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). In many domains, like healthcare, accurate TTEs can be difficult to obtain. For example, with some diseases, identifying TTEs can require manual chart review by a clinical expert (e.g., sepsis [1]–[3]), making it challenging to efficiently label large datasets. Thus, automated pragmatic labeling tools based on the structured components of the dataset (e.g., tables from an electronic health record) are often used instead [4], [5]. For example, one could label sepsis onset as 1) when the CDC (Center for Disease Control and Prevention) definition is met [6] or 2) when the Sepsis3 definition is met [7]. However, this could mislabel *who experiences the event* and *when the event occurs*. In the absence of ground truth TTEs for most patients, one can potentially learn an accurate survival model by combining noisy proxies from 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 [8], [9], or 2) inference time [10]–[13]. Survival analysis differs from standard classification in that some individuals may have censored outcomes (i.e., an individual is only known to be event free up until a certain point). Work addressing noisy labels in survival analysis is largely limited to the single labeler case [14]. In our setting with multiple noisy labels, one could naïvely take the average of the noisy TTEs during preprocessing 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. This addresses issues around aggregating censored labels but still requires assumptions on how to aggregate the noisy predictions (e.g., the ground truth TTE is the average of the noisy ones).

Our Idea. To address the limitations of past work and naïve solutions, 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 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.

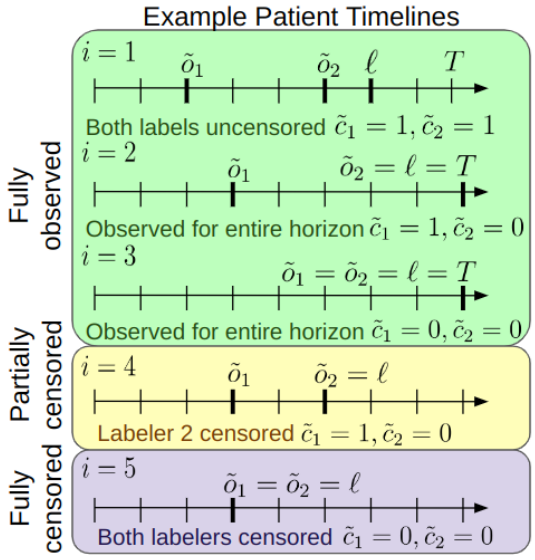

Fig. 1: Demonstration with example timelines. Instead of observing ground truth time-to-events (TTEs) in our dataset, we observe $m > 1$ noisy proxies. In this figure we have $m = 2$ labelers for five example individuals. Notation) ℓ: last time observed, $\tilde{c}_j$: indicator for whether an event 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.

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 [15]. 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., $\hat{e} = argmin_t \hat{S}(t|\mathbf{x}) \leq 0.5$) [16], 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 labelers (i.e., 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 if none are recorded). The second is **fully censored**: $\mathcal{C} = \{\mathcal{D}^{(i)} | (\tilde{c}_j^{(i)} == 0 \forall j) \wedge ([\max_j o_j^{(i)}] < T)\}$, 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 [17] 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 $P(\mathcal{D}^{(i)} \in \mathcal{P}) = P(\mathcal{D}^{(i)} \in \mathcal{P}|\mathbf{x}^{(i)})$) and that ℓ is uncorrelated with e.

B. Proposed Approach

Our approach (**Figure 2**) 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 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

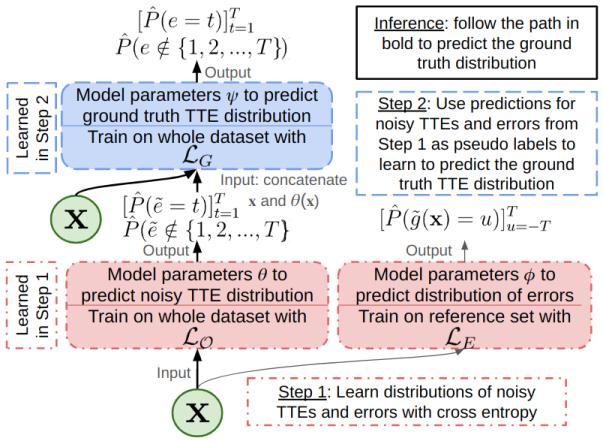

Fig. 2: Overview of proposed approach. Based on $\mathbf{x}$, we first predict the noisy TTEs and their errors as intermediate outputs. We then concatenate the noisy TTE prediction with $\mathbf{x}$ and map this to a prediction of the ground truth TTE as the final output of the approach. The approach 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.

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.

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 $\mathbf{x}$. Note that the $\tilde{e}_j$ values themselves are not necessarily from the same distribution. For example, given $\mathbf{x}$, $\tilde{e}_1$ 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 [18]–[21]. 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 cross-entropy loss to learn the ground truth survival distribution. This is represented in the first term of $\mathcal{L}_G$ below.

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

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 [22], 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 [23]. 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.

III. Experiments and Discussion

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. For our synthetic data, we used $T = 200$, $n = 5,000$, and $d = 100$. We obtained e by drawing from a normal distribution centered around a function of the features. We drew e from a normal distribution to contrast with how we generated the noisy TTEs. TTE noise was drawn from a skewed distribution based on another function of the features. Censorship status was assigned randomly. From our real data, MIMIC-III, we predicted sepsis onset since multiple definitions of sepsis exist, and there is currently no ground truth definition. We used demographic and vital sign features from the first seven hours of admission to the intensive care unit with a time horizon of 24 hours starting at $t = 7$. To approximate ground truth sepsis onset, we used a composite definition based on 1) the CDC definition and 2) the Centers for Medicare and Medicaid Services (CMS) definition [6], [24] like past work [25]. As our noisy labels, we considered Sepsis1 [26], Sepsis3 [7], and an analogous composite definition based on Sepsis3 and Sepsis1. Our baselines, *Naïve Average*, *Voting Average*, and *Independent*, adapted work from classification [8]–[13] to survival analysis. Both averaging baselines aggregated the noisy labels at preprocessing time, and *Independent* aggregated at inference time. We considered three metrics [16]: 1) For TTE prediction accuracy, we measured the signed difference between the predicted TTE and ground truth TTE (i.e., $\hat{e} - e$). 2) We measured discriminative performance with the time-dependent C-index [27]. 3) We measured calibration error on the synthetic data with the integrated mean squared error (IMSE) between the predicted and ground truth survival curves. For MIMIC-III, we used the distributional divergence for calibration (DDC) [20]. More detail about the implementation and experimental setup can be found with the code.

We conducted sensitivity analyses and ablation studies on synthetic data where we tested the following hypotheses:

- our approach can learn in the presence of partial censorship more effectively than the baselines and
- our approach is effective when the assumed aggregation function is misspecified through Steps 1b and 2.

To test these hypotheses, we varied 1) the amount of partial censorship in the dataset and 2) the amount by which the assumed and ground truth aggregation functions differed. 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 both experiments, ground truth labels were uncensored at test time.

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%. Here, 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. 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 censorship, these baselines had a high C-index at the cost of an extremely biased TTE estimate. *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 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 slightly higher C-index than *Proposed*, likely because they were more sample efficient.

Robustness to Misspecified 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, while *Proposed -Error*

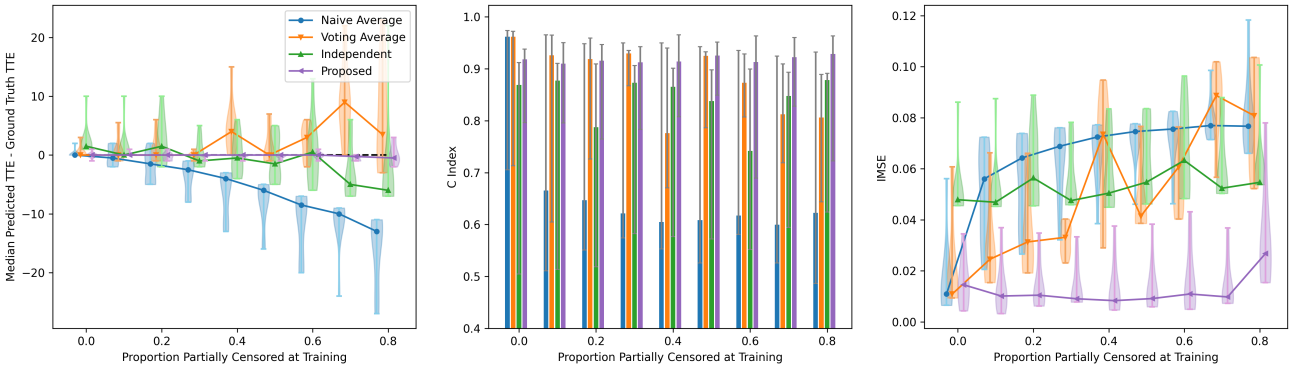

Fig. 3: We varied the rate of partial censorship while keeping the rate of full censorship at 10% during training. The proposed approach degraded the least as the rate of censorship increased. At each point along the x-axis, we plot the median and distribution of values for each approach across all replications.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

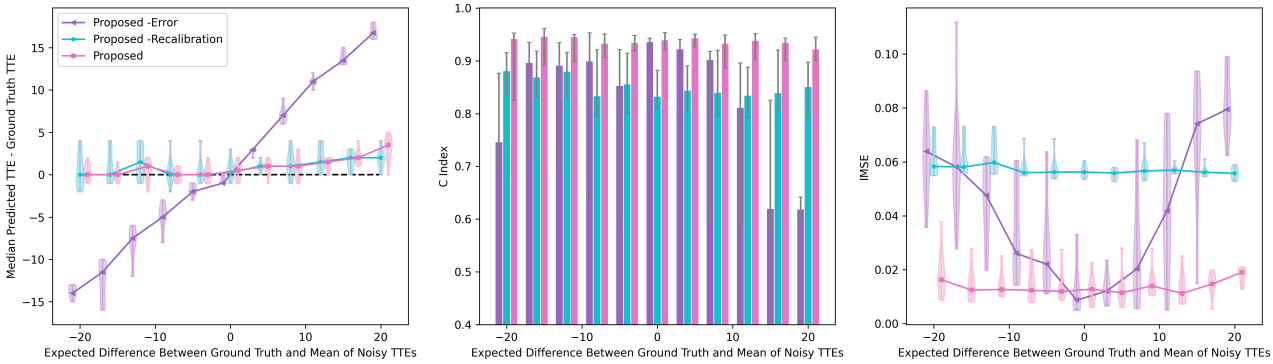

Fig. 4: We varied the TTE noise mean. Our approach maintained good performance across all metrics, while its ablations did not. At each point along the x-axis, we plot the median and distribution of values for each approach across all replications.

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**). *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.

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. Here, 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 I**. 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, 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. 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.

IV. 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

TABLE I: 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]

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 [28]. 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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.