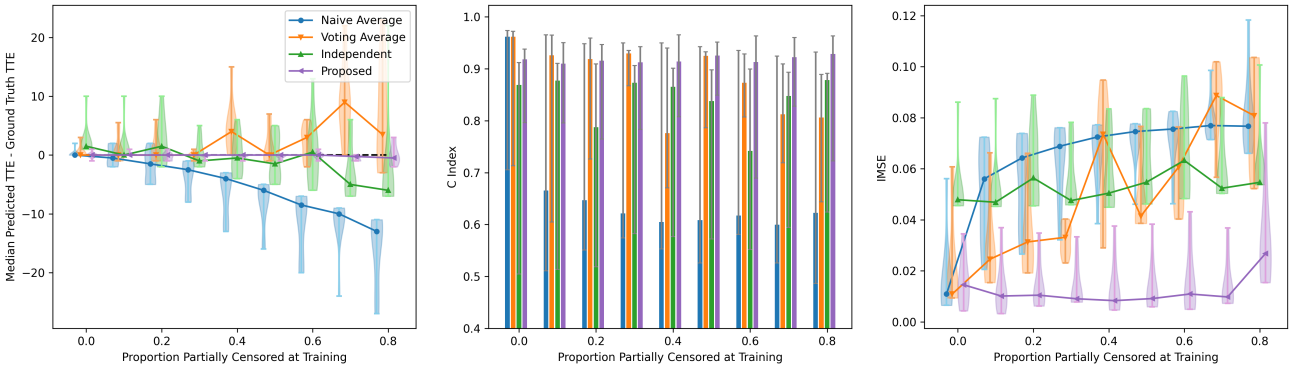

Fig. 3: We varied the rate of partial censorship while keeping the rate of full censorship at 10% during training. The proposed approach degraded the least as the rate of censorship increased. At each point along the x-axis, we plot the median, and distribution of values for each approach across all replications. We show the C-index results as a bar plot for clarity with error bars representing the range in values across runs. The outer plots have jitter along the x-axis for clarity.

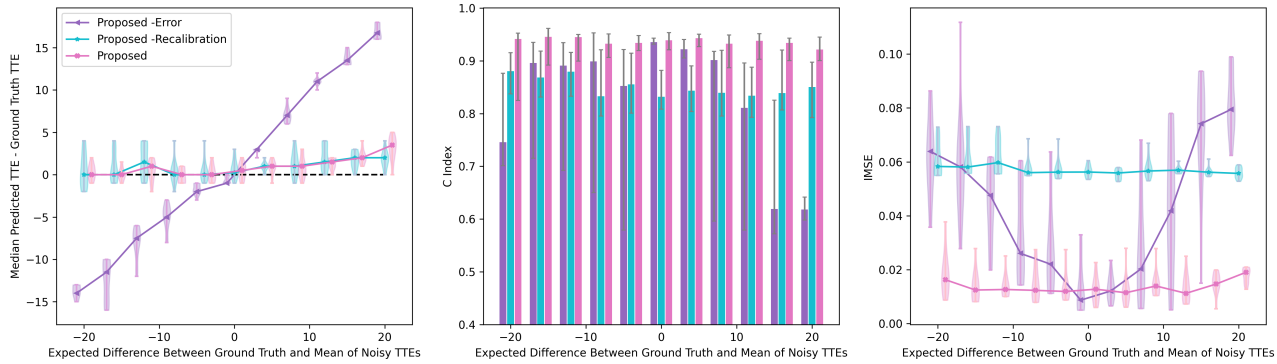

Fig. 4: We varied the TTE noise mean. Our approach maintained good performance across all metrics, while its ablations did not. At each point along the x-axis, we plot the median, and distribution of values for each approach across all replications.

with expectations, since our synthetic dataset was constructed such that the ground truth and noisy TTE distributions did not match, and *Independent* did not have a re-calibration step like our approach. At low rates of censorship, both averaging approaches had a slight advantage over *Proposed* with respect to discriminative performance. This is likely because they were more sample efficient than *Proposed* since the ground truth TTE could be accurately recovered by taking the average of the noisy TTEs during preprocessing. When we increased the training set size 20 fold, the C-Index of *Proposed* improved to 0.94 [IQR 0.93-0.95] from 0.92 [IQR 0.87-0.93].

Robustness to Differences from the Aggregation Function. Here, we relaxed the setting where the ground truth and assumed aggregation function matched on the synthetic dataset and compared to ablations of the proposed approach while varying the expected difference between the averaged noisy TTEs and ground truth TTE (i.e., the noise mean). We fixed the rates of full and partial censorship at training to 0.1 and 0.4, respectively. Here, *Proposed -Recalibration* implemented Steps 1a and 1b of the approach and obtained predictions of the ground truth TTE distribution through convolution of the predictions from these steps (**Appendix E**), while *Proposed -Error* implemented Steps 1a and 2 and did not learn the error pattern among labelers.

Proposed and *Proposed -Recalibration* were able to accurately predict the TTE across a variety of noise means, while *Proposed -Error* was not (**Figure 4**). While TTE error degraded for *Proposed -Error*, we note that it was more stable across replications than the other metrics (i.e., C-index and IMSE) since it depended less on the variability of how accurate the predictions were in the test set within a replication. At 0-mean noise, *Proposed -Error* achieved similar TTE error and discriminative performance to *Proposed*, but with slightly better calibration. This is likely because, at 0-mean noise, imperfections from the error prediction component may lead to a propagation of error when learning the ground truth survival distribution. When it is known in advance that the noise is 0-mean, it would be more efficient to use *Proposed -Error*. However, this experiment demonstrates that the error prediction component makes the proposed approach applicable to more

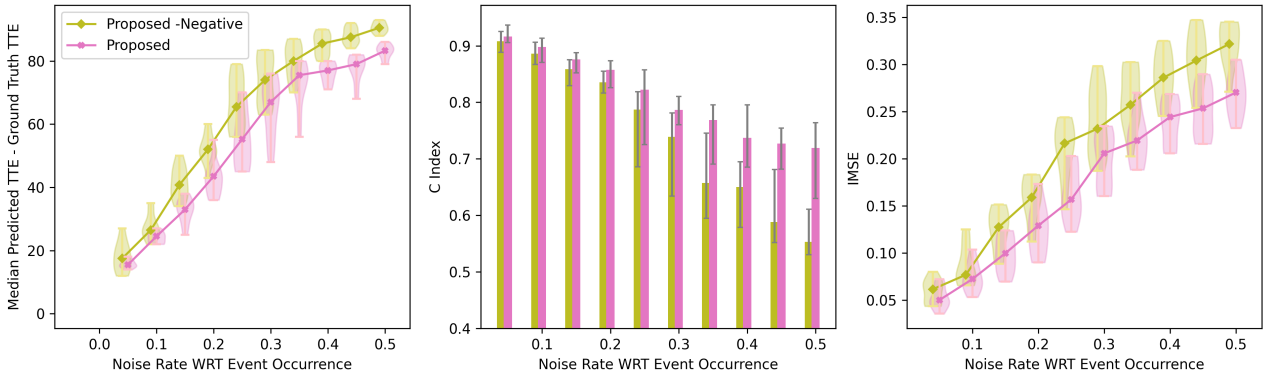

Fig. 5: We varied the noise rate at training with respect to event occurrence. The proposed approach with explicit supervision over negative individuals degraded the least as the noise rate increased. We show the distribution of results over 10 replications.

settings. *Proposed -Recalibration* suffered with respect to both discriminative performance and calibration error, with C-index values between 0.83-0.88 and IMSE values between 0.055-0.060, while those of *Proposed* were 0.92-0.95 and 0.011-0.019, respectively. This highlights the limitation of *Proposed -Recalibration* of assuming that the noisy and ground truth TTE distributions match.

Robustness to Occurrence Level Noise In this experiment, we focused on the setting where some individuals did not experience the event by T time steps, and there was additional noise on whether event occurrence was correctly labeled. We fixed the rates of full and partial censorship during training at 0.1 and 0.4, respectively, used 0-mean noise, and fixed the proportion of negative individuals at 50%. In **Figure 5**, we varied the noise rate with respect to event occurrence (e.g., a noise rate of 20% corresponded to a 20% FP and FN rate across all labelers). We performed an ablation study on the proposed approach, where *Proposed -Negative* is the proposed approach without the second term of $\mathcal{L}_G$ that provided explicit supervision over negative individuals in the reference set. As shown in the results, we observed modest improvements in both the TTE and calibration error, with greater improvement in discriminative performance (compare 0.72 [IQR 0.69-0.73] to 0.55 [IQR 0.53-0.59] at a noise rate of 0.5). Overall, this experiment shows that providing direct supervision on negative individuals from the reference set in $\mathcal{L}_G$ is beneficial.

B. Performance on Sepsis Prediction

Finally, we examined the proposed approach and baselines on the clinical task of predicting sepsis onset in the intensive care unit, where 62.44% of patients with noisy TTEs were administratively censored for at least one of them and the 'ground truth' TTE was not the average of the noisy TTEs. This dataset contained many false negative patients, so noisy TTEs were biased to indicating that the event did not occur. The results are shown in **Table II**. The proposed approach achieved the best performance relative to the baselines with respect to all metrics. However, it is important to note that, although the proposed approach outperformed the baselines,

TABLE II: Sepsis prediction in MIMIC-III. The proposed approach achieves the best performance across all metrics. Here, TTE error corresponds to the 'Median Predicted TTE - Ground Truth TTE'. Entries are of the form: median [IQR].

Approach	TTE Error (Hours)	C-Index (↑)	DDC (↓)
Naïve Average	17.50 [16.25-18.00]	0.48 [0.45-0.52]	1.14 [1.08-1.29]
Voting Average	17.50 [16.25-18.00]	0.45 [0.42-0.48]	1.13 [1.08-1.23]
Independent	18.50 [17.25-19.00]	0.48 [0.46-0.50]	0.91 [0.84-1.12]
Proposed	**14.50** [13.25-15.75]	**0.59** [0.54-0.62]	**0.86** [0.82-1.23]

there is still significant room for improvement. For example, the proposed approach predicted sepsis onset 14.5 hours too late, which would not be sufficient in a hospital setting. This is because sepsis onset is difficult to predict, especially with the high false negative rates of our noisy proxies and the limited set of features that we used. We chose our features to align with past work [26] and excluded those that might lead to label leakage [30], [35]. In addition, the composite definition we used as ground truth based on the CDC and CMS definitions is itself a noisy proxy, since a consensus on how to label ground truth sepsis onset has not yet been reached. Despite these limitations, the improvement of our approach over the baselines is promising and serves as a starting point. We include additional results on varying the number of reference patients in the training set in **Appendix F**. Overall, we found that the proposed approach improved over the baselines when there were at least 450 reference patients in the training set.

V. Related Work

We summarize work from classification, survival analysis, and semi-supervised learning where we discuss 1) approaches for multiple noisy labels, 2) approaches for noisy labels from a single labeler and interval censoring, and 3) pseudo-labels.

Classification. Past work addressing multiple noisy labels tends to focus on the setting of multiple annotators from crowdsourcing [36] and mainly builds on the Dawid and Skene approach [11], which models each labeler individually. For example, follow up work has proposed to learn a mapping from the ground truth label to the noisy label for each labeler

[37], [38] or to learn a model for each labeler separately based on the noisy labels [39]. More recently, instead of learning a separate model for each labeler, Jiang et al. proposed to learn the distribution of observed labels using a single model [40]. This approach is most similar to Step 1a of our approach.

Alternative to modeling labelers separately, other works propose to aggregate the labels at preprocessing time. In classification, this often means taking the majority vote among labelers. Past work has built on this by learning how to weight labelers, since some may be more reliable than others [9], [10]. Due to the presence of censorship in survival data, our approach did not aggregate at preprocessing time.

Survival Analysis. Past work studying label noise in survival analysis focuses on the single labeler case. Recent work in this space has shifted to a scenario where, for some individuals, reference TTEs can be obtained via chart review with an expert. Note that this concept has been used in other parts of the noisy label literature that aim to learn with noisy data in classification [41], and such a subset of data is often referred to as an anchor set. This subset is then used to help learn the relationship between the noisy and ground truth TTE [42]. The relationship between the ground truth and noisy TTEs can be learned in several ways, such as learning the sensitivities and specificities of the noisy TTEs for various time points [43] and learning a mapping between the noisy and ground truth TTEs [15]. The latter is the most similar to our approach. However, the multiple noisy labels setting brings novel challenges in that individuals can be partially censored.

Finally, another field that tackles a similar problem to ours is interval censoring [44], [45]. In this setting, it is known that the event occurs after some time point, but before another. These time points can be considered as two noisy TTEs. We differ in that we do not assume that at least one noisy TTE occurs before and after the ground truth TTE.

Semi-Supervised Learning. While this paper focuses on supervised learning, the idea of pseudo-labels has been explored in the context of semi-supervised learning [46], [47], where one aims to learn from a dataset where some instances are labeled and some are not. For some semi-supervised approaches, a classifier is learned from the labeled instances and then pseudo-labels from that classifier are applied to the unlabeled instances to fine tune it.

VI. Conclusion

We tackled the problem of multiple noisy labels in survival analysis. We highlight the limitations of existing approaches (adapted from the classification setting with multiple noisy labels) and propose a novel approach tailored to survival analysis. Leveraging a small reference set of expertly-labeled examples, our approach predicts the distribution of noisy TTEs and their errors and then maps that to a prediction of the ground truth TTE. Going forward our work could be extended to consider a setting in which you have multiple noisy non-overlapping labelers - that is not all labelers label all examples. In addition, one could also consider cases where the reference set is biased (i.e., not representative of the target population) or absent altogether. In assuming that the reference set was selected at random, we implicitly assumed that it was unbiased. Finally, future work could consider cases of dependent censoring [48]. Overall, our work brings together concepts from the noisy labels literature in classification and survival analysis to address an important problem.

Acknowledgments

This work was supported by Cisco Research and the National Science Foundation (NSF award no. IIS 2124127). The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Cisco Research or the National Science Foundation.

References

[1] M. Shankar-Hari, G. S. Phillips, M. L. Levy, C. W. Seymour, V. X. Liu, C. S. Deutschman, D. C. Angus, G. D. Rubenfeld, M. Singer, *et al.*, "Developing a new definition and assessing new clinical criteria for septic shock: for the third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 775–787, 2016.

[2] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[3] H. A. Lindner, S. Schamoni, T. Kirschning, C. Worm, B. Hahn, F.-S. Centner, J. J. Schoettler, M. Hagmann, J. Krebs, D. Mangold, *et al.*, "Ground truth labels challenge the validity of sepsis consensus definitions in critical illness," *Journal of Translational Medicine*, vol. 20, no. 1, p. 27, 2022.

[4] D. Tjandra, R. Q. Migrino, B. Giordani, and J. Wiens, "Cohort discovery and risk stratification for alzheimer's disease: an electronic health record-based approach," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 6, no. 1, p. e12035, 2020.

[5] R. Pace, T. Peters, E. Rahme, and K. Dasgupta, "Validity of health administrative database definitions for hypertension: a systematic review," *Canadian Journal of Cardiology*, vol. 33, no. 8, pp. 1052–1059, 2017.

[6] J. E. Yoo, D. W. Shin, K. Han, D. Kim, S.-P. Lee, S.-M. Jeong, J. Lee, and S. Kim, "Blood pressure variability and the risk of dementia: a nationwide cohort study," *Hypertension*, vol. 75, no. 4, pp. 982–990, 2020.

[7] C. Rhee, R. B. Dantes, L. Epstein, and M. Klompas, "Using objective clinical data to track progress on preventing and treating sepsis: Cdc's new 'adult sepsis event'surveillance strategy," *BMJ quality & safety*, 2018.

[8] C. W. Seymour, V. X. Liu, T. J. Iwashyna, F. M. Brunkhorst, T. D. Rea, A. Scherag, G. Rubenfeld, J. M. Kahn, M. Shankar-Hari, M. Singer, *et al.*, "Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 762–774, 2016.

[9] N. Parde and R. Nielsen, "Finding patterns in noisy crowds: Regression-based annotation aggregation for crowdsourced data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1907–1912, 2017.

[10] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," in *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, vol. 11, p. 269, NIH Public Access, 2017.

[11] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[12] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: whom to trust when everyone lies a bit," in *Proceedings of the 26th Annual international conference on machine learning*, pp. 889–896, 2009.

[13] J. Li, H. Sun, and J. Li, "Beyond confusion matrix: learning from multiple annotators with awareness of instance features," *Machine Learning*, vol. 112, no. 3, pp. 1053–1075, 2023.
[14] Z. Chu, J. Ma, and H. Wang, "Learning from crowds by modeling common confusions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 5832–5840, 2021.
[15] E. J. Oh, B. E. Shepherd, T. Lumley, and P. A. Shaw, "Raking and regression calibration: Methods to address bias from correlated covariate and time-to-event error," *Statistics in medicine*, vol. 40, no. 3, pp. 631–649, 2021.
[16] D. M. Vock, J. Wolfson, S. Bandyopadhyay, G. Adomavicius, P. E. Johnson, G. Vazquez-Benitez, and P. J. O'Connor, "Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting," *Journal of biomedical informatics*, vol. 61, pp. 119–131, 2016.
[17] S.-a. Qi, N. Kumar, M. Farrokh, W. Sun, L.-H. Kuan, R. Ranganath, R. Henao, and R. Greiner, "An effective meaningful way to evaluate survival models," in *ICML*, 2023.
[18] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–36, 2019.
[19] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[20] K. Ren, J. Qin, L. Zheng, Z. Yang, W. Zhang, L. Qiu, and Y. Yu, "Deep recurrent survival analysis," in *Proc. AAAI*, pp. 1–8, 2019.
[21] F. Kamran and J. Wiens, "Estimating calibrated individualized survival curves with deep learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 240–248, 2021.
[22] D. Tjandra, Y. He, and J. Wiens, "A hierarchical approach to multi-event survival analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 591–599, 2021.
[23] S. Hu, E. Fridgeirsson, G. van Wingen, and M. Welling, "Transformer-based deep survival analysis," in *Survival Prediction-Algorithms, Challenges and Applications*, pp. 132–148, PMLR, 2021.
[24] K. J. O'malley, K. F. Cook, M. D. Price, K. R. Wildes, J. F. Hurdle, and C. M. Ashton, "Measuring diagnoses: Icd code accuracy," *Health services research*, vol. 40, no. 5p2, pp. 1620–1639, 2005.
[25] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.
[26] M. Moor, M. Horn, B. Rieck, D. Roqueiro, and K. Borgwardt, "Early recognition of sepsis with gaussian process temporal convolutional networks and dynamic time warping," in *Proceedings of the 4th Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research, pp. 2–26, PMLR, 2019.
[27] A. Kalantari, H. Mallemat, and S. D. Weingart, "Sepsis definitions: the search for gold and what cms got wrong," *Western Journal of Emergency Medicine*, vol. 18, no. 5, p. 951, 2017.
[28] A. K. Venkatesh, T. Slesinger, J. Whittle, T. Osborn, E. Aaronson, C. Rothenberg, N. Tarrant, P. Goyal, D. M. Yealy, and J. D. Schuur, "Preliminary performance on the new cms sepsis-1 national quality measure: early insights from the emergency quality network (e-qual)," *Annals of emergency medicine*, vol. 71, no. 1, pp. 10–15, 2018.
[29] C. Rhee, Z. Zhang, S. S. Kadri, D. J. Murphy, G. S. Martin, E. Overton, C. W. Seymour, D. C. Angus, R. Dantes, L. Epstein, *et al.*, "Sepsis surveillance using adult sepsis events simplified esofa criteria versus sepsis-3 sofa criteria," *Critical care medicine*, vol. 47, no. 3, p. 307, 2019.
[30] F. Kamran, D. Tjandra, A. Heiler, J. Virzi, K. Singh, J. E. King, T. S. Valley, and J. Wiens, "Evaluation of sepsis prediction models before onset of treatment," *NEJM AI*, vol. 1, no. 3, 2024.
[31] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis," *Chest*, vol. 101, no. 6, pp. 1644–1655, 1992.
[32] A. E. W. Johnson, D. J. Stone, L. A. Celi, and T. J. Pollard, "The mimic code repository: enabling reproducibility in critical care research," *Journal of the American Medical Informatics Association*, vol. 25, no. 1, pp. 32–39, 2018.
[33] C. Y. Sean, K. D. Betthauser, A. Gupta, P. G. Lyons, A. M. Lai, M. H. Kollef, P. R. Payne, and A. P. Michelson, "Comparison of sepsis definitions as automated criteria," *Critical care medicine*, vol. 49, no. 4, pp. e433–e443, 2021.
[34] L. Antolini, P. Boracchi, and E. Biganzoli, "A time-dependent discrimination index for survival data," *Statistics in Medicine*, vol. 24, no. 24, pp. 3927–3944, 2005.
[35] A. Wong, E. Otles, J. P. Donnelly, A. Krumm, J. McCullough, O. DeTroyer-Cooley, J. Pestrue, M. Phillips, J. Konye, C. Penoza, *et al.*, "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients," *JAMA Internal Medicine*, vol. 181, no. 8, pp. 1065–1070, 2021.
[36] V. S. Sheng and J. Zhang, "Machine learning with crowdsourcing: A brief summary of the past research and future directions," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 9837–9843, 2019.
[37] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and M. Yang, "Deep learning for continuous multiple time series annotations," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pp. 91–98, 2018.
[38] J. Jacob, O. Ciccarelli, F. Barkhof, and D. C. Alexander, "Disentangling human error from the ground truth in segmentation of medical images," ACL, 2021.
[39] M. Guan, V. Gulshan, A. Dai, and G. Hinton, "Who said what: Modeling individual labelers improves classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.
[40] L. Jiang, H. Zhang, F. Tao, and C. Li, "Learning from crowds with multiple noisy label distribution propagation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6558–6568, 2021.
[41] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 447–461, 2015.
[42] A. S. Magaret, "Incorporating validation subsets into discrete proportional hazards models for mismeasured outcomes," *Statistics in Medicine*, vol. 27, no. 26, pp. 5456–5470, 2008.
[43] S. Hunsberger, P. S. Albert, and L. Dodd, "Analysis of progression-free survival data using a discrete time survival model that incorporates measurements with and without diagnostic error," *Clinical Trials*, vol. 7, no. 6, pp. 634–642, 2010.
[44] D. Rabinowitz, A. Tsiatis, and J. Aragon, "Regression with interval-censored data," *Biometrika*, vol. 82, no. 3, pp. 501–513, 1995.
[45] J. C. Lindsey and L. M. Ryan, "Methods for interval-censored data," *Statistics in medicine*, vol. 17, no. 2, pp. 219–238, 1998.
[46] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, p. 896, Atlanta, 2013.
[47] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, "Meta pseudo labels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11557–11568, 2021.
[48] W. Zhang, C. K. Ling, and X. Zhang, "Deep copula-based survival analysis for dependent censoring with identifiability guarantees," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 20613–20621, 2024.
[49] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M. W. Sjoding, and J. Wiens, "Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data," *Journal of the American Medical Informatics Association*, 2020.
[50] H. Kvamme, Ø. Borgan, and I. Scheel, "Time-to-event prediction with neural networks and cox regression," *Journal of machine learning research*, vol. 20, no. 129, pp. 1–30, 2019.

APPENDIX

A. Summary of Proposed Approach

We summarize the approach in **Algorithm 1**.

B. Extended Theory

Using Markov's inequality, we can express the quantity $P(|e - \hat{e}| \geq \varepsilon)$ for positive individuals with some error tolerance value ε as follows. We only consider positive individuals because TTEs are not defined for negative individuals

$$P(|e - \hat{e}| \geq \varepsilon) \text{ (here, we consider positive individuals)}$$

Algorithm 1 Training procedure for proposed approach.

Input: $\{\mathbf{x}^{(i)}, \tilde{\mathbf{c}}^{(i)}, \tilde{\mathbf{o}}^{(i)}\}_{i \notin \mathcal{A}}$, $\{\mathbf{x}^{(i)}, \tilde{\mathbf{c}}^{(i)}, \tilde{\mathbf{o}}^{(i)}, c^{(i)}, o^{(i)}\}_{i \in \mathcal{A}}$, θ, ϕ, ψ (dataset and initial model parameters)
Output: θ, ψ (trained model parameters)

Train

while ¬(step 1 stopping criteria) **do** ▷ Step 1
  $[P(\tilde{e} = t|\mathbf{x})]_{t=1}^{T}, [P(\tilde{e} \notin \{1, 2, ..., T\}|\mathbf{x})] = \theta(\mathbf{x})$ ▷ 1a
  Compute $\mathcal{L}_{\mathcal{O}}$ for whole training set ▷ 1a Loss
  $[P(\tilde{g}(\mathbf{x}) = t|\mathbf{x})]_{t=T}^{T} = \phi(\mathbf{x})$ ▷ 1b
  Compute $\mathcal{L}_{\mathcal{E}}$ for reference set $\mathcal{A}$ ▷ 1b Loss
  Update model parameters
  Compute step 1 stopping criteria
Freeze θ, ϕ
while ¬(step 2 stopping criteria) **do** ▷ Step 2
  $[P(e = t|\mathbf{x})]_{t=1}^{T}, [P(e \notin \{1, 2, ..., T\}|\mathbf{x})] = \psi(\theta(\mathbf{x}), \mathbf{x})$
  Compute $\hat{e}_{avg}$ ▷ Pseudo-label
  Compute $\mathcal{L}_{\mathcal{G}}$ for whole training set ▷ Step 2 Loss
  Update model parameters
  Compute step 2 stopping criteria
return θ, ψ ▷ Use θ and ψ at inference time: $\psi(\theta(\mathbf{x}), \mathbf{x})$

$$
\begin{aligned}
&\leq \frac{1}{\varepsilon}\mathbb{E}_P[|e - \hat{e}|] \text{ (Markov's inequality)} \\
&= \frac{1}{\varepsilon}\left((1-\alpha)\mathbb{E}_{TP}[|e - \hat{e}|] + \alpha\mathbb{E}_{FN}[|e - \hat{e}|]\right) \\
&\approx \frac{1}{\varepsilon}\alpha\mathbb{E}_{FN}[|e - \hat{e}|] \text{ (if we assume (*))} \\
&\leq \frac{1}{\varepsilon}\alpha T \text{ (since the absolute error is at most T)}
\end{aligned}
$$

where subscripts on $\mathbb{E}$ denote the set of individuals to consider (i.e., P: positive, FN: false negative, TP: true positive). (*)We assume that we can accurately learn the distribution of noisy TTEs and errors for true positive individuals, so the expected error for $e - \hat{e}$ will mainly depend on false negative individuals.

Similarly, with $P(|P(e \notin \{1, 2, ..., T\}) - \hat{P}(e \notin \{1, 2, ..., T\})| \geq \phi)$ for all individuals with some error tolerance value ϕ ($e \not\leq T$ is shorthand for $e \notin \{1, 2, ..., T\}$),

$$
\begin{aligned}
&P(|P(e \not\leq T) - \hat{P}(e \not\leq T)| \geq \phi) \\
&\leq \frac{1}{\phi}\mathbb{E}_P[|P(e \not\leq T) - \hat{P}(e \not\leq T)|] \text{ (Markov)} \\
&\approx \frac{1}{\phi}\Big(\rho\alpha\mathbb{E}_{FN}[|P(e \not\leq T) - \hat{P}(e \not\leq T)|] \\
&+ (1-\rho)\beta\mathbb{E}_{FP}[|P(e \not\leq T) - \hat{P}(e \not\leq T)|]\Big) \text{ (assume (**))} \\
&\leq \frac{1}{\phi}(\rho\alpha + (1-\rho)\beta) \text{ (since the absolute error} \leq 1)
\end{aligned}
$$

TN: true negative, FP: false positive, β: false positive rate, ρ: positive rate. (**)We assume that we can accurately learn from TP and TN individuals, so the expected error for $P(e \not\leq T) - \hat{P}(e \not\leq T)$ will mainly depend on FNs and FPs.

TABLE III: For each hyperparameter, the lower bound is shown on top, and the upper bound is shown on the bottom. For hyperparameters we did not tune, we show one row.

Hyperparameter	Synthetic	MIMIC-III
Layer Size	100	100
Learning Rate	0.001	0.0001
	0.01	0.01
L2 Constant	0.001	0.001
	0.1	0.1
Number Layers	2	3
Censor Weight	1	0.01
	1	0.1

C. MIMIC-III Preprocessing

Data from the MIMIC-III dataset were extracted from the following tables: PRESCRIPTIONS, CHARTEVENTS, LABEVENTS, PATIENTS, INPUTEVENTS_MV, INPUTEVENTS_CV, OUTPUTEVENTS, ADMISSIONS, MICROBIOLOGYEVENTS, ICUSTAYS, DIAGNOSES_ICD, and were processed using the FlexIble Data Driven pipeLinE (FIDDLE), [49], a publicly available pre-processing tool for electronic health record data. From these tables, we extracted features relating to demographics (ethnicity, race, age at admission, gender, martial status) and vital signs (temperature, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, SpO2 [oxygen saturation]). We then used the default settings of FIDDLE (theta_1 = 0.001, theta_2 = 0.001, and theta_freq = 1.0) to preprocess the features. As model input, we used the first seven slices of the output corresponding to the first seven hours of the ICU visit.

D. Other Implementation Details

All networks were trained on Intel(R) Xeon(R) CPUs, E7-4850 v3 @ 2.20GHz and Nvidia GeForce GTX 1080 GPUs. All layers were initialized with He initialization from a uniform distribution. We divide our training data into five batches during training. All random seeds were initialized with 123456789+x where x is the replication number. The runtime for the results in **Figures 3**, **4**, **5** was 48 hours for each plot, 30 minutes for **Table II**, and 2.5 hours for **Figure 6**.

We randomly split the data into 80/20% training/test sets, reporting results on the held-out test set. From the training set, we randomly selected 500 individuals from the synthetic dataset and 1,000 patients from MIMIC-III to use as the reference set, where half of the reference set remained in the training set and the other half was set aside as a validation set for model selection. For model selection, we used early stopping (patience=20) based on validation set performance. Validation set performance was measured as the difference between the predicted and reference TTEs. We also used the validation set for hyperparameter selection. For hyperparameters, we mainly focused on tuning the learning rate and L2 regularization constant using random search, with a budget of 5. The ranges we used and the values of the other hyperparameters are listed in **Table III**. For the MIMIC-III dataset, where there was class

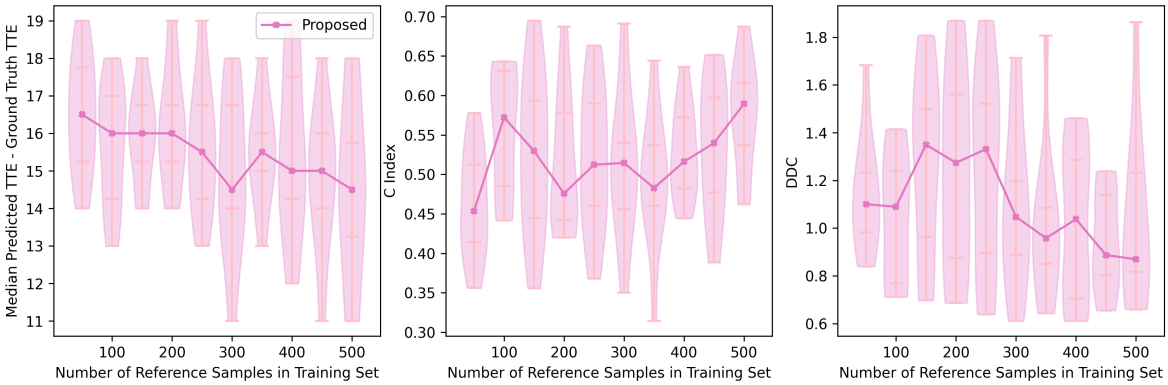

Fig. 6: We varied the number of reference individuals in the training set. We observed a general trend of improvement as the number of reference individuals increased. We show the full distribution of results over 10 replications and mark the IQRs.

imbalance with respect to occurrence of sepsis onset (i.e., most of them do not), we found that down-weighting the loss of the fully censored patients improved performance. Thus, the hyperparameter 'Censor Weight' represents a constant that was multiplied to $\mathcal{L}_{\mathcal{C}}$ during training.

For our model architectures, we trained feed forward neural networks whose architecture was similar to DeepHit [19], [22]. Both averaging approaches were implemented as a single feed forward that took as input individual features and outputted the probability of event occurrence at each time point. Independent was implemented as a multitask feed forward neural network, where each head took as input individual features and outputed the probability of event occurrence at each time point. The proposed approach was implemented as three feed forward neural networks whose structure mirrored **Figure 2**. θ was a feed forward network. ϕ was a feed forward network that took as input individual features and outputted the probability of observing each error value over all possible error values. ψ was a feed forward network, which took as input a concatenation of individual features and the predicted distribution of noisy TTEs and outputted the probability of the ground truth event occurrence at each time point in the horizon. We chose a feed forward implementation based on DeepHit and cross entropy loss since it requires fewer assumptions than other objectives (e.g., proportional hazards [50]). All models were trained in Python3.7 and Pytorch1.7.1, using Adam.

E. Convolution Operation for Ablation

Our convolution-like operation for *Proposed -Recalibration* below. Here, $\mathbb{I}$ is the indicator function.

$$\hat{P}(e=t) = \sum_{u=1}^{T} \sum_{v=-T}^{T} \mathbb{I}(u+v == t)\hat{P}(\tilde{e}=u)\hat{P}(\tilde{g}(\mathbf{x})=v)$$

In the two summations, for u and v combinations that exceeded T or were below 0, we clipped the values at T and 0 where appropriate so that the output of the convolution was a proper probability distribution (i.e., it sums to 1).

F. Varying the Reference Set Size for Sepsis Prediction

We performed a sensitivity analysis on MIMIC-III, where we varied the number of reference individuals in the training set from 50 to 500 patients. Like **Appendix D**, individuals were chosen at random to be in the reference set, and we kept the number of reference individuals in the validation set constant at 500. We evaluated the proposed approach only since none of the other baselines explicitly used a reference set. In our results (**Figure 6**), we observed a general trend of improvement as the number of reference individuals in the training set increased. Due to the difficulty of the task and nature by which we defined the labels, the trends were not strong. However, the results potentially suggest that, even with 250 reference individuals in the training set, the proposed approach can improve over the baselines with respect to two out of three metrics (i.e., TTE accuracy and discriminative performance) when comparing to **Table II**. At 250 individuals, the proposed approach achieved a TTE accuracy of 15.5 hours [IQR 14.25-16.75] compared to the next best baseline with 500 reference individuals in the training set (17.5 hours [IQR 16.25-18.00]) and a C-index of 0.51 [IQR 0.46-0.59] compared to the next best baseline with 500 reference individuals in the training set (0.48 [IQR 0.45-0.52]). The proposed approach did not improve over the baselines with respect to calibration error until there were 450 reference individuals in the training set, and this may be because the ψ, the recalibration component of the approach, is trained based on the output of θ and ϕ. Since ϕ is trained using only reference individuals, having fewer reference individuals likely hurts the accuracy of ϕ, leading to a propagation of error when training ψ. At 450 reference individuals in the training set, the proposed approach achieved a DDC of 0.89 [IQR 0.80-1.13] compared to the next best baseline with 500 reference individuals in the training set (0.91 [IQR 0.84-1.12]). In general, the optimal size of the reference set is likely to be task dependent. In spite of this, observing an improvement over the baselines with respect to all metrics at a reference size of 450 in the training set shows promise due to the difficulty of the task.