

**Predicting the Timing of Clinical Outcomes in the Presence of  
Multiple Events and Noisy Labels**

by

Donna Tjandra

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
(Computer Science and Engineering)  
in the University of Michigan  
2024

Doctoral Committee:

Associate Professor Jenna Wiens, Chair  
Assistant Professor Benjamin Fish  
Professor Bruno Giordani  
Professor Emily Mower Provost

Donna Tjandra  
dotjandr@umich.edu  
ORCID iD: 0000-0001-7303-9123  
© Donna Tjandra 2024

## ACKNOWLEDGEMENTS

I would like to thank the following individuals (human or otherwise) for their guidance and support over the past six years.

- My advisor Jenna Wiens ☆
- My committee Ben Fish, Bruno Giordani, and Emily Mower Provost
- My collaborators Bruno Giordani, Ray Migrino, Chase Irwin, and the sepsis team
- My internship mentors Nina Mishra and Milind Shyani
- My labmates and other member of MLD3: Ian Fox, Jeeheh Oh, Haozhu Wang, Jiaxuan Wang, Yanqi Wang, Ben Li, Fahad Kamran, Shengpu Tang, Harry Rubin-Falcone, Meera Krishnamoorthy, Sarah Jabbour, Yifei He, Erkin Otles, Jung Min Lee, Trenton Chang, Michael Ito, Winston Chen, Greg Kondas, Stephanie Shepard, and Rebekah Clark. Special mentions to Donnell Tang, Maine Tang, Maine's six kittens (Alpha, Beta, Gamma, Delta, Epsilon, Zeta) Marvin Catzen-Rubin-Falcone, and Taylor Ito
- My family
- My undergraduate mentors Natasha Pascoe, Charlie Boone, and Ken Jackson
- My stuffed toys ☺ and favorite fictitious characters ♡

Graduate school was an eye-opening experience, thanks for the memories!



# TABLE OF CONTENTS

|                              |     |
|------------------------------|-----|
| ACKNOWLEDGEMENTS . . . . .   | ii  |
| LIST OF FIGURES . . . . .    | vi  |
| LIST OF TABLES . . . . .     | ix  |
| LIST OF APPENDICES . . . . . | x   |
| LIST OF ACRONYMS . . . . .   | xi  |
| ABSTRACT . . . . .           | xii |

## CHAPTER

|                                                               |           |
|---------------------------------------------------------------|-----------|
| <b>1 Introduction . . . . .</b>                               | <b>1</b>  |
| 1.1 Challenges and Opportunities . . . . .                    | 3         |
| 1.2 Practical Value . . . . .                                 | 5         |
| 1.3 Contributions . . . . .                                   | 6         |
| <b>2 Background . . . . .</b>                                 | <b>7</b>  |
| 2.1 Survival Analysis . . . . .                               | 7         |
| 2.1.1 Problem Setup and Notation . . . . .                    | 7         |
| 2.1.2 Predicting Survival Curves . . . . .                    | 9         |
| 2.1.3 Settings with Multiple Events . . . . .                 | 11        |
| 2.1.4 Noisy Labels in Survival Analysis . . . . .             | 12        |
| 2.1.5 Evaluation of Survival Models . . . . .                 | 12        |
| 2.2 Classification . . . . .                                  | 13        |
| 2.2.1 Problem Setup and Notation . . . . .                    | 13        |
| 2.2.2 Evaluation . . . . .                                    | 14        |
| 2.2.3 Ordinal Regression . . . . .                            | 15        |
| <b>3 Hierarchical Multi-Event Survival Analysis . . . . .</b> | <b>16</b> |
| 3.1 Introduction . . . . .                                    | 16        |
| 3.2 Related Work . . . . .                                    | 18        |
| 3.2.1 Problem Setup and Definitions . . . . .                 | 18        |
| 3.2.2 Multi-Event Survival Analysis . . . . .                 | 18        |
| 3.2.3 Evaluation with Multiple Events . . . . .               | 19        |

|          |                                                               |           |
|----------|---------------------------------------------------------------|-----------|
| 3.3      | Methods . . . . .                                             | 20        |
| 3.3.1    | Notation . . . . .                                            | 20        |
| 3.3.2    | Overall Architecture . . . . .                                | 21        |
| 3.3.3    | Training Loss . . . . .                                       | 25        |
| 3.3.4    | Censorship Due to (Semi-)Competing Events . . . . .           | 26        |
| 3.4      | Experimental Setup . . . . .                                  | 27        |
| 3.4.1    | Datasets . . . . .                                            | 27        |
| 3.4.2    | Baselines . . . . .                                           | 30        |
| 3.4.3    | Evaluation . . . . .                                          | 32        |
| 3.4.4    | Implementation Details . . . . .                              | 32        |
| 3.5      | Results and Discussion . . . . .                              | 33        |
| 3.5.1    | How Does the Proposed Approach Perform? . . . . .             | 33        |
| 3.5.2    | Do Hierarchical Predictions Help? . . . . .                   | 33        |
| 3.6      | Conclusion . . . . .                                          | 35        |
| <b>4</b> | <b>Learning with Instance-Dependent Label Noise . . . . .</b> | <b>37</b> |
| 4.1      | Introduction . . . . .                                        | 37        |
| 4.2      | Related Work . . . . .                                        | 39        |
| 4.2.1    | Identifying Mislabeled Data . . . . .                         | 39        |
| 4.2.2    | Noise-Robust Loss Functions . . . . .                         | 40        |
| 4.2.3    | Label Noise and Bias . . . . .                                | 40        |
| 4.2.4    | Anchor Points for Addressing Label Noise . . . . .            | 41        |
| 4.2.5    | How We Differ . . . . .                                       | 41        |
| 4.3      | Methods . . . . .                                             | 41        |
| 4.3.1    | Notation and Setting . . . . .                                | 42        |
| 4.3.2    | Justification and Desired Properties . . . . .                | 42        |
| 4.3.3    | Proposed Approach . . . . .                                   | 43        |
| 4.4      | Experimental Setup . . . . .                                  | 46        |
| 4.4.1    | Datasets . . . . .                                            | 47        |
| 4.4.2    | Adding Label Noise . . . . .                                  | 49        |
| 4.4.3    | Evaluation Metrics . . . . .                                  | 49        |
| 4.4.4    | Baselines . . . . .                                           | 50        |
| 4.4.5    | Implementation Details . . . . .                              | 50        |
| 4.5      | Results and Discussion . . . . .                              | 51        |
| 4.5.1    | Robustness to Noise Rate . . . . .                            | 51        |
| 4.5.2    | Sensitivity to Alignment Set Composition . . . . .            | 53        |
| 4.5.3    | Ablation Study and Hyperparameter Sensitivity . . . . .       | 55        |
| 4.6      | Conclusion . . . . .                                          | 56        |
| <b>5</b> | <b>Survival Analysis with Multiple Noisy Labels . . . . .</b> | <b>57</b> |
| 5.1      | Introduction . . . . .                                        | 57        |
| 5.2      | Related Work . . . . .                                        | 59        |
| 5.2.1    | Classification . . . . .                                      | 59        |
| 5.2.2    | Survival Analysis . . . . .                                   | 60        |
| 5.2.3    | Semi-Supervised Learning. . . . .                             | 61        |

|          |                                                  |           |
|----------|--------------------------------------------------|-----------|
| 5.3      | Methods . . . . .                                | 61        |
| 5.3.1    | Notation and Problem Setup . . . . .             | 61        |
| 5.3.2    | Proposed Approach . . . . .                      | 63        |
| 5.3.3    | Practical Guidelines on Expected Error . . . . . | 66        |
| 5.4      | Experimental Setup . . . . .                     | 67        |
| 5.4.1    | Datasets . . . . .                               | 68        |
| 5.4.2    | Baseline Approaches . . . . .                    | 70        |
| 5.4.3    | Evaluation . . . . .                             | 71        |
| 5.4.4    | Implementation Details . . . . .                 | 71        |
| 5.5      | Results and Discussion . . . . .                 | 73        |
| 5.5.1    | Experiments on Synthetic Data . . . . .          | 73        |
| 5.5.2    | Performance on Sepsis Prediction . . . . .       | 77        |
| 5.6      | Conclusion . . . . .                             | 78        |
| <b>6</b> | <b>Conclusion . . . . .</b>                      | <b>79</b> |
| 6.1      | Summary of Contributions . . . . .               | 79        |
| 6.2      | What Next? . . . . .                             | 81        |
| 6.3      | Final Conclusion . . . . .                       | 82        |
|          | APPENDICES . . . . .                             | 83        |
|          | BIBLIOGRAPHY . . . . .                           | 90        |

# LIST OF FIGURES

## FIGURE

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                    |    |
|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1 | An overview of this thesis. We build on existing work in survival analysis and address shortcomings. First, we leverage the relatedness among TTEs in the multi-event setting to more effectively learn the dependencies among events. Then, we consider label noise with respect to whether the event occurred, and then label noise with respect to when the event occurred in the context of multiple labelers.                 | 3  |
| 2.1 | Example survival curve. . . . .                                                                                                                                                                                                                                                                                                                                                                                                    | 8  |
| 3.1 | Proposed architecture. A multitask network that takes $\mathbf{x}$ as input and outputs $\hat{\mathcal{P}}$ , a set of probability distributions for each event. Event specific subnetworks, $\epsilon_k$ , estimate the probability distribution, $\hat{\mathbf{p}}_k$ , hierarchically. At granularity $m$ , $\hat{\mathbf{p}}_{km}$ is constructed from $\phi_{km}(\theta(\mathbf{x}))$ and $\hat{\mathbf{p}}_{km-1}$ . . . . . | 21 |
| 3.2 | Example prediction. We predict if the event will occur within the horizon, and then iteratively predict at finer time scales until reaching the original time scale.                                                                                                                                                                                                                                                               | 22 |
| 3.3 | Consistency example. Individuals $i = 1, 2$ are shown by the blue and red curves, with TTE $o^{(1)}$ and observed time $o^{(2)}$ , respectively. Left: curves are consistent at $o^{(1)}$ and $o^{(2)}$ . Right: curves are inconsistent at $o^{(2)}$ . . . . .                                                                                                                                                                    | 26 |
| 3.4 | Local consistency. The red and blue curves represent individuals $i = 1, 2$ , and the dashed and solid curves represent events $E = 1, 2$ , respectively. The curves are globally (top), but not locally (bottom) consistent. $i = 1$ : inconsistent at $o_2^{(1)}$ . $i = 2$ : inconsistent at $o_1^{(2)}$ and $o_2^{(2)}$ . We assume that both events occur in both individuals. . . . .                                        | 31 |
| 3.5 | Comparison to existing approaches. Our approach outperforms state-of-the-art approaches in different multi-event settings across different metrics. Error bars represent empirical 95% confidence intervals. In order, the datasets represent the general multi-event setting, semi-competing risks, semi-competing risks, and competing risks. . . . .                                                                            | 33 |
| 3.6 | Comparison of survival curves. For seven out of eight randomly selected instances from the test set, the survival curves from our approach are similar to or more closely match the ground truth compared to predicting the original task directly.                                                                                                                                                                                | 34 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |    |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 4.1 | Our Approach. a) The model predicts, $\hat{y}$ , at training and inference time using $\theta$ . At training time, it also predicts whether the observed label is correct using $\phi$ . $\theta$ and $\phi$ are pre-trained using $A$ and then fine tuned with the complete dataset. b) We pretrain the model using the alignment points, then train on the noisy data. $\mathcal{L}_\theta$ , $\mathcal{L}_\phi$ , and $\mathcal{L}'_\theta$ are the objectives for the alignment points, label confidence score ( $\hat{\beta}_\phi$ ), and noisy data, respectively. $\alpha_1$ , $\alpha_2$ , $\gamma$ are scalar hyperparameters. . . . . | 43 |
| 4.2 | Robustness to noise rate in an instance dependent setting. We plot the mean and standard deviation for 10 random seeds. As the noise rate increases, the proposed approach generally shows the least degradation up to a minority noise rate of 80%. . . . .                                                                                                                                                                                                                                                                                                                                                                                    | 52 |
| 4.3 | Robustness to noise disparity in an instance dependent setting. We plot the mean and standard deviation for 10 random seeds. As the noise disparity increases, the proposed approach generally shows the least degradation up to a minority noise rate of 80%. . . . .                                                                                                                                                                                                                                                                                                                                                                          | 53 |
| 4.4 | Robustness to varying alignment set size and bias. Mean HM and standard deviation for 10 random seeds. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 54 |
| 4.5 | Ablation study of proposed approach. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 55 |
| 4.6 | Sensitivity analysis of proposed approach to objective function hyperparameters. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 56 |
| 5.1 | Comparison of distributions. We show examples of mismatches between the underlying survival distribution and distribution of noisy TTEs, and how these can hurt discriminative performance (a) and calibration (b). We then show how we address this (c). . . . .                                                                                                                                                                                                                                                                                                                                                                               | 59 |
| 5.2 | 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) $\ell$ : 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 $\ell$ if $\tilde{c}_j == 0$ or noisy TTE $\tilde{e}_j$ otherwise. . . . .                                                                                                                                                                         | 62 |
| 5.3 | 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. . . . .                                                                                                                                                                                           | 64 |
| 5.4 | 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 partial 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. . . .                                                                                                                                  | 74 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |    |
|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 5.5 | 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. We show the C-index results as a bar plot for clarity with error bars representing the range in values across runs. The other plots have jitter along the x-axis for clarity. . . . .                                                                  | 75 |
| 5.6 | 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. 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 other plots have jitter along the x-axis for clarity. | 76 |
| B.1 | Robustness with respect to AUROC when varying alignment set size and bias. We plot the mean and standard deviation for 10 random seeds. . . . .                                                                                                                                                                                                                                                                                                                                                          | 86 |
| B.2 | Robustness with respect to AUEOC when varying alignment set size and bias. We plot the mean and standard deviation for 10 random seeds. . . . .                                                                                                                                                                                                                                                                                                                                                          | 87 |

## LIST OF TABLES

### TABLE

|     |                                                                                                                                                                                                                                                                                                                                                 |    |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1 | Dataset summaries. $T$ is in months for ADNI and SEER, and hours for MIMIC-III. ‘D’, ‘A’, ‘S’, and ‘DC’/‘DP’ mean death, ARF, shock, and death from larynx cancer/pulmonary disease, respectively. . . . .                                                                                                                                      | 27 |
| 3.2 | Ablation study. The highest values are bold. Error bars show empirical 95% CIs. Abbreviations are as follows: DA: Adapted DeepHit for multiple events, -HA: Proposed without hierarchical component, A: DA with hierarchical architecture, S: DA with hierarchical supervision, P: Proposed approach . . . . .                                  | 35 |
| 4.1 | Notation. A summary of notation used throughout. Superscripts in parentheses specify instances (e.g., $\mathbf{x}^{(i)}$ ). Subscripts specify indexes into a vector (e.g., $\mathbf{x}_i$ ) . . . . .                                                                                                                                          | 41 |
| 5.1 | 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 . . . . . | 70 |
| 5.2 | 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. . . . .                                                                                                                                                                             | 72 |
| 5.3 | 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]. . . . .                                                                                                            | 77 |

## LIST OF APPENDICES

|                                      |    |
|--------------------------------------|----|
| A Supplement for Chapter 3 . . . . . | 83 |
| B Supplement for Chapter 4 . . . . . | 85 |
| C Supplement for Chapter 5 . . . . . | 88 |

## LIST OF ACRONYMS

**ML** machine learning

**AD** Alzheimer’s disease

**EHR** electronic health record

**CI** confidence interval

**ICU** intensive care unit

**TTE** time-to-event

**C-index** concordance index

**AUROC** area under the receiver operating characteristic curve

**AUEOC** area under the equalized odds curve

**EO** equalized odds

**HM** harmonic mean

**IMSE** integrated mean squared error

**IMAE** integrated mean absolute error

**IQR** interquartile range

## ABSTRACT

In survival analysis, one aims to predict the probability of a new event occurring over time, given time-to-event (TTE) training data. An event can refer to any outcome of interest, like disease onset, where ‘survival’ refers to being ‘event free’. Unlike standard regression tasks, outcomes for individuals in the training set may be censored. That is, an individual can be lost to follow-up such that it is known that the event does not occur up until some time point, but it is unknown when or whether the event occurs thereafter. While standard regression approaches exclude such individuals, survival analysis approaches account for censorship. However, current approaches often make limiting assumptions during training, such as 1) the timing of different events is independent and 2) TTEs are labeled accurately. In this thesis, we tackle the limitations of these assumptions and propose novel approaches that specifically address 1) potential dependencies among events and 2) label noise.

We first tackle the challenge of accounting for potential dependencies among TTEs during training in the multi-event setting. The majority of past approaches for multi-event survival analysis learn to model each event separately, thus assuming that TTEs among events are independent. However, this is unlikely to hold in practice and to address this, we propose a novel hierarchical approach that models multi-event TTE prediction at different resolutions of time. We begin with coarser resolutions (e.g., weeks rather than days), where relationships among events are likely easier to learn given the reduction in heterogeneity across time in TTE labels at this resolution, and use predictions from these coarser resolutions to guide predictions at the original resolution. We empirically show that our approach is more effective than existing baselines across a variety of tasks.

We then tackle the challenge of learning from noisy labels with respect to whether the event occurs. Past approaches are limited in that they often do not explicitly account for differences in noise rates among subsets of the data (e.g., female patients are sometimes underdiagnosed for heart disease compared to male patients and thus have a higher noise rate). We propose an approach to address this by learning the underlying pattern of label noise using a small subset of data for which we know the ground truth and noisy labels. We assume that this subset is representative of the target population so that patterns learned from it generalize. We then use this pattern to re-weight the objective function during

training on the entire dataset and demonstrate that our approach outperforms several state-of-the-art baselines across many tasks.

Finally, we tackle label noise with respect to when the event occurs in the context of multiple labelers. For a given event, there are often multiple ways to label the TTE. Each has its own strengths and weaknesses, and we hypothesize that combining multiple labeling tools could improve performance. However, adapting approaches for multiple noisy labels from classification to survival analysis is limited since it is not always straightforward how to aggregate the noisy labels. We propose to learn in the presence of multiple noisy TTEs by first learning to predict the noisy labels and then mapping that prediction to a prediction of the ground truth. In doing so, we implicitly learn an aggregation function.

By addressing key challenges in survival analysis, we demonstrate that one can train more accurate models. TTE labels potentially contain dependencies among events, since many events, such as disease onset and death, are often related. In addition, our approaches can also enable researchers to use a wider variety of datasets. For example, in healthcare, many models are trained from EHR (electronic health record) data. EHR-derived datasets are likely to contain noisy labels due to the use of automated labeling tools. Overall, our approaches improve the applicability of survival analysis in realistic scenarios containing messy data.

# CHAPTER 1

## Introduction

Survival analysis is a setting within supervised learning where one aims to predict the probability of an event occurring over time given time-to-event (TTE) data. Due to the widespread use of survival analysis across different fields of study, events can refer to any outcome of interest, and ‘survival’ refers to not having experienced the event. For example, previous work has studied events like mortality [149, 6, 40], device failure [199], political transition [1], exoneration [48], and animal recapture [21] using survival analysis. Unlike standard regression, individuals in a survival dataset may have censored outcomes. For example, in estimating mortality post-hospital discharge, a patient may only be observed for the first 15 days before they drop out of the study. For this individual, we have partial information in that we know they did not experience mortality within the first 15 days, but not what happened after 15 days. While standard regression approaches might exclude this individual from the training set given an incomplete label, approaches for survival analysis learn TTEs (i.e., when the event occurred) while accounting for censorship (e.g., by minimizing the probability of event occurrence before the last point of observation in individuals with censored outcomes during training). Despite current progress in survival analysis, there remain several opportunities for improvement. We focus on two of them, as described below.

The first challenge we highlight is modeling multiple events. While we previously described survival analysis in the context of a single event, we may also be interested in examining multiple events at once [94, 40]. For example, clinicians are sometimes interested in multiple downstream outcomes when deciding whether to prescribe medication [90, 8]. Considering multiple events at once presents new opportunities compared to the single-event setting in that there are likely dependencies among the events that one could leverage during training. The second challenge we highlight is the setting in which TTE labels are not always correct (i.e., the noisy labels setting). This presents new opportunities for survival analysis in that there is an implicit need to learn to map between the noisy and ground truth labels. Generally, noisy labels arise when it is infeasible to obtain accurate labels for the entire dataset due to time constraints, and there exist faster ways of obtaining labels at the

cost of accuracy [126, 168]. For example, some datasets from healthcare are labeled using automated, but error-prone, labeling tools since it would be infeasible for a clinical expert to chart review the entire dataset. In this thesis, we explore open problems in survival analysis relating to 1) modeling dependencies among TTEs in the multi-event setting and 2) noisy labels with respect to a) whether the event occurs within the prediction horizon and b) when the event occurs and develop novel approaches to address them.

The multi-event and noisy labels settings have been studied by past work [176, 45, 160]. Many works in the multi-event setting propose to model each event separately, thereby treating them as independent [6, 40]. Others build on this and learn to account for dependencies using deep learning with a multitask setup [96], where a shared layer is included before branching off into event-specific models. However, this does not explicitly take advantage of the observation that at coarser time resolutions/granularities than the one used to collect the data, temporal relationships/dependencies among TTEs for different events may be more predictable due to the lower variance in TTEs at the coarser grain. Past work addressing label noise mainly focuses on classification, which corresponds to label noise with respect to whether the event occurred in survival analysis. Earlier works assume that the noise is independent of individual/instance features [160]. However, this is not always true [179], and when the noise does depend on instance features, this can lead to different noise rates within subsets of the data, which in turn, can lead to biased model performance in favor of subsets with less noise. Although instance-dependent label noise has recently received more attention [29, 187, 173], the effect of these approaches on model bias is relatively understudied [105]. Outside of classification, past work studying label noise in survival analysis [109, 122] considers cases when the noisy TTE in the dataset does not match the TTE that occurred in reality. Previous work considers TTE label noise in the context of a single labeler, and label noise in survival analysis with multiple labelers remains understudied [127]. However, this setting is still relevant in domains like healthcare since there are often multiple ways to develop automated labeling tools for a specific condition, each with its own strengths and weaknesses, that can potentially be combined.

To address these gaps, we propose the following techniques. The first focuses on accounting for dependencies among events in the multi-event setting. To handle this, we propose a hierarchical approach that learns to model TTEs at several time granularities since it is likely that relationships among events are more predictable, and thus easier to learn, at coarser granularities due to their lower variance. The second and third focus on label noise. We begin with label noise with respect to whether the event occurs and consider the case where differences in noise rates in subsets of the data can bias model performance. We propose to learn the underlying pattern of mislabeling and use that pattern to reweight the objective

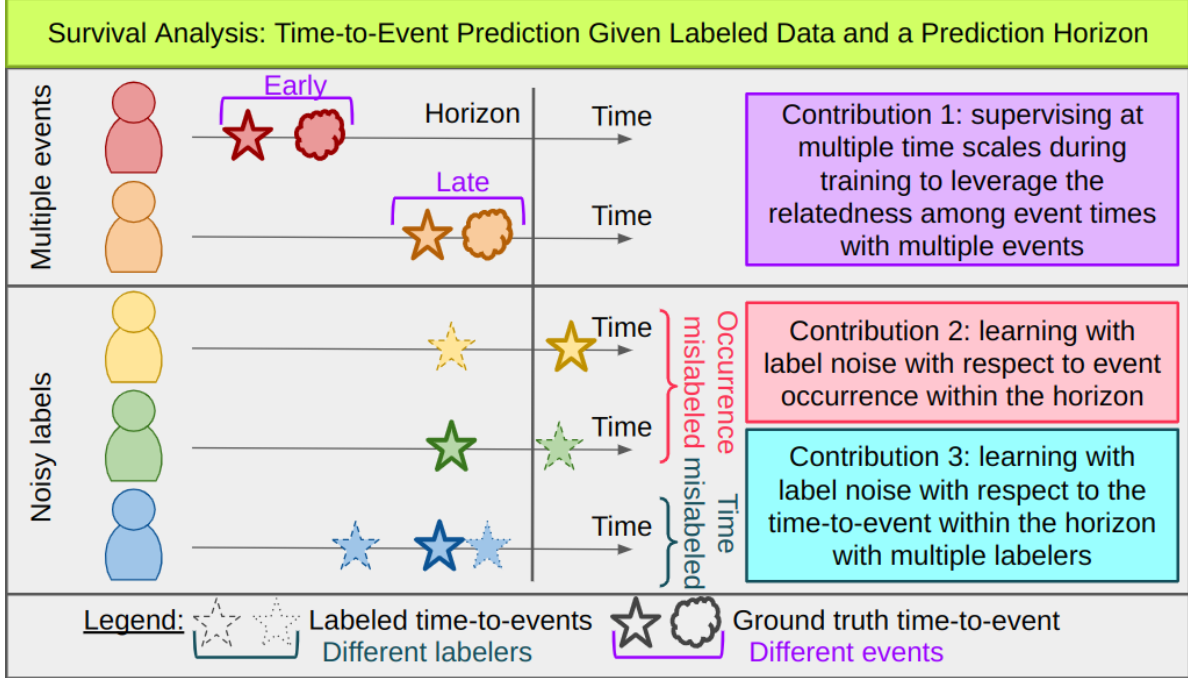


Figure 1.1: An overview of this thesis. We build on existing work in survival analysis and address shortcomings. First, we leverage the relatedness among TTEs in the multi-event setting to more effectively learn the dependencies among events. Then, we consider label noise with respect to whether the event occurred, and then label noise with respect to when the event occurred in the context of multiple labelers.

function. We then consider label noise with respect to when the event occurs in the context of multiple labelers, where we propose to learn to predict 1) the noisy TTEs and 2) their associated error values and map those predictions to a prediction of the ground truth TTE.

## 1.1 Challenges and Opportunities

We tackle open challenges in survival analysis in 1) the multi-event setting and 2) the label noise setting through the development of novel approaches that address the limitations of past work by augmenting the training scheme. We describe relevant background in **Chapter 2**. Our contributions are illustrated in **Figure 1.1** and summarized below.

**Accounting for dependencies among TTEs with multiple events.** We first tackle the challenge of leveraging the dependencies among TTEs in the context of multi-event survival analysis in **Chapter 3**. We focus on modeling multiple events at once [3, 67, 96] due to the added difficulty in learning the potential dependencies among TTEs both within an event and among different events. While past work has examined ways to learn general

dependencies among events [96, 95], they do not explicitly take advantage of the temporal relationships among TTEs at coarser time grains. For example, the time between two events may appear more predictable under a coarser time granularity than the one used to collect the data due to its lower variance. Incorporating this easier task (i.e., predicting at the coarse granularity) into the training scheme may help to improve performance. Our approach does so by breaking down the main task of predicting TTEs at a fixed time granularity into a series of subtasks that build on each other. Each subtask represents a simpler version of the main task and learns to predict the probability of event occurrence at a coarser time granularity. By first considering a more simple task, we build up to the main task by using the predictions from previous tasks to guide predictions at subsequent tasks. We empirically show that our approach is more effective than existing baselines on a variety of tasks.

**Instance-dependent label noise with respect to whether the event occurs.** We then tackle label noise in **Chapter 4** in the context of whether the event occurs, which is analogous to a classification task. Past work from classification often assumes that label noise is independent of the features [7, 38, 160]. However, this is not always true, and label noise can be instance-dependent [179]. Despite the growing popularity of instance-dependent label noise [188, 187, 29, 184], studies that take into account potential differences in the noise rates (i.e., the proportion of mislabeled instances) among subsets of the data are limited [105]. Because of this, approaches that account for label noise are often biased so that performance favors subsets with lower noise rates. We propose an approach to address this issue by learning the underlying pattern of noise using a small subset of data for which we know the ground truth and noisy labels. We assume that this subset is representative of the target population so that patterns learned from it can be generalized to the rest of the dataset. We then use that pattern to re-weight the objective function during subsequent training such that it is unbiased in expectation, where we consider the entire dataset. Empirically, we show that, as we vary the overall noise rate and the difference in noise rates among subsets of the data, our approach is robust compared to state-of-the-art baselines. By accounting for different noise rates in subsets of the data, our approach applies to a wide range of instance-dependent label noise settings.

**Label noise in when the event occurs with multiple labelers.** Finally, we continue to tackle label noise, but this time focus on noise with respect to when the event occurs in the context of multiple labelers in **Chapter 5**. The majority of approaches proposed for the multiple noisy labels setting are proposed in the context of classification, and these approaches are limited when adapted to survival analysis. For example, some approaches

handle multiple labels in classification via aggregation [80, 165, 73, 127], and when adapted to survival analysis, may require assumptions on how the ground truth and noisy TTEs are related that do not always hold (e.g., the ground truth TTEs is the average of the noisy TTEs). Furthermore, labels for survival analysis may be censored, making label aggregation more complex. As a result, it becomes unclear how to aggregate in a straightforward manner. In light of this, we propose to learn in the presence of multiple noisy TTEs by learning to predict 1) the noisy TTEs and 2) their associated error values, and then mapping those predictions to a prediction of the ground truth. Like with the previous contribution, we assume that we have a small subset of ground truth labels that we can leverage to learn this relationship. In doing so, our approach does not require strict assumptions on the relationship between the noisy and ground truth TTEs and can be applied to individuals with censored labels. We show that our approach improves over approaches from classification adapted to survival analysis. By learning with multiple noisy labels, we combine the strengths and weaknesses of the individual labelers to improve predictions.

## 1.2 Practical Value

The challenges we explore are largely informed by tasks from the healthcare domain, where we aimed to predict clinical outcomes, such as Alzheimer’s disease (AD) and sepsis, using electronic health record (EHR) data (i.e., hospital records from patients at a particular health system). For these outcomes, a common yet relevant task is to learn when outcome onset occurs [149, 6, 40]. For example, AD is the sixth leading cause of death in Americans 65 years and older [8], and sepsis is estimated to be present in 30%-50% of hospitalizations that end in death [146]. As a result, early prediction can help clinicians plan appropriately.

The task of predicting onset fits naturally into the survival analysis framework since events directly correspond to the occurrence of onset, and TTEs directly correspond to the time of onset. Moreover, since patients may leave a health system at any time, censored outcomes are common among individuals in EHR-based datasets, and thus the task of predicting when onset occurs cannot be solved by standard regression alone. In our work on predicting the onset of AD [167], patient death was often another event of interest since death either prevents AD onset or is expected to occur four to eight years after AD onset [8]. In our work on predicting both AD and sepsis onset [167, 77], there was no way to label 1) which patients had the condition and 2) when onset occurred that was both reliable and pragmatic. As a result, we relied on automated labeling tools based on the structured components of the EHR, which allowed us to label large datasets conveniently at the cost of accuracy. Thus, both the multi-event and noisy label settings have relevance in real-world tasks.

Although we rely on a small subset of data for which we have both ground truth and noisy labels, our approaches for noisy labels are still applicable in many settings. As mentioned above, we relied on automated labeling tools to label patients with AD and sepsis. When developing automated labeling tools, it is necessary to obtain chart-reviewed labels from a clinical expert for a subset of the data to evaluate the performance of these labeling tools. Thus, we can take advantage of such a subset during training to learn the relationship between the ground truth and noisy labels.

### 1.3 Contributions

In survival analysis, and supervised learning more broadly, we make the following contributions:

- ✓ We propose an approach for multi-event survival analysis that leverages the relatedness among TTE labels. Our approach breaks down the overall task into a series of simpler coarse-grained sub-tasks, where temporal relationships among TTEs are more predictable due to lower variance. The sub-tasks build on each other, allowing for predictions at finer grains to be informed by predictions at coarser grains. We show that our approach leads to more accurate predictions on a variety of datasets.
- ✓ We propose an approach that addresses instance-dependent label noise with respect to whether the event occurs. Our approach uses a small subset of the data for which we have observed and ground truth labels to learn the underlying noise pattern. This pattern is then used to re-weight the loss function and account for potential differences in the noise rates between subsets of the data. We show that our approach is robust to the noise rate across a variety of noise conditions.
- ✓ We propose an approach that addresses label noise with respect to when the event occurs in the context of multiple noisy labels. Our approach first predicts the noisy TTEs and their associated errors and maps that to a prediction of the ground truth TTE. Like the previous contribution, we use a small subset of data for which we have both noisy and ground truth TTEs so that we do not need to place assumptions on how the noisy and ground truth TTEs are related. Compared to baselines adapted from the classification literature, our approach provides a performance improvement.

The rest of this thesis is structured as follows. Chapter 2 introduces background related to concepts from survival analysis and noisy labels. Chapter 3 describes our contributions of work published in AAAI 2021. Chapter 4 describes our contributions of work published in CHIL 2023. Chapter 5 describes our contributions of work in submission.

## CHAPTER 2

# Background

In this chapter, we review topics from survival analysis and classification. For survival analysis, we provide general overviews on modeling a single event, modeling multiple events, noisy (i.e., incorrect) labels, and common evaluation metrics. We include a discussion on classification since our contribution for label noise with respect to whether the event occurs is analogous to a classification task. There, we provide overviews of the problem setup, common evaluation metrics, and ordinal regression.

### 2.1 Survival Analysis

In survival analysis, one aims to model TTEs, i.e., the time it takes until an event occurs. Events can refer to any outcome of interest, and survival in this context, refers to being event free. Survival analysis is used in several fields, such as healthcare [149, 6, 40], engineering [140, 199], economic and political studies [59, 1], sociology [48], and biology [136]. We focus on tasks from the healthcare space, where events usually correspond to clinical outcomes such as disease onset or mortality, and survival usually corresponds to being disease free or not having experienced mortality, respectively. In this section, we describe the problem setup, past work, and introduce the settings of multiple events and noisy labels.

#### 2.1.1 Problem Setup and Notation

Unless otherwise indicated, our main notation is as follows. We denote instance features as  $\mathbf{x} \in \mathbb{R}^d$ , where  $d$  is the number of features, and the TTE as  $e$ . We assume that  $e$  is not deterministic and is drawn from a distribution  $D$  based on  $\mathbf{x}$ . We aim to model TTEs by predicting the survival function,  $S(t)$ . At time point  $t = 1, 2, \dots, T$  over a prediction horizon of length  $T$ , the survival function describes the probability of being event free as  $S(t) = P(e > t)$ . If we plot the survival function with respect to time over the length of the horizon, we obtain the survival curve (**Figure 2.1**). The probability of survival

can alternatively be written as  $S(t|\mathbf{x}) = P(e > t|\mathbf{x})$  if we make individualized predictions conditioned on instance features. Let  $\ell$  be the time-to-followup, or the time at which an instance was last observed (i.e., we do not know what happened after time  $\ell$ ). Let  $c$  be an indicator variable for censorship status, with  $c = 1$  if the instance’s outcome is censored and  $c = 0$  otherwise. Given  $e$ ,  $\ell$ , and  $c$ , the observed time  $o$  is the recorded time in the dataset, which is  $e$  if  $c = 0$  and  $\ell$  if  $c = 1$  (i.e., event free up until time  $\ell$ ). In some parts of the literature,  $c$  is an indicator for event occurrence and its values take the opposite meaning to what we described above. Throughout, unless otherwise indicated, let superscripts in parentheses denote instances (e.g.,  $\mathbf{x}^{(i)}$ ), subscripts denote indexes into vectors (e.g.,  $x_k$ ), and  $\mathbb{I}$  denote the indicator function. We also assume that censored outcomes occur at random with respect to  $\mathbf{x}$  unless otherwise specified, which is a common assumption [79, 31, 96].

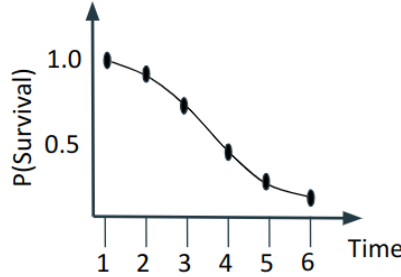


Figure 2.1: Example survival curve.

### 2.1.1.1 Types of Censorship

The description above describes right censorship. That is, we know the event did not occur up until time  $\ell$  but not when it occurred after time  $\ell$ , if at all. In addition to right censorship, instances can also have their outcomes be left censored [47, 49]. With left censorship, we know that the event had already occurred by time  $\ell$ , but we do not know when before  $\ell$  it occurred. In other words, we know that  $e < \ell$  for instances with a left censored outcome. Intermediate of right and left censorship is interval censorship [138, 102], where the event is known to have occurred after one time point but before another. However, it is not known when, during the interval, the event occurred. Instances with interval censored outcomes would then be associated with two points of observation  $\ell_1$  and  $\ell_2$  with  $\ell_1 < e < \ell_2$ . For this thesis, we focus on right censorship.

Since we defined our problem setup as predicting survival relative to some horizon  $T$ , there may be cases where outcomes are right censored in that the instance is known to be event free for the entire horizon (i.e., the event does not occur in the horizon). Here,  $\ell = T$

and  $c = 1$ . This specific case of right censoring is referred to as administrative censoring [72]. In our setup, the label for an instance’s outcome is complete if 1) the event occurred at time  $e \leq T$ , or 2) the event did not occur within the horizon  $e \notin [1, T]$ . Therefore, instances with an administratively censored outcome have complete label information, since we know that  $e \notin [1, T]$ . This is in contrast to an instance with a censored outcome at time  $\ell < T$  with  $c = 1$ . Here, we know that the event did not occur for some portion of the horizon, but we do not know whether it occurred in the remaining portion of the horizon.

## 2.1.2 Predicting Survival Curves

Past work has modeled survival at both the population level [79, 119] and at the instance (i.e., individual) level [31, 96, 144]. Population-level survival curves were proposed in the context of comparing groups of individuals [11, 39, 35] and do not condition on individual features (i.e., they model the survival curve as  $S(t)$ ). Individual-level survival curves condition on individual features (i.e., they model the survival curve as  $S(t|\mathbf{x})$ ).

### 2.1.2.1 Population-Level Survival Curves

Early approaches to predict survival curves focused on doing so at the population level using count-based methods [79, 119]. For  $n$  individuals with dataset  $D = \{o^{(i)}, c^{(i)}\}_{i=1}^n$ , these approaches count, at each time point  $t$ , the number of individuals who experienced the event (i.e.,  $\sum_{i=1}^n \mathbb{I}(o^{(i)} = t \wedge c^{(i)} = 0)$ ) and number of individuals who are event free (i.e.,  $\sum_{i=1}^n \mathbb{I}(o^{(i)} > t \vee (c^{(i)} = 1 \wedge o^{(i)} = t))$ ). These quantities are then used to construct the survival curve. For example, the Kaplan-Meier approach [79] models survival as

$$S(t) = \prod_{u=1}^t \left( 1 - \frac{\sum_{i=1}^n \mathbb{I}(o^{(i)} = u \wedge c^{(i)} = 0)}{\sum_{i=1}^n \mathbb{I}(o^{(i)} > u \vee (c^{(i)} = 1 \wedge o^{(i)} = u))} \right)$$

To compare two groups of individuals, denoted with superscripts <sup>(1)</sup> and <sup>(2)</sup>, using their population-level survival curves, one typically uses the logrank test [111, 16]. The logrank test measures whether the slopes of the ‘failure’ curves are significantly different at each recorded TTE. The failure curve can be considered as a counterpart the survival curve in that each point on the failure curve is 1 minus the corresponding point on the survival curve. To perform the logrank test, we compute the term  $Z$  without loss of generality (WLOG)

$$Z = \frac{\sum_{t \in \tau} \lambda_t^{(1)} - \mu_t^{(1)}}{\sqrt{\sum_{t \in \tau} \nu_t^{(1)}}}$$

where  $\tau$  is the set of distinct TTEs recorded in the dataset,  $\lambda_t^{(1)}$  is the number of instances from group 1 whose event occurred by time  $t$  and  $\mu_t^{(1)} = (\lambda_t^{(1)} + \lambda_t^{(2)}) \frac{\xi_t^{(1)}}{\xi_t^{(1)} + \xi_t^{(2)}}$  is the expected value of  $\lambda_t^{(1)}$  with  $\xi_t^{(i)}$  being the number of instances from group  $i$  whose event had not yet occurred by time  $t$ .  $\nu_t^{(1)}$  is the variance of  $\lambda_t^{(1)}$  and can be written as

$$\nu_t^{(1)} = \mu_t^{(1)} \left( \frac{(\xi_t^{(1)} + \xi_t^{(2)}) - (\lambda_t^{(1)} + \lambda_t^{(2)})}{\xi_t^{(1)} + \xi_t^{(2)}} \right) \left( \frac{\xi_t^{(2)}}{\xi_t^{(1)} + \xi_t^{(2)} - 1} \right)$$

Under the null hypothesis that the the slopes of the failure curves are the same,  $\lambda_t^{(1)}$  follows a hypergeometric distribution parameterized by  $\xi_t^{(1)} + \xi_t^{(2)}$ ,  $\xi_t^{(1)}$ , and  $\lambda_t^{(1)} + \lambda_t^{(2)}$ , resulting in the expressions of  $\mu_t^{(1)}$  and  $\nu_t^{(1)}$  above. If the null hypothesis holds, the distribution of  $Z$  converges to a standard normal as the size of  $\tau$  approaches infinity by the central limit theorem. As a result, one can check for statistical significance by comparing  $Z$  to some threshold significance value based on quantiles of the standard normal distribution. Note that  $Z$  can be computed with respect to either group, hence WLOG above.

### 2.1.2.2 Individual-Level Survival Curves

Count-based approaches generally model survival at the population level, and thus, cannot produce individualized survival curves that are conditioned on instance features. As a result, other types of approaches have been developed for this setting and learn from datasets of the form  $D = \{\mathbf{x}^{(i)}, o^{(i)}, c^{(i)}\}_{i=1}^n$  [31, 180, 42, 37]. One of the most common approaches used to model individualized survival curves is the Cox proportional hazards model [31]. The Cox approach models the survival curve, conditioned on instance features, as  $S(t|\mathbf{x}) = S_0(t)^{\exp(\mathbf{w} \cdot \mathbf{x})}$ , where  $\mathbf{w}$  are learned coefficients describing the linear contribution of each feature to overall survival,  $S_0(t)$  is the baseline survival function, and  $\exp()$  is the exponential function. The baseline survival function represents the survival function when all normalized feature values are equal to 0 (e.g., features are 0-mean normalized so that 0 represents a ‘baseline value’) and can be estimated via counting as shown below [20]. Note that, in order to estimate the baseline survival function, we must first learn the coefficients  $\mathbf{w}$ .

$$S_0(t) = \exp(-\Lambda_0(t)) \quad \text{where} \quad \Lambda_0(t) = \sum_{i=1}^n \frac{\mathbb{I}(o^{(i)} < t \wedge c^{(i)} = 0)}{\sum_{j=1}^n \mathbb{I}(o^{(i)} < o^{(j)}) \exp(\mathbf{w} \cdot \mathbf{x}^{(j)})}$$

Coefficients  $\mathbf{w}$  of the Cox model are learned by maximizing the partial likelihood shown below [32]. Note that, if the goal is to learn the feature contributions to survival and not to

predict survival curves, then learning only  $\mathbf{w}$  is sufficient.

$$PartialLikelihood(\mathbf{w}) = \prod_{i=1}^n \left( \frac{\exp(\mathbf{w} \cdot \mathbf{x}^{(i)})}{\sum_{j=1}^n \mathbb{I}(o^{(i)} < o^{(j)}) \exp(\mathbf{w} \cdot \mathbf{x}^{(j)})} \right)^{\mathbb{I}(c^{(i)}=0)}$$

The Cox approach assumes that individualized survival curves can be produced by learning how much to stretch or shrink the baseline survival curve. This is called the proportional hazards assumption, which states that the ratio of the hazard functions between two instances is the same at all time points, where the hazard function at time point  $t$  is defined as the slope of the failure curve at time  $t$ . Variants of the Cox model have been proposed that do not require the feature contributions to be linear, but they still require the proportional hazards assumption [3, 67, 81].

More recently, others have proposed approaches to model survival without requiring the proportional hazards assumption [96, 95, 144]. For example, some instead assume that the survival curve follows a specific form, such as the lognormal distribution [9]. Alternatively, some have framed the problem as multi-class classification [96], where each time point represents a class, and the model is trained to predict the probability of event occurrence at each time point (i.e., for time point  $t = 1, 2, \dots, T$ , predict  $P(e = t|\mathbf{x})$ ). From these predictions, the survival function can then be calculated as  $S(t|\mathbf{x}) = 1 - \sum_{i=1}^t P(e = i|\mathbf{x})$ . These models are learned by minimizing cross entropy loss, shown below, which is able to recover the empirical survival distribution given sufficient data and only requires the assumption that the ground truth survival distribution is the empirical survival distribution. For uncensored outcomes, cross entropy maximizes the probability of event occurrence at the observed TTE in the dataset. For censored outcomes, cross entropy minimizes the probability of event occurrence before the last point of observation. Here  $\hat{\cdot}$  refers to a model's prediction.

$$CE = -\frac{1}{n} \sum_{i=1}^n \mathbb{I}(c^{(i)} = 0) \log \hat{P}(e^{(i)} = o^{(i)}) + \mathbb{I}(c_k^{(i)} = 1) \log \left( 1 - \left( \sum_{t=1}^{o^{(i)}} \hat{P}(e^{(i)} = o^{(i)}) \right) \right)$$

### 2.1.3 Settings with Multiple Events

So far, we described approaches that model a single event. However, there may be cases where we are interested in modeling more than one event at a time. For example, we may be interested in modeling multiple diseases at once, or comparing the likelihood of disease onset versus mortality over time. In these cases, the setting then becomes survival analysis with multiple events. When there are multiple events, the events may occur in any order, and or there may be settings where there are constraints on the order in which events are allowed

to occur (e.g., competing and semi-competing risks). In the competing risk setting [42], we aim to model multiple events, but from these events, only one is allowed to occur. That is, when one event occurs, it is impossible for the others to occur (e.g., mortality caused by cancer versus mortality caused by a traffic accident). The semi-competing risks setting [43] is slightly relaxed from competing risks in that only some, but not all, events prevent the occurrence of the other events (e.g., mortality and onset of AD).

#### 2.1.4 Noisy Labels in Survival Analysis

Finally, we review work in survival analysis addressing label noise (i.e., when the labeled TTEs can be inaccurate). Earlier work in the field has proposed extensions to the Kaplan-Meier estimator [139] and Cox proportional hazards model [158, 114]. More recent work has shifted to considering the scenario where TTEs that are known to be recorded correctly are available for a small subset of the data [109]. Works from this space usually refer to such a set as a ‘validation’ set, though we do not use this terminology in order not to confuse this set with the validation set typically used for model selection during training.

#### 2.1.5 Evaluation of Survival Models

To evaluate survival analysis models, past work has used the concordance index (C-index) [56], a metric similar to the area under the receiver operating characteristic curve (AUROC) from classification, to measure discriminative performance. Briefly, the C-index represents the probability that a randomly selected pair of instances, where one is known to experience the event before the other, are correctly ranked. That is, if  $e^{(1)} < e^{(2)}$  then  $S(e^{(1)}|\mathbf{x}^{(1)}) < S(e^{(1)}|\mathbf{x}^{(2)})$  is a correct ranking. We use the time-dependent version proposed by Antolini et al. [5], which formulates the C-index as (where  $\hat{\cdot}$  denotes a prediction)

$$C = \frac{\sum_{i,j=1}^n \mathbb{I}((o^{(i)} < o^{(j)} \vee (o^{(i)} = o^{(j)} \wedge c^{(j)} = 1)) \wedge c^{(i)} = 0) \mathbb{I}(\hat{S}(o^{(i)}|\mathbf{x}^{(i)}) < \hat{S}(o^{(j)}|\mathbf{x}^{(j)}))}{\sum_{i,j=1}^n \mathbb{I}((o^{(i)} < o^{(j)} \vee (o^{(i)} = o^{(j)} \wedge c^{(j)} = 1)) \wedge c^{(i)} = 0)}$$

The denominator represents the number of instance pairs that can be ranked, and the numerator represents the number of instance pairs that are correctly ranked.

In addition to discriminative performance, researchers measure whether the predicted survival curves are calibrated. That is, do the predicted survival curves match the ground truth survival curve. The most straightforward way to measure this would be to take the integrated mean squared error (IMSE) between the ground truth survival curve and the predicted survival curve as  $IMSE = \frac{1}{T} \sum_{t=1}^T (S(t|\mathbf{x}) - \hat{S}(t|\mathbf{x}))^2$ . Alternatively, one could also use the integrated mean absolute error (IMAE). However, in nearly all real-world

tasks, we do not have access to ground truth survival curves. Thus, an alternative method to measure calibration error is based on the following interpretation of the survival curve: if the predicted probability of survival at time  $e$  is  $a\%$  for some individual, then similar individuals (who would have similar survival curves) experience the event after time  $e$   $a\%$  of the time [52]. Therefore, for  $a\%$  of individuals, the probability of survival at their TTE should be less than or equal to  $\frac{a}{100}$ . This means that we should expect a histogram of the probabilities of survival at each individual’s TTE to follow a uniform distribution. Past work uses this observation to quantify the difference between these distributions to measure calibration [52, 78, 137].

Finally, past work has also evaluated the accuracy of the predicted TTEs, where the predicted TTE can be taken as the median value of the predicted survival function (i.e.,  $\hat{e} = \operatorname{argmin}_t \hat{S}(t|\mathbf{x}) \leq 0.5$ ) or the mean value of the survival function (i.e.,  $\hat{e} = \sum_{t=1}^T tP(e = t)$ ). [13, 137]. TTE accuracy can then be measured as the squared difference (i.e.,  $(e - \hat{e})^2$ ), absolute difference (i.e.,  $|e - \hat{e}|$ ) or signed difference (i.e.,  $e - \hat{e}$ ).

## 2.2 Classification

In a standard classification task, one aims to assign a label to an instance from one of several categories, or classes. This is usually done through estimating the probability for being in each class. In **Chapter 4**, we focus on classification in the context of label noise, where observed class labels in the dataset do not always reflect an instance’s ground truth label. Incorrect, or noisy, labels are widely studied [45, 160] and can be found in many applications, such as automated labeling tools [168, 121, 197, 66] and crowdsourced labeling [33, 143, 64]. In this section, we introduce the problem setup and describe some metrics used for evaluating classification models. We also briefly introduce ordinal regression as a more specialized type of classification due to its connection to **Chapter 3**.

### 2.2.1 Problem Setup and Notation

The main notation we use throughout is as follows. Let  $\mathbf{x} \in \mathbb{R}^d$  denote instance features, where  $d$  is the number of features. Each instance is associated with a ground truth label, denoted  $y$ , that indicates in which class it belongs. In addition to a ground truth label, each instance is associated with an observed label  $\tilde{y}$ . The observed label also serves as an indicator for the class in which the instance belongs, but unlike the ground truth label, it can sometimes be incorrect (i.e., sometimes  $\tilde{y} \neq y$ ). If label noise is instance-independent, then whether an instance is mislabeled does not depend on its features (i.e.,  $P(\tilde{y} == y) = P(\tilde{y} == y|\mathbf{x})$ ).

Otherwise, the label noise is said to be instance-dependent. When label noise is instance-dependent, some instances are more likely to be mislabeled than others based on their features. Thus, this introduces bias with respect to which types of instances are mislabeled, resulting in differences in the noise rates (i.e., percentage of mislabeled instances) among subsets of the data. Datasets used to learn in the presence of label noise are generally of the form  $D = \{\mathbf{x}^{(i)}, \tilde{y}^{(i)}\}_{i=1}^n$ , where  $n$  is the number of instances in the dataset.

### 2.2.2 Evaluation

For classification models, we consider two forms of evaluation. The first is discriminative performance, which is usually measured using the AUROC. This describes the probability that a randomly chosen pair of instances are correctly ranked. For example, in binary classification, a positive instance should have a higher predicted probability for being in the positive class than a negative instance. The second is performance with respect to model bias. This can be measured using the area under the equalized odds curve (AUEOC) [34]. For classification threshold  $\tau$ , we calculate the equalized odds (EO) [55], denoted  $EO(\tau)$ , between two groups, called 1 and 2, as shown below.  $TP_a(\tau)$  and  $FP_a(\tau)$  denote true and false positive rates for group  $a$  at threshold  $\tau$ , respectively. The AUEOC is obtained by plotting the EO value for all possible  $\tau$  and calculating the area under the curve.

$$EO(\tau) = \frac{2 - |TP_1(\tau) - TP_2(\tau)| - |FP_1(\tau) - FP_2(\tau)|}{2}$$

In general, metrics for measuring bias from the fairness literature fall into three main categories: independence, sufficiency, and separation [12]. The first evaluates whether the proportion of predicted positive examples is the same among groups (e.g., demographic parity), which results in independence between group membership and the proportion of identified positive examples. We did not use metrics from this category because in our setting of focus (i.e., healthcare), many conditions have established differences in prevalence among patient groups. For example, the prevalence of Alzheimer’s disease is higher in countries with a larger proportion of individuals above 65 years [206]. The second category considers sufficiency (i.e., are the model’s predictions well calibrated within groups) and evaluates whether the positive and negative predictive values are the same among groups (e.g., predictive parity). However, this measure is sensitive to class balance, which can differ among groups in healthcare (e.g., women are more likely to suffer from breast cancer compared to men). In contrast, EO (the metric we rely on in later chapters) falls into ‘separation’, which evaluates whether the true and false positive rates among groups are the same. We selected this measure because it is applicable to many settings in healthcare [132, 141, 133] in that the consequences of failing

to treat a patient in need [135, 18], or giving an inappropriate treatment [17, 118] can be serious. However, it is important to note that any given bias/fairness metric is unlikely to apply to all scenarios. Therefore, finding the most appropriate metric for a specific use case is task-dependent and would require collaboration with a subject expert.

### 2.2.3 Ordinal Regression

Beyond standard classification, some tasks may involve classifying objects into categories such that there is an inherent ordering among classes. For example, items could be rated on a likert scale of 1 to 5, where a rating of 1 is poor, a rating of 5 is excellent, and the quality of the item increases with each numerical score. This setting is referred to as ordinal regression and has been used in many applications [14, 22, 198, 131]. For a more thorough review, see Gutierrez et al. [51]

While ordinal regression could be framed as standard regression [89], this setup is limited in that standard regression assumes that the difference between labels corresponds to a measure of distance, when this is not always the case with the classes from ordinal regression. From our example with the likert scale, although a rating of 2 is better than a rating of 1, there may not be a way to meaningfully quantify the distance between a ‘1’ and a ‘2’. This limitation is addressed by learning to regress on a continuous latent variable and then learning the thresholds that map the latent variable to the ordinal classes [171]. This way, the concept of distance between the ordinal variables becomes more flexible, as they do not need to be defined before model training.

Similarly, framing ordinal regression as a standard (i.e., nominal) classification problem [57] is also limited in that standard classification does not leverage the inherent ordering among classes, thus leading to sample inefficiency. Alternative to regressing on a latent continuous variable and learning the mapping thresholds, another class of approaches from ordinal regression builds on the nominal classification framing. For example, approaches learn to model different parts of the overall task and then combine predictions for each part to obtain a prediction for the overall task [44]. Continuing with our example of the 5-point likert scale, approaches from this class could first learn to model  $P(y > 1)$ ,  $P(y > 2)$ , ...,  $P(y > 4)$  for  $y \in \{1, 2, 3, 4, 5\}$ , and then use these predictions to obtain  $P(y = 5) = P(y > 4)$ ,  $P(y = 4) = P(y > 3) - P(y > 4)$ ,  $P(y = 3) = P(y > 2) - P(y > 3)$ ,  $P(y = 2) = P(y > 1) - P(y > 2)$ , and  $P(y = 1) = 1 - P(y > 1)$ . Here, it is assumed that the tasks  $P(y > 1)$ ,  $P(y > 2)$ , ...,  $P(y > 4)$  are easier to solve than predicting each ordinal class individually.

## CHAPTER 3

# Hierarchical Multi-Event Survival Analysis

### 3.1 Introduction

**Motivation and Problem Setting.** Survival analysis, commonly used in fields such as healthcare, aims to estimate the probability of an event occurring over some time horizon [93, 183]. In contrast to single-event analysis, multi-event survival analysis is more challenging due to the potentially complex interdependence among events. For example, in addition to learning which patient characteristics contribute to early disease onset, one would also need to learn how this relates to the patient’s likelihood of experiencing some other outcome, like mortality. In light of this challenge, researchers commonly make the simplifying assumption that the events are independent and apply existing single-event analysis techniques [149]. However, when this assumption does not hold, it can lead to poor performance [23].

**Gaps in Existing Work.** Recently, Lee et al. [96] used multitask learning to jointly model events. Each task within the multitask framework corresponds to a specific event and frames survival prediction as a multiclass classification problem, where each class corresponds to the probability of event occurrence at a specific time point. While the multitask framing addresses the independence assumption in part, it remains limited in that it does not explicitly leverage the relationships among TTE labels at different time granularities both within an event and among different events. For example, the difference in timing between two events may appear more predictable under a coarser time granularity than the one used to collect the data due to lower variance. Incorporating the coarser granularity, which we hypothesize is easier to learn, into the training scheme may help to improve performance.

**Our Idea.** To address the issue described above, we propose a novel hierarchical approach that jointly models events. Our approach improves performance by solving a series of simpler tasks (i.e., survival prediction at a coarser granularity), building up to the original target task (i.e., survival prediction at the original time granularity). Within an event, we first divide the original survival prediction task into a series of simpler tasks that serve as guides to help

complete the original task. Our hierarchical approach *iteratively predicts survival at multiple granularities of time* (e.g., months, weeks, then days). Predictions at coarser time scales (e.g., will the event happen during the prediction horizon or not) are used to guide predictions at finer time scales (a more complex task), improving the overall accuracy on the original task. Although our approach can be used with any numbers of events, we hypothesize that we will observe the most improvement in the multi-event setting since dependencies are likely to exist among TTEs both within and among events. We do not assume any constraints on the order in which events may occur (e.g., competing or semi-competing events) [60], but we outline ways to account for such constraints if needed.

Applied to publicly available data, we also present *a comprehensive evaluation of the discriminative performance of our proposed approach* relative to existing baselines. To date, evaluation of survival analysis models in the multi-event setting has been limited. Common evaluation techniques like the C-index [56, 5] focus on discriminative performance at the global level (i.e., comparing the risk of two different individuals for the same event) but not at the local level (i.e., comparing the risk for two different events for the same individual). We propose new evaluation metrics to evaluate discriminative performance at the local level.

Overall, we introduce a novel, hierarchical approach for multi-event survival analysis. Across several datasets, both synthetic and real, we show that the proposed method consistently outperforms several well-accepted baselines. Our contributions are as follows:

- We propose a novel, hierarchical, deep learning approach for multi-event survival analysis that models survival hierarchically along the prediction horizon, using predictions at coarser time scales to guide predictions at finer time scales
- We highlight the shortcomings of evaluating discriminative performance in the setting of multi-event survival analysis, present a novel evaluation framework, and use it for additional supervision during training

**Organization.** The rest of this chapter is as follows. First, we present some background information on how previous work 1) addresses multi-event survival analysis and 2) evaluates models in the context of multi-event survival analysis. Then, we describe our hierarchical approach and how the relatedness among time points can be leveraged to improve predictions. Finally, we describe experiments on synthetic and real data in which we evaluate our approach against well-accepted baselines from the literature.

## 3.2 Related Work

Here, we set up the problem of multi-event survival analysis and review the current approaches as well as the existing evaluation methods, highlighting limitations.

### 3.2.1 Problem Setup and Definitions

We aim to predict the probability of survival (i.e., the probability of an event *not* occurring) over time for multiple events based on an individual’s features or covariates. At training time, we assume that we are provided with the covariates, observed times, and censoring statuses for each individual. At test time, we are provided with only the covariates. For each event, an individual’s *observed time* is the time at which the event occurred or the time at which the individual was last observed, depending on the censoring status.

### 3.2.2 Multi-Event Survival Analysis

The majority of past work in multi-event survival analysis focuses on extending the Cox proportional hazards model [31]. Among the most common Cox-based approaches to multi-event survival analysis is that proposed by Larson [94], in which each event is modeled separately. This approach is widely accepted in practical applications of multi-event survival analysis [6, 159], but it fails to capture the dependencies among events. For example, if our events are disease onset and mortality and disease onset is likely, then the probability for mortality may increase. Alternative to modeling events separately, others have proposed the multi-state model, which learns to model event transitions separately. For example, a model is learned for experiencing mortality conditioned on previously experiencing disease onset [3]. However, by separately modeling each event transition, this approach also fails to leverage shared information among the events.

The limitation of leveraging dependencies among events has been addressed for Cox models by using a shared frailty term across models (i.e., an individual specific variable whose effect on the hazard is multiplicative) [67], additional forms of regularization [175], or by modeling the joint survival distribution [60]. However, these approaches require alternative assumptions on the distribution of the frailty term, the proportional hazards assumption, or assumptions on the form of the joint survival distribution, respectively. Although recent approaches [96, 37] have been able to model dependencies without placing alternative assumptions by using deep learning in a multi-task framework, they are also limited in that they do not leverage easier subtasks at coarser time granularities during training that can potentially be used to learn relationships among events more efficiently. Although this short-

coming is also relevant in the single-event setting, it is especially important in the multi-event setting since time points are not only related to each other within an event, but also to the time points of other events. We note here that, although one could use techniques from ordinal regression [51] to account for the ordering among time points, such approaches do not assume that classes have any quantifiable distance between them. Time points, on the other hand, have quantifiable differences that can be leveraged during training (e.g., the number of days between weeks). Despite the difference in settings, our approach leverages concepts similar to some approaches from ordinal regression, as described below.

Our proposed approach leverages a series of simpler subtasks to guide the predictions of the original task. To date, the concept of using simpler tasks to aid with predictions at more complex, finer-grained tasks has not been explored in survival analysis. However, such an approach has been used for classification [88, 115, 110, 151]. For example, in text classification, Kowsari et al. showed how training a classifier to learn to discriminate among super-classes (e.g., novels and research papers) can aid in learning to discriminate among sub-classes (e.g., fiction novels and review papers). By breaking the overall task into a series of nested subtasks, the model can more effectively accomplish the original task. Similar concepts have been explored in ordinal regression [57], where approaches first learn to predict easier tasks that represent coarser grained versions of the original task and then combine predictions. These approaches differ from ours in the types of coarse grained tasks that they explore. Generally, coarse grained tasks from these approaches do not nest within each other and iteratively build their way up to the original task as with our approach.

### 3.2.3 Evaluation with Multiple Events

Methods for evaluating the discriminative performance of approaches for multi-event survival analysis have been limited. Many report the learned coefficients of the Cox model [2, 130]. However, this requires knowledge of how all possible combinations of the features contribute to the outcome. In the context of multiple events, alternative metrics, like the C-index [56], have their own issues. Here, we focus on two shortcomings of existing metrics: 1) they can result in biased estimates with respect to portions of the curve, and 2) they only consider rankings among individuals within a single event.

The C-index is perhaps the most commonly used evaluation metric for discriminative performance in survival analysis [56]. The C-index is a weighted average of the AUROC at different time points [58]. Within an event, the C-index assumes that rankings are *consistent* over time (i.e, if two individuals are correctly ranked at time 1, then they are correctly ranked at all other time points). However, for deep approaches to survival analysis, there

is no guarantee that the curves will not cross. To address this assumption, Antolini et al. proposed an extension that compares the probabilities of survival at the *earlier* point of observation within each pair instead of some fixed time point for all pairs [5]. However, this comparison can result in an incomplete estimate of performance, since it does not consider the *later* points of observation within the pair, i.e., the time points when the first individual has experienced the event but the second individual has not. We propose an alternative consistency score that compares individuals at *all* relevant time points.

For multiple events, the C-index is typically computed for each event separately. As a result, this compares individuals within the context of a specific event. While this captures discriminative performance *among* individuals (inter-individual, intra-event) at the *global* level, it does not capture discriminative performance *within* an individual (inter-event, intra-individual) at the *local* level (i.e., it does not evaluate the survival curves among events within an individual). We go beyond the notion of global rankings and consider local rankings.

### 3.3 Methods

We introduce a novel approach for multi-event survival analysis that improves the accuracy of previous work. Our architecture contains a hierarchical component that uses predictions at coarser time scales (a simpler task) to guide predictions at finer time scales (a more complex task). During training, we rely on a composite loss function composed of both established and novel ranking penalties that provide supervision across the entire prediction horizon. Unless otherwise mentioned, we assume censoring is non-informative.

#### 3.3.1 Notation

Let  $n$  be the number of individuals,  $K$  be the number of events, and  $T$  be the size of the discrete time horizon over which we consider survival. We define the ‘original task’ as predicting the probability of event occurrence at each time point  $t = 1, 2, \dots, T$ .

Each individual  $i = 1, 2, \dots, n$  is associated with three variables:  $\{\mathbf{x}^{(i)}, \mathbf{c}^{(i)}, \mathbf{o}^{(i)}\}$ . Let  $\mathbf{x}^{(i)} \in \mathbb{R}^D$  be the vector of covariates. Let  $\mathbf{c}^{(i)} \in \{0, 1\}^K$  be the vector of censoring statuses for each event, where  $c_k^{(i)}$  is 1 if event  $k$  is censored and 0 otherwise. Let  $\mathbf{o}^{(i)} \in \{1, 2, \dots, T\}^K$  be the vector of observed times for each event. Let  $\mathbf{e}^{(i)} \in \{1, 2, \dots, T\}^K$  be the vector of TTEs, where  $o_k^{(i)} = e_k^{(i)}$  if  $c_k^{(i)} = 0$ , and  $o_k^{(i)} < e_k^{(i)}$  otherwise.  $\mathbf{e}^{(i)}$  is not provided in the data. Let  $\mathcal{S}^{(i)} = \{\mathbf{S}_1^{(i)}, \mathbf{S}_2^{(i)}, \dots, \mathbf{S}_K^{(i)}\}$  be a matrix of survival curves for individual  $i$ .  $\mathbf{S}_k^{(i)} = (P(e_k^{(i)} > 1), P(e_k^{(i)} > 2), \dots, P(e_k^{(i)} > T))$  describes the probability of individual  $i$  *not* experiencing event  $k$  until time point  $t$  for  $t = 1, 2, \dots, T$ . Let  $\mathcal{P}^{(i)} = \{\mathbf{p}_1^{(i)}, \mathbf{p}_2^{(i)}, \dots, \mathbf{p}_K^{(i)}\}$  be a set

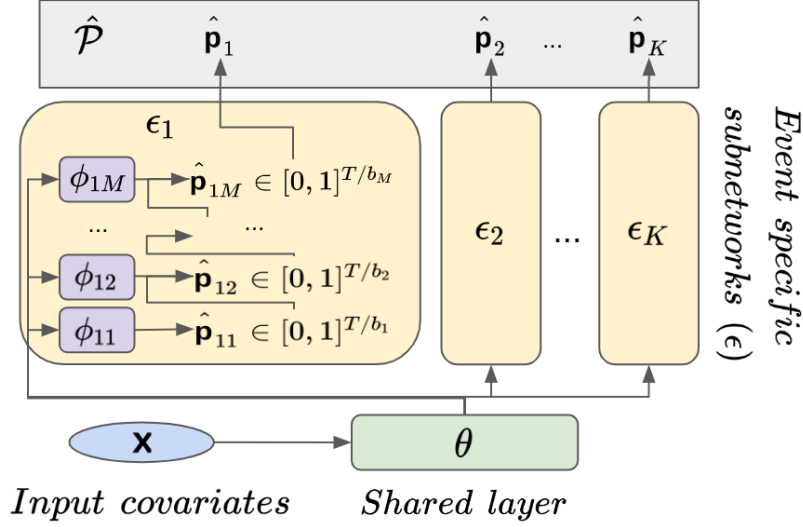


Figure 3.1: Proposed architecture. A multitask network that takes  $\mathbf{x}$  as input and outputs  $\hat{\mathcal{P}}$ , a set of probability distributions for each event. Event specific subnetworks,  $\epsilon_k$ , estimate the probability distribution,  $\hat{\mathbf{p}}_k$ , hierarchically. At granularity  $m$ ,  $\hat{\mathbf{p}}_{km}$  is constructed from  $\phi_{km}(\theta(\mathbf{x}))$  and  $\hat{\mathbf{p}}_{km-1}$ .

of probability distributions for individual  $i$ .  $\mathbf{p}_k^{(i)} = (P(e_k^{(i)} = 1), P(e_k^{(i)} = 2), \dots, P(e_k^{(i)} = T))$  describes the probability of  $i$  experiencing event  $k$  at  $t = 1, 2, \dots, T$ . It then follows that, given  $\mathcal{P}^{(i)}$ , we can calculate the corresponding  $\mathcal{S}^{(i)}$ .

Let  $(b_1, b_2, \dots, b_M)$  be a vector of size  $M$ , where  $M$  is the number of temporal granularities. At granularity  $m$ , for  $m = 1, 2, \dots, M$ , we split the horizon into  $T/b_m$  non-overlapping time bins, where the number of time points from the original task contained in each time bin of the task at granularity  $m$  is  $b_m$ .  $b_1 = T$  is the coarsest granularity (i.e., it has one time bin containing all time points from the original task), and  $b_M = 1$  is the granularity of the original task. We assume that  $b_m$  are monotonically decreasing, and  $b_{m-1} \bmod b_m \equiv 0$ . The time bin in which  $o_k^{(i)}$  falls at granularity  $m$  is then denoted as  $o_{km}^{(i)}$ . Let  $\tau_z^{(m)}$  be a set of time points from the original task, regrouped under granularity  $m$ , for coarse bin indexes  $z = 1, 2, \dots, T/b_m$ . For example, if  $T = 6$  and  $b_m = 3$ , then we split the horizon in half with  $\tau_1^{(m)} = \{1, 2, 3\}$  and  $\tau_2^{(m)} = \{4, 5, 6\}$ .  $p_{km}^{(i)}[z] = \sum_{u \in \tau_z^{(m)}} p_k^{(i)}[u]$  is then the probability of event occurrence in  $\tau_z^{(m)}$ . We denote indexes into probability vectors with square brackets (e.g.  $p[t]$ ) and use a ‘hat’ for the model’s estimates (e.g.,  $\hat{\mathbf{p}}$ ).

### 3.3.2 Overall Architecture

The components of our approach are incorporated into a multitask architecture (**Figure 3.1**) with a shared layer  $\theta$  to account for dependencies from different granularities among

events. It contains  $K$  event specific subnetworks,  $\epsilon_1, \epsilon_2, \dots, \epsilon_K$ , where  $\epsilon_k$  takes the output from  $\theta$  as input and constructs  $\hat{\mathbf{p}}_k$  in a hierarchical manner, as described below. From  $\hat{\mathcal{P}}$ , the survival curves,  $\hat{\mathcal{S}}$ , can be derived as  $S_k^{(i)}(t|\mathbf{x}^{(i)}) = 1 - \sum_{u=1}^{t-1} p_k^{(i)}[u]$  for time point  $t$ .

### 3.3.2.1 Hierarchical Component

We iteratively predict whether the event will occur at increasingly finer time scales and hypothesize that it improves performance by more accurately predicting the coarser grained task. In the example from **Figure 3.2** ( $T = 4$ ,  $M = 3$ ), we aim to predict the discrete probability of event occurrence at  $t = 1$  (i.e.,  $P(e = 1)$ ). Instead of directly predicting  $P(e = 1)$ , we begin at the coarsest granularity ( $m = 1$ ,  $b_1 = 4$  and  $\tau_1^{(1)} = \{1, 2, 3, 4\}$ ), and predict  $P(e \in \tau_1^{(1)})$ . Then, we move to  $m = 2$  ( $b_2 = 2$ ,  $\tau_1^{(2)} = \{1, 2\}$ , and  $\tau_2^{(2)} = \{3, 4\}$ ). We predict the probability that the event occurs in the first half of the horizon,  $\tau_1^{(2)}$ , conditioned on it occurring within the horizon (i.e.,  $P(e \in \tau_1^{(2)}|e \in \tau_1^{(1)})$ ). Next, we move to  $m = 3$  (the original task), where  $b_3 = 1$ . We then predict  $P(e \in \tau_1^{(3)}|e \in \tau_1^{(2)}) = P(e = 1|e \in \tau_1^{(2)})$  and recover  $P(e = 1) = P(e = 1|e \in \tau_1^{(2)})P(e \in \tau_1^{(2)}|e \in \tau_1^{(1)})P(e \in \tau_1^{(1)})$ .

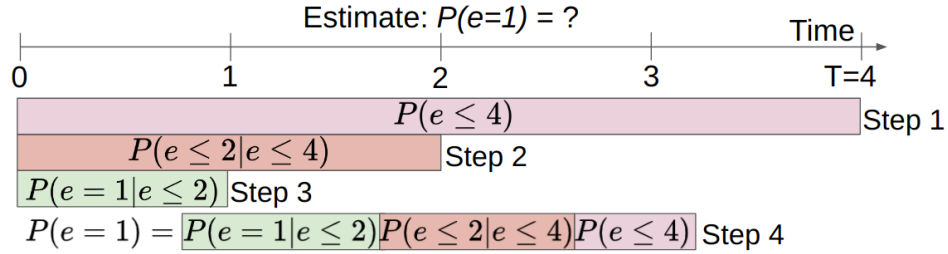


Figure 3.2: Example prediction. We predict if the event will occur within the horizon, and then iteratively predict at finer time scales until reaching the original time scale.

*Claim 1.* Predicting the coarser grained guiding task is a lower variance task compared to predicting the original, finer grained task directly.

*Justification.* Assume for the original task that an individual's TTE,  $e$ , is a random variable drawn from some discrete distribution conditioned on their covariates  $\mathbf{x}$ , where the probability of each time point occurring is non-zero. Let  $\mathbb{E}_O(e)$  and  $\text{Var}_O(e)$  be the expected value and variance of  $e$ , respectively, under the original task,  $O$ . Next, consider a guiding task,  $G$ , that uses coarser grained time bins, where each coarse grained time bin contains  $b$  original time bins. For coarse grained time bin  $\tau_z$ ,  $z = 1, 2, \dots, T/b$ ,  $P(e \in \tau_z) = \sum_{u \in \tau_z} P(e = u)$ . Time bins from the original task that fall into the same coarse grained time bin are equivalent under the guiding task, and thus  $\text{Var}_G(e) < \text{Var}_O(e)$ , where  $\text{Var}_G(e)$  is

the variance of  $e$  under  $G$ .

$$\begin{aligned}
\mathbb{E}_O(e) &= \left[ \sum_{t=1}^T tP(e=t) \right] \\
\mathbb{E}_G(e) &= \left[ \sum_{t=1}^T \left( \sum_{u \in \tau_{\lceil t/b \rceil}} \frac{u}{b} \right) P(e=t) \right] \\
\text{Var}_O(e) &= \sum_{t=1}^T P(e=t)(t - \mathbb{E}_O(e))^2 \\
&= \sum_{t=1}^T P(e=t) \left( t - \left[ \sum_{t=1}^T tP(e=t) \right] \right)^2 \\
\text{Var}_G(e) &= \sum_{t=1}^T P(e=t) \frac{1}{b} \left( \sum_{u \in \tau_{\lceil t/b \rceil}} u - \mathbb{E}_G(e) \right)^2 \\
&= \sum_{t=1}^T P(e=t) \frac{1}{b} \left( \sum_{u \in \tau_{\lceil t/b \rceil}} u - \left[ \sum_{t=1}^T \left( \sum_{u \in \tau_{\lceil t/b \rceil}} \frac{u}{b} \right) P(e=t) \right] \right)^2
\end{aligned}$$

The maximum value of  $\text{Var}_O(e)$  is greater than the maximum value of  $\text{Var}_G(e)$  because there are more 0-valued terms in the multiplier of  $P(e=t)$  in  $\text{Var}_G(e)$  than in  $\text{Var}_O(e)$ . Therefore, we hypothesize that we can learn the coarse grained probabilities more accurately. In doing so, we constrain the maximum error of the resulting survival curve for the original task more tightly, compared to predicting the original task directly. We show this below and provide more detail in **Appendix A**. By lowering the upper bound on the error, we hypothesize that this will, in turn, improve performance.

*Claim 2.* If we can more accurately predict the coarser grained probabilities, then we can more tightly constrain the maximum error of the resulting survival curve.

*Justification.* We define error with respect to the survival curve as the IMAE between the predicted and ground truth survival curve.

$$\text{IMAE} = \frac{1}{T} \sum_{t=1}^T |\hat{S}(t|\mathbf{x}) - S(t|\mathbf{x})|$$

If we assume that the coarser grained predictions are more accurate, then the IMAE will be lower at the coarse grained bin edges

$$\frac{1}{T/b} \sum_{z=1}^{T/b} |\hat{S}_G(zb|\mathbf{x}) - S(zb|\mathbf{x})| < \frac{1}{T/b} \sum_{z=1}^{T/b} |\hat{S}_O(zb|\mathbf{x}) - S(zb|\mathbf{x})|$$

where  $\hat{S}_G$  and  $\hat{S}_O$  are the predicted survival curves for the guiding and original tasks, respectively, and  $S$  is the ground truth survival curve. Then within each coarse grained bin

$\tau_z$ , the error can be upper bounded for the coarse grained and direct tasks by

$$\begin{aligned} & |\hat{S}_G(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| + |S(zb+1|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \\ & + |\hat{S}_G(z(b+1)|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \text{ (Guiding task)} \end{aligned}$$

and

$$\begin{aligned} & |\hat{S}_O(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| + |S(zb+1|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \\ & + |\hat{S}_O(z(b+1)|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \text{ (Original task)} \end{aligned}$$

Then if we take the sum of the upper bounds over all segments  $\tau_z$  for  $z = 1, 2, \dots, T/b$ , and compare the guiding task with the original task, (note that we can ignore the  $|S(zb+1|\mathbf{x}) - S(z(b+1)|\mathbf{x})|$ — term since it is identical for both tasks)

$$\begin{aligned} & \frac{1}{T/b} \sum_{z=1}^{T/b} |\hat{S}_G(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| + |\hat{S}_G(z(b+1)|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \\ & < \frac{1}{T/b} \sum_{z=1}^{T/b} |\hat{S}_O(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| + |\hat{S}_O(z(b+1)|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \end{aligned}$$

We show why this is in **Appendix A**  $\square$  Overall, this shows that predicting more accurately at the coarser grain allows us to anchor our predictions at the finer grain to what we learned at the coarser grained. Similar ideas apply to the IMSE.

The hierarchical component is implemented within each  $\epsilon$  subnetwork (**Figure 3.1**) as described in **Algorithm 1**. For event  $k$ , we first predict at the coarsest granularity (i.e., the horizon) via  $\phi_{k1}$  (line 1). We then consider grains  $m = 2, 3, \dots, M$  iteratively (line 2). For granularity  $m$ , we compute the conditional probability distribution (line 3; e.g.,  $P(e_k \in \tau_1^{(m)} | e_k \in \tau_1^{(m-1)})$ ), and recover the marginal probabilities for each time bin (e.g.,  $P(e_k \in \tau_1^{(m)})$ ) via multiplication with the corresponding marginal probability from granularity  $m-1$  (e.g.,  $P(e_k \in \tau_1^{(m-1)})$ ) (lines 4-6). After reaching granularity  $M$ , we will have computed corresponding the probabilities of the original task (lines 7-8).

For each event  $e$ , we account for individuals who do not experience event  $e$  by time  $T$  by predicting  $\hat{P}(e \in \tau_1^{(1)}) = \hat{P}(e \leq T)$ . We implement this within the  $\epsilon$  subnetworks as part of the probability predictions via our first step (i.e., coarsest grain) predictions.

---

**Algorithm 1** Hierarchical prediction for event  $k$ .  $\hat{\mathbf{p}}_{km}$  is the predicted probability distribution at granularity  $m$ .

---

**Input:**  $\theta(\mathbf{x})$ , learned representation from  $\theta$

**Output:**  $\hat{\mathbf{p}}_k$ , estimated distribution of occurrence at the original time scale

Hierarchical\_Prediction( $\theta(\mathbf{x})$ )

```

1:  $\hat{\mathbf{p}}_{k1} \leftarrow \phi_{k1}(\theta(\mathbf{x}))$   $\triangleright \hat{P}(e \leq T)$ 
2: for  $m = 2$  to  $M$  do
3:    $\hat{\mathbf{p}}_{km} \leftarrow \phi_{km}(\theta(\mathbf{x}))$   $\triangleright$  conditional probabilities
4:   for  $z = 1$  to  $T/b_m$  do
5:      $z_{m-1} \leftarrow \lfloor (z-1)b_m/b_{m-1} \rfloor + 1$   $\triangleright$  time index from previous granularity
6:      $\hat{\mathbf{p}}_{km}[z] \leftarrow \hat{\mathbf{p}}_{km}[z] \times \hat{\mathbf{p}}_{km-1}[z_{m-1}]$   $\triangleright$  combine predictions to get bin probability
7:  $\hat{\mathbf{p}}_k \leftarrow \hat{\mathbf{p}}_{kM}$ 
8: return  $\hat{\mathbf{p}}_k$ 

```

---

### 3.3.3 Training Loss

We learn the model parameters by minimizing a composite loss  $\mathcal{L} = \mathcal{L}_{TTE} + \alpha \mathcal{L}_g$ . Scalar hyperparameter  $\alpha \in \mathbb{R}^+$  controls the tradeoff between the losses.

$\mathcal{L}_{TTE} = \left( \sum_{m=1}^M -\mathcal{L}_{TTE}^m \right)$  maximizes the likelihood of each event when it occurs, for uncensored individuals, and the likelihood of survival for each individual at each event until their last observation time, for censored individuals. Unlike previous work [96, 144], we compute  $\mathcal{L}_{TTE}$  over multiple time scales (**Eq. 3.1**).

$$\mathcal{L}_{TTE}^m = \sum_{i=1}^n \sum_{k=1}^K \mathbb{I}(c_k^{(i)} = 0) \log \left( \hat{p}_{km}^{(i)}[o_{km}^{(i)}] \right) + \mathbb{I}(c_k^{(i)} = 1) \log \left( 1 - \left( \sum_{t=1}^{o_{km}^{(i)}} \hat{p}_{km}^{(i)}[t] \right) \right) \quad (3.1)$$

$\mathcal{L}_g = \sum_{k=1}^K \mathcal{L}_g^k$  maximizes global consistency and further improves performance by encouraging the model to correctly rank the relevant pairs of individuals. For event  $k$ , let  $\mathcal{C}_k$  be the set of comparable individuals (defined below). We compute  $\mathcal{L}_g$  at the finest time scale for each event (**Eq. 3.2**) where  $F(s_1, s_2, \sigma) = \exp((s_1 - s_2)/\sigma)$ ,  $o_k^{*(j)} = o_k^{(j)}$  if  $c_k^{(j)} = 1$  and  $o_k^{*(j)} = o_k^{(j)} - 1$  otherwise (explained below).  $\sigma_g$  is a scalar hyperparameter.

$$\begin{aligned} \mathcal{L}_g^k &= \sum_{i,j \in \mathcal{C}_k} F \left( \hat{s}_k^{(i)}[o_k^{(i)}], \hat{s}_k^{(j)}[o_k^{(j)}], \sigma_g \right) + F \left( \hat{s}_k^{(i)}[o_k^{*(j)}], \hat{s}_k^{(j)}[o_k^{*(j)}], \sigma_g \right) \\ \mathcal{C}_k &:= \{(i, j) | c_k^{(i)} = 0 \wedge (o_k^{(i)} < o_k^{(j)} \vee (o_k^{(i)} = o_k^{(j)} \wedge c_k^{(j)} = 1))\} \end{aligned} \quad (3.2)$$

Previous work [96] only uses the forward consistency term of  $\mathcal{L}_g^k$  (i.e., the first term of  $\mathcal{L}_g^k$ ), which corresponds to the rankings considered by the C-index. When evaluating global

rankings, comparable individuals are compared at the earlier observation point only. This ignores the time steps between the two observation points, and could lead to inconsistent rankings (**Figure 3.3**).

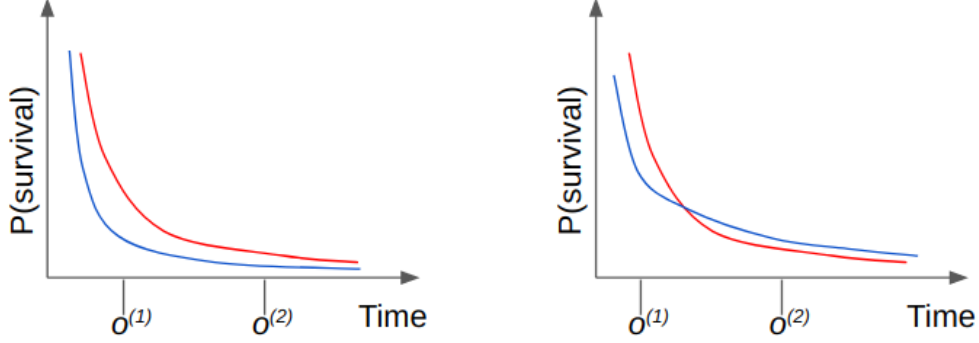


Figure 3.3: Consistency example. Individuals  $i = 1, 2$  are shown by the blue and red curves, with TTE  $o^{(1)}$  and observed time  $o^{(2)}$ , respectively. Left: curves are consistent at  $o^{(1)}$  and  $o^{(2)}$ . Right: curves are inconsistent at  $o^{(2)}$ .

*Claim 3.* Let  $i = 1, 2$  be a comparable pair of individuals with observed times,  $o^{(1)}$  and  $o^{(2)}$ , respectively, such that  $o^{(1)} < o^{(2)}$ . Comparing at time  $o^{(1)}$  only is insufficient.

*Justification.* During the time between  $o^{(1)}$  and  $o^{(2)}$ ,  $i = 1$  experienced the event, while  $i = 2$  did not. The last time point at which this is known to be true is  $o^{(2)}$  if  $c^{(2)} = 1$  and  $o^{(2)} - 1$  otherwise (i.e.,  $o^{*(2)}$ ). Thus, at  $o^{*(2)}$ , we expect that  $s^{(1)}[o^{*(2)}] < s^{(2)}[o^{*(2)}]$ . This is captured by the second term of **Eq. 3.2**. Thus, we encourage the network to correctly rank individuals throughout the horizon.

### 3.3.4 Censorship Due to (Semi-)Competing Events

Until now, we considered the general multi-event case, where there were no constraints on the ordering of events. This assumed that all events occur eventually (i.e.,  $s[t] \rightarrow 0$  as  $t \rightarrow \infty$ ). However, this is not true in some cases of informative censoring, where the occurrence of one event can be prevented by another, such as with competing [42] and semi-competing [43] events. For example, cancer relapse and death are semi-competing events, where death prevents the occurrence of relapse but not vice versa. For relapse, we account for this constraint by predicting the probability that it will never occur, even past  $t = T$ . We denote events that do not occur within the prediction horizon as  $e \notin \{1, 2, \dots, T\}$ . Our approach accounts for these cases, or cases in general where an event may not occur within

the prediction horizon, through the first level of granularity in our predictions, where we predict  $P(e \leq T)$ . This implicitly also predicts  $P(e \notin \{1, 2, \dots, T\}) = 1 - P(e \leq T)$ .

## 3.4 Experimental Setup

We evaluate our proposed approach across different publicly available datasets and compare to several baselines. We hypothesize that our hierarchical approach will lead to more accurate survival curves and consistent improvements across a range of evaluation metrics that measure discriminative performance, compared to existing approaches.

### 3.4.1 Datasets

We used four datasets. One was a multi-event synthetic dataset for which we know ground truth, facilitating evaluation. The remainder were real, publicly available datasets from the health domain. Of the real datasets, one of them contained competing risks while the other two were semi-competing risks. Outcome rates are reported in **Table 3.1**.

| Dataset       | N      | D     | K | T   | Event Sequence: %                                                                                      |
|---------------|--------|-------|---|-----|--------------------------------------------------------------------------------------------------------|
| Synthetic     | 5,000  | 15    | 2 | 20  | $o_1$ : 20.9, $o_2$ : 20.4<br>$o_1, o_2$ : 4.4, $o_2, o_1$ : 4.3                                       |
| ADNI          | 1,604  | 965   | 2 | 60  | AD: 19.3, D: 0.5<br>AD, D 0.6                                                                          |
| MIMIC<br>-III | 13,801 | 3,822 | 3 | 12  | A: 5.2, S: 1.5, D: 4.3<br>A, S: 5.0, A, D: 0.7<br>A, S, D: 1.9<br>S, A: 0.3, S, D: 0.5<br>S, A, D: 0.2 |
| SEER          | 11,374 | 689   | 2 | 120 | DC: 6.9, DP: 8.2                                                                                       |

Table 3.1: Dataset summaries.  $T$  is in months for ADNI and SEER, and hours for MIMIC-III. ‘D’, ‘A’, ‘S’, and ‘DC’/‘DP’ mean death, ARF, shock, and death from larynx cancer/pulmonary disease, respectively.

#### 3.4.1.1 Synthetic

This dataset is based on a synthetic dataset from previous work [96] and contains two events, where both can occur in any order. It serves as a sanity check. We generated 5,000 individuals, each with 15 covariates. The covariates and TTEs were drawn from the distributions described below.  $\mathbf{X}_{y:z}$  denotes  $\mathbf{X}$  from covariate indexes  $y$  to  $z$ , inclusive.  $U$

denotes a uniform distribution. Given the TTEs, a time horizon was chosen at the 50<sup>th</sup> percentile among the times of the first event experienced to allow for an equal proportion of individuals who experience no event and at least one event.

$$\begin{aligned}
\mathbf{X}_{1:5} &\sim U(-5, 5)^5, \mathbf{X}_{6:15} \sim U(-10, 10)^{10} \\
u_1 &\sim \text{Lognormal}((\mathbf{1}^T \mathbf{X}_{1:5})^2 + (\mathbf{1}^T \mathbf{X}_{6:10})^2, 0.4) \\
u_2 &\sim \text{Lognormal}((\mathbf{1}^T \mathbf{X}_{1:5})^2 + (\mathbf{1}^T \mathbf{X}_{11:15})^2, 0.4) \\
v_1 &\sim u_1 + \text{Lognormal}(0.5(\mathbf{1}^T \mathbf{X}_{1:5})^2, 0.4) \\
v_2 &\sim u_2 + \text{Lognormal}(0.5(\mathbf{1}^T \mathbf{X}_{1:5})^2, 0.4) \\
e_1 &= u_1 \text{ if } u_1 < u_2, \text{ } v_1 \text{ otherwise} \\
e_2 &= u_2 \text{ if } u_2 < u_1, \text{ } v_2 \text{ otherwise}
\end{aligned}$$

The distributions from which the TTEs were drawn were chosen such that the skew of the TTE distributions could be controlled, since TTEs in real data is often skewed. Given our initial continuous-valued TTEs, we discretized them into 20 equally spaced temporal segments. We included more variability in the last 10 covariates of  $\mathbf{X}$  to allow for more variation in the TTE distribution. The squared terms introduce non-linear relationships between the covariates and TTE. The dataset was also constructed such that the time to the second event is conditioned on the first event, introducing dependencies among events. More specifically, the second time to event is taken as the corresponding  $v$  value instead of the corresponding  $u$  value had it occurred first.

### 3.4.1.2 ADNI

This dataset is a publicly available dataset containing data on AD<sup>1</sup>. We considered the semi-competing outcomes of AD onset and death as events. We selected a cohort of 1,604 participants, excluding left censored individuals. We considered a prediction horizon of five years, where time was measured relative to the individual’s first encounter with the ADNI study, and we predicted at the monthly level ( $T = 60$ ). After preprocessing, described below, each individual had 965 covariates. Overall, 28.7% of the population was event free by  $t = 60$ , and 50.8% of the population were censored due to loss of followup.

**Preprocessing.** Unless otherwise mentioned, all data was taken from the ‘AD-NIMERGE’ table. Individuals were labeled as follows. AD onset was taken as the first encounter where the value of the ‘DX’ column was ‘Dementia’, where the time to onset was taken as the value in the ‘Month’ column. Date of death was taken as the value of the

‘NPDOD’ column in the ‘NEUROPATH.04.12.18’ table, and the time to death was taken as the difference between the date of death and the date of the first visit in months. The date of the first visit was taken as the value of the ‘update\_timestamp’ column at which the corresponding value in the ‘Month’ column was zero. For those who did not experience an event throughout their period of observation or by 60 months, the last point of observation was the minimum of the largest corresponding value in the ‘Months’ column and 60.

Covariates, including cognitive test scores, AD-specific biomarkers, medical imaging data, and demographics, were pre-processed such that continuous covariates were discretized based on quintiles, discrete covariates were encoded as one-hot vectors, and infrequent (frequency=1) or overly-frequent (frequency=1,603) covariates were removed. Additionally, we discarded the symbols ‘<’ and ‘>’ from all numerical covariates whose value included such a symbol. For example, if a value was recorded as ‘> 1.1’, we took the value of 1.1. This was done before binning continuous covariates into quintiles.

### 3.4.1.3 MIMIC-III

This dataset is a publicly available dataset of EHR data [75]. Experiments from this dataset were not regulated as human subjects research since the data were de-identified. Unlike ADNI, the data cover a smaller time scale (e.g., hours rather than years). We considered three semi-competing outcomes: 1) acute respiratory failure (ARF), 2) shock, and 3) in hospital mortality (death). MIMIC-III includes data pertaining to vital signs, medications, diagnostic and procedure codes, and laboratory measurements. Each event was defined as described by Oh et al. [124]. We considered a prediction horizon of 12 hours, where time was measured relative to six hours after the start of an intensive care unit (ICU) encounter, and we predicted at the hourly level ( $T = 12$ ). We used data from the first 30 minutes of the ICU encounter to avoid label leakage, excluding encounters where an event occurred within the first six hours. This resulted in a sample of 13,801 ICU encounters corresponding to 10,947 patients. After preprocessing, described below, each encounter was associated with 3,822 covariates. Overall, 76.3% of the population was event free by  $t = 12$ , and 4.1% of the population did not experience an event due to loss of followup.

**Preprocessing.** Data were extracted from the following tables: PATIENTS, ADMISSIONS, ICUSTAYS, CHARTEVENTS, LABEVENTS, MICROBIOLOGYEVENTS,

---

<sup>1</sup>Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD.

OUTPUTEVENTS, DATETIMEEVENTS, INPUTEVENTS.MV, and PROCEDUREEVENTS.MV, and were processed using the Flexible Data Driven pipeLine (FIDDLE), [162], a publicly available pre-processing tool for EHR data. To preprocess the tables from MIMIC-III into input for FIDDLE, we used the procedure from previous work. We then used the default settings of FIDDLE ( $\text{theta}_1 = 0.001$ ,  $\text{theta}_2 = 0.001$ , and  $\text{theta\_freq} = 1.0$ ). In addition, we used  $T = 5$  hours, and  $dt = 0.5$  hours. As model input, we used the first slice of the output corresponding to the first 30 minutes of the ICU visit.

#### 3.4.1.4 SEER (<https://seer.cancer.gov/data/>)

This is a publicly available dataset containing cancer incidence data from population-based registries, used by previous work to study competing risks [96, 200]. We considered the competing outcomes of death due to larynx cancer and death due to pulmonary disease. We selected a cohort of 11,374 participants between ages 60-65, excluding left censored individuals. We considered a prediction horizon of 10 years, where time was measured relative to the individual’s age, and we predicted at the monthly level ( $T = 120$ ). Covariates included features relating to demographics, tumor behavior, and tumor characteristics. After preprocessing, see below, each individual was associated with 689 covariates. Overall, 21.0% of the population was event free by  $t = 120$ , and 63.9% of the population did not experience an event due to loss of followup.

**Preprocessing.** An individual’s last observed time was taken according to the ‘Survival months’ feature. Individuals whose value was 9999 were excluded, as this indicated an unknown observed time. Cause of death was recorded from the ‘Cause of Death to SEER site re-code’ feature. Only individuals whose value was ‘00000’, ‘22020’, or ‘50130’ were included, as these codes corresponded to being alive, experiencing death due to larynx cancer, and death due to pulmonary disease, respectively. That is, we excluded individuals who died from causes other than the ones we considered. We used the ‘yr1975\_2016.seer9/RESPIR.TXT’ file as our source of data. Covariates were pre-processed such that continuous covariates were discretized based on quintiles, discrete covariates were encoded as one-hot vectors.

### 3.4.2 Baselines

We compare our approach with the following baselines. We also perform an ablation study as outlined in **Section 5.2**.

**Independent:** This approach learns a separate model for each event and assumes that the time-to-events among all events are independent. This baseline, though simplistic, is most commonly used in clinical research [40]. Our implementation uses a series of separate

feed forward networks. It is trained to predict at the given time scale only, minimizing the forward consistency term of  $\mathcal{L}_g$ .

**DeepHit:** This approach [96] models all events simultaneously in a single model, but was designed for competing risks (where the occurrence of one event prevents the occurrence of the other events) in that it assumes  $\sum_{t=1}^T \sum_{k=1}^K p_k[t] = 1$ . It accounts for dependencies among events via  $\theta$  but not for individuals who experience no event by time  $T$ . It is trained using the forward consistency term of  $\mathcal{L}_g$  and predicts at the time scale of the original task only. For the non-competing risks datasets, we adapt DeepHit by instead assuming  $\sum_{t=1}^T p_k[t] = 1$  for  $k = 1, 2, \dots, K$ . For the semi-competing risks datasets, we also predict  $P(e \notin \{1, 2, \dots, T\})$ . This baseline allows us to explore whether approaches for competing risks generalize to other multi-event settings.

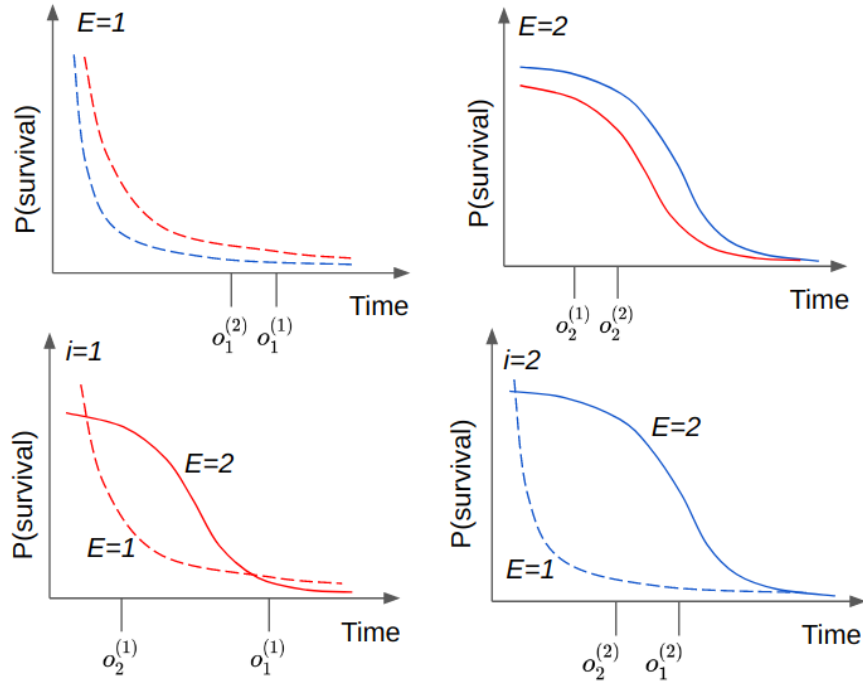


Figure 3.4: Local consistency. The red and blue curves represent individuals  $i = 1, 2$ , and the dashed and solid curves represent events  $E = 1, 2$ , respectively. The curves are globally (top), but not locally (bottom) consistent.  $i = 1$ : inconsistent at  $o_2^{(1)}$ .  $i = 2$ : inconsistent at  $o_1^{(2)}$  and  $o_2^{(2)}$ . We assume that both events occur in both individuals.

**Oracle:** An ‘approach’ with access to the ground truth TTE distributions that we use on the synthetic data only. It provides an upper bound for performance.

### 3.4.3 Evaluation

We evaluate using measures of both global and local consistency. For global evaluation, we report the average C-index, as in previous work [96, 81], and average global consistency across all events. For event,  $k$ , we compute the global consistency as in **Eq. 3.3**.  $|\mathcal{C}_k|$  is the size of  $\mathcal{C}_k$  (i.e., the number of comparable individuals). Let  $\mathcal{T}_k(i, j)$  (defined below) be the set of comparable time points for individuals  $i$  and  $j$  with size  $|\mathcal{T}_k(i, j)|$ .

$$\frac{1}{|\mathcal{C}_k|} \sum_{i, j \in \mathcal{C}_k} \frac{1}{|\mathcal{T}_k(i, j)|} \sum_{t \in \mathcal{T}_k(i, j)} \mathbb{I}(\hat{s}_k^{(i)}[t] < \hat{s}_k^{(j)}[t]) \quad (3.3)$$

$$\mathcal{T}_k(i, j) := \{t | c_k^{(i)} = 0 \wedge (o_k^{(i)} \leq t < o_k^{(j)} \vee (o_k^{(i)} = t = o_k^{(j)} \wedge c_k^{(j)} = 1))\}$$

Unlike past work, we measure local consistency, as it is critical that multi-event models can discriminate among events *within* an individual (**Figure 3.4**). We compute it for individual  $i$  as in **Eq. 3.4**. Let  $\mathcal{C}^i$  be the set of comparable events with size  $|\mathcal{C}^i|$ , and let  $\mathcal{T}^i(j, k)$  be the set of comparable time points for events  $j$  and  $k$  with size  $|\mathcal{T}^i(j, k)|$  (defined below). We report the average among all individuals.

$$\frac{1}{|\mathcal{C}^i|} \sum_{j, k \in \mathcal{C}^i} \frac{1}{|\mathcal{T}^i(j, k)|} \sum_{t \in \mathcal{T}^i(j, k)} \mathbb{I}(\hat{s}_j^{(i)}[t] < \hat{s}_k^{(i)}[t]) \quad (3.4)$$

$$\mathcal{C}^i := \{(j, k) | c_j^{(i)} = 0 \wedge (o_j^{(i)} < o_k^{(i)} \vee (o_j^{(i)} = o_k^{(i)} \wedge c_k^{(i)} = 1))\}$$

$$\mathcal{T}^i(j, k) := \{t | c_j^{(i)} = 0 \wedge (o_j^{(i)} \leq t < o_k^{(i)} \vee (o_j^{(i)} = t = o_k^{(i)} \wedge c_k^{(i)} = 1))\}$$

### 3.4.4 Implementation Details

For each dataset, we randomly split the data into 60/20/20% training/validation/test, and data from the same individual did not appear across splits. All models were trained in Python3.6 and Pytorch [128], using Adam [84]. Hyperparameters, including the learning rate, L2 regularization constant, and objective function scalars (e.g.,  $\alpha$ ), were tuned using a random grid search, with a budget of 20. We used early stopping based on validation set performance, where we aimed to maximize the average of the proposed global and local consistencies. All network layers were initialized with Xavier initialization from a uniform distribution. We report results on the held-out test set, with error bars representing empirical 95% confidence interval (CI)s from 1,000 bootstrapped samples. More details about the implementation are here: [https://github.com/MLD3/Hierarchical\\_Survival\\_Analysis](https://github.com/MLD3/Hierarchical_Survival_Analysis).

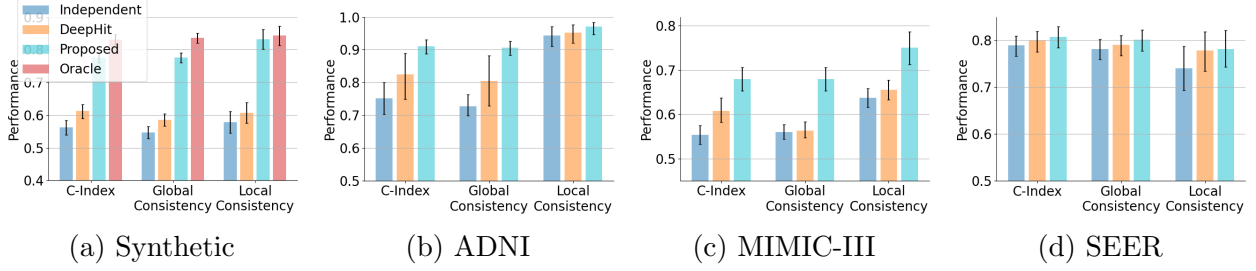


Figure 3.5: Comparison to existing approaches. Our approach outperforms state-of-the-art approaches in different multi-event settings across different metrics. Error bars represent empirical 95% confidence intervals. In order, the datasets represent the general multi-event setting, semi-competing risks, semi-competing risks, and competing risks.

## 3.5 Results and Discussion

Here, we evaluate the proposed approach on our four datasets.

### 3.5.1 How Does the Proposed Approach Perform?

As shown in **Figure 3.5**, the proposed method noticeably outperforms both baselines on all datasets. For SEER, the proposed method achieved a global and local consistency of 0.80 (95% CI=0.77-0.82) and 0.78 (95% CI=0.74-0.82), while using DeepHit achieved a global and local consistency of 0.79 (95% CI=0.77-0.82) and 0.77 (95% CI=0.73-0.81), respectively. The poorer performance of our adaptation of DeepHit to the non-competing risks datasets suggests that it does not generalize well to other settings. We also include a qualitative comparison of the survival curves from our approach with those of when we predict the original task directly (e.g., DeepHit) on the synthetic data. From eight randomly selected examples from the held-out test set, we observe that there is only one sample on which the survival curve from our approach resembles the ground truth less closely (**Figure 3.6**).

### 3.5.2 Do Hierarchical Predictions Help?

To assess the extent to which the hierarchical component improved performance, we performed an ablation study. Our ablations are as follows. To control for the number of network parameters, the overall network size was kept constant. **Adapted DeepHit (DA)**: We generalized previous work to the multi-event setting [96] and implemented it as described in our adaptation of DeepHit to the non-competing risks datasets. **Proposed minus hierarchical (-HA)**: We built from DA such that  $\mathcal{L}_g$  penalizes at the first *and last* relevant time points, and we account for individuals who remain event free by time  $T$ . It assesses

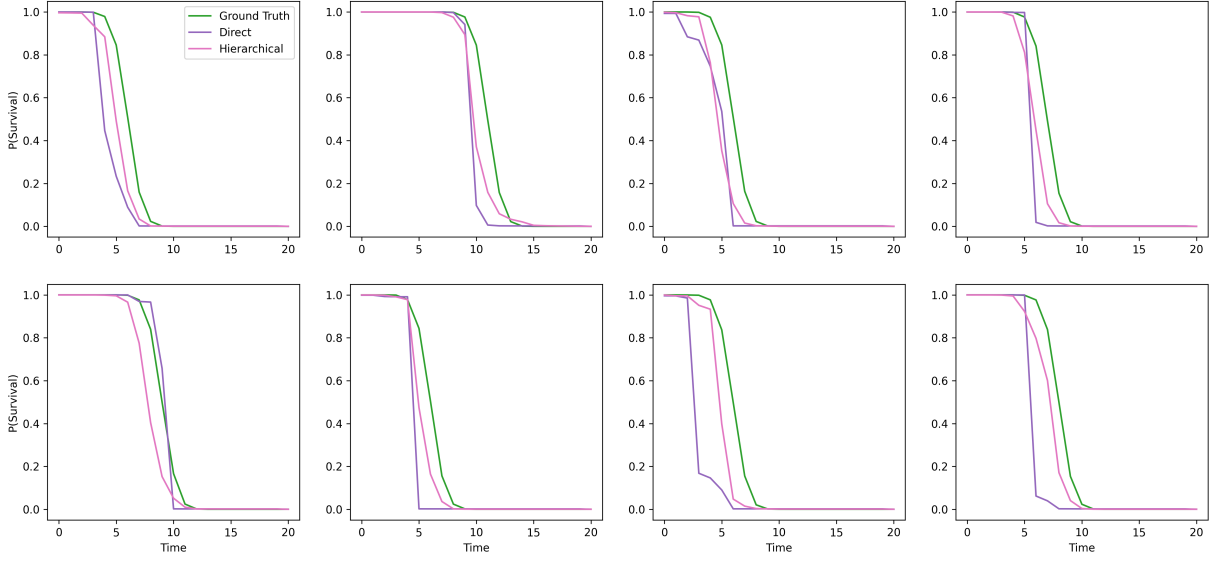


Figure 3.6: Comparison of survival curves. For seven out of eight randomly selected instances from the test set, the survival curves from our approach are similar to or more closely match the ground truth compared to predicting the original task directly.

the proposed approach without the hierarchical component. **DA +architecture (A)**: This ablation builds from -HA in that it utilizes the proposed architecture as opposed to a vanilla feed forward network. It serves to assess the effectiveness of the hierarchical architecture. **DA +supervision (S)**: This ablation builds from -HA in that we supervise along multiple time scales instead of that from the original task. It serves to assess the effectiveness of the hierarchical supervision. Our **proposed method (P)** combines A and S builds from -HA in that we utilize the hierarchical architecture and loss function.

Results are shown in **Table 3.2**. First, notice that the addition of *the hierarchical component improves* over the baselines across all datasets and all evaluation metrics. The magnitude of improvement is larger in MIMIC-III than the other real datasets. We hypothesize that this is due to the increased difficulty from modeling three events instead of two. The larger performance gains in local consistency compared to the other metrics on most datasets also show that the proposed approach is effective in modeling interdependencies among events. Second, increased supervision over  $\mathcal{L}_g$  and accounting for individuals who remain event free by time  $T$  led to large improvements in performance. This is likely because over 20% of individuals in each dataset were event free throughout the horizon. In addition, we observe that both the hierarchical architecture (A) and supervision along multiple time scales (S) improve performance compared to -HA.

| Dataset   | Approach | C-index              | Global Consistency   | Local Consistency    |
|-----------|----------|----------------------|----------------------|----------------------|
| Synthetic | DA       | .61(.59-.63)         | .58(.56-.60)         | .60(.57-.63)         |
|           | -HA      | .73(.71-.75)         | .73(.72-.75)         | .77(.74-.81)         |
|           | A        | .76 (.74-.77)        | .76 (.74-.77)        | .82 (.79-.85)        |
|           | S        | .73 (.71-.74)        | .73 (.72-.75)        | .79 (.76-.82)        |
|           | P        | <b>.77</b> (.76-.79) | <b>.77</b> (.76-.79) | <b>.83</b> (.80-.86) |
| ADNI      | DA       | .82(.74-.88)         | .80(.72-.78)         | .95(.92-.97)         |
|           | -HA      | .90(.88-.92)         | .89(.87-.91)         | .95(.90-.98)         |
|           | A        | <b>.91</b> (.89-.93) | <b>.90</b> (.89-.92) | .95 (.90-.98)        |
|           | S        | .90 (.88-.92)        | <b>.90</b> (.88-.92) | .96 (.93-.98)        |
|           | P        | <b>.91</b> (.88-.93) | <b>.90</b> (.88-.92) | <b>.97</b> (.94-.98) |
| MIMIC-III | DA       | .60(.58-.63)         | .56(.54-.58)         | .65(.63-.67)         |
|           | -HA      | .66(.63-.69)         | .66(.63-.69)         | .74(.71-.78)         |
|           | A        | .67 (.65-.70)        | .67 (.65-.70)        | <b>.75</b> (.71-.78) |
|           | S        | .67 (.64-.70)        | .67 (.64-.70)        | <b>.75</b> (.71-.78) |
|           | P        | <b>.68</b> (.65-.70) | <b>.68</b> (.65-.70) | <b>.75</b> (.71-.78) |
| SEER      | DA       | .79(.76-.81)         | .78(.75-.80)         | .75(.71-.80)         |
|           | -HA      | .79(.77-.81)         | .79(.76-.81)         | .76(.72-.80)         |
|           | A        | <b>.80</b> (.78-.82) | .79 (.77-.82)        | .77 (.72-.81)        |
|           | S        | <b>.80</b> (.78-.82) | <b>.80</b> (.77-.82) | .77 (.73-.81)        |
|           | P        | <b>.80</b> (.78-.83) | <b>.80</b> (.77-.82) | <b>.78</b> (.74-.82) |

Table 3.2: Ablation study. The highest values are bold. Error bars show empirical 95% CIs. Abbreviations are as follows: DA: Adapted DeepHit for multiple events, -HA: Proposed without hierarchical component, A: DA with hierarchical architecture, S: DA with hierarchical supervision, P: Proposed approach

## 3.6 Conclusion

We introduced a novel approach for multi-event survival analysis. It utilizes a novel hierarchical structure that divides the task of predicting event occurrence over a time horizon  $T$  into a series of simpler, yet practically relevant, subtasks that iteratively build on one other. On both synthetic and real datasets, we showed that the proposed method led to improvements over well-accepted baselines across different performance metrics. Through ablations, we demonstrated the effectiveness of the hierarchical component.

Going forward, one might consider extensions for recurrent events. It may also be worth considering how to optimally partition the horizon to improve performance. Although our approach focused on the general concept of dividing the horizon into a series of easier subtasks, how exactly the horizon should be divided has yet to be answered (e.g., should we divide the horizon in half or into thirds). In addition, it may also be interesting to explore

how different partitions can be combined, since the approach currently has the limitation that it cannot take advantage of temporal relationships at the bin edges. Nonetheless, this work represents an important foundation for future work in multi-event survival analysis.

## CHAPTER 4

# Learning with Instance-Dependent Label Noise

### 4.1 Introduction

**Motivation and Problem Setting.** Datasets used to train machine learning models can contain incorrect labels (i.e., label noise), which can lead to overfitting. In survival analysis, there are two layers of label noise. The first is noise with respect to whether the event occurred, and the second is noise with respect to the timing of the event. Here, we consider noise with respect to whether the event occurred, which is analogous to a classification task. Thus, we consider classification more broadly in this chapter. While label noise in classification is widely studied, the majority of past work focuses on instance-independent label noise (i.e., when the noise is independent from an instance’s features) [160]. However, label noise can depend on instance features [179, 24], leading to different noise rates within subsets of the data. Furthermore, in settings where the noise rates differ with respect to a sensitive attribute, this can lead to harmful disparities in model performance [105]. For example, consider the task of predicting cardiovascular disease among patients admitted to a hospital. Compared to male patients, female patients may be more likely to be under-diagnosed [113] and thus mislabeled, potentially leading to worse predictions for female patients. Although instance-dependent label noise has recently received more attention [29, 187, 173], the effect of these approaches on model bias has been relatively understudied [105]. Here, we address current limitations and propose a novel method for learning with instance-dependent label noise in a setting inspired by healthcare, specifically examining how modeling assumptions affect existing issues around potential model bias.

**Gaps in Existing Work.** Broadly, current work addressing instance-dependent label noise takes one of two approaches: 1) learn to identify mislabeled instances [28, 186, 204], or 2) learn to optimize a noise-robust objective function [41, 178]. In the first category, instances identified as mislabeled are either filtered out [83] or relabeled [15]. In some settings, this

approach can have a negative effect on model bias. From our example on cardiovascular disease, approaches that filter out mislabeled individuals could ignore more female patients, since they have a potentially higher noise rate. While relabeling approaches use all available data, they can be sensitive to assumptions around the noise distribution [92]. In the second category, current approaches rely on objective functions that are less prone to overfitting to the noise and use all of the data [25]. However, past work has empirically shown that these improve discriminative performance the most when used to augment filtering approaches, and thus, the limitations described above still potentially hold.

**Our Idea.** In light of these limitations, we propose an approach that addresses instance-dependent label noise, makes no assumptions about the noise distribution, and uses all data during training. We focus on a setting that frequently arises in healthcare, where we are given observed labels for a condition of interest (e.g., cardiovascular disease) and have a clinical expert who can evaluate whether the observed labels are correct for a small subset of the data (e.g., by manual chart review). Using this subset, which we refer to as the ‘alignment’ set, we learn the underlying pattern of label noise in a pre-training step. We then minimize a weighted cross-entropy over all the data. Note that our alignment set is a special case of anchor points [104], with the added requirement that the set contains instances for which the ground truth and observed labels do *and* do not match.

On synthetic and real data, we demonstrate that our approach improves on state-of-the-art baselines from the noisy labels and fairness literature, such as stochastic label noise [25] and group-based peer loss [174]. Overall, our contributions include:

- A novel approach to learn from datasets with instance-dependent noise that highlights a setting frequently found in healthcare
- A systematic examination of different settings of label noise, evaluating discriminative performance and bias mitigation
- Empirical results showing that the proposed approach is robust to both to the noise rate and amount of noise disparity between subgroups, reporting the model’s ability to maintain discriminative performance and mitigate potential bias
- A demonstration of how performance of the proposed approach changes when assumptions about the alignment set are violated

**Organization.** The rest of this chapter is as follows. First, we present some background information on how previous work addresses label noise 1) in the context of general classification, 2) in the context of fairness, and 3) with the use of anchor points. Then, we describe our approach and how anchor or alignment sets can be used to address label noise

and mitigate bias. Finally, we describe experiments on synthetic and real data in which we evaluate our approach to state-of-the-art baselines from the literature.

## 4.2 Related Work

Past approaches generally fall into two categories. The first aims to identify mislabeled instances via filtering heuristics or by estimating the probability of mislabeling [129, 99, 29]. The second adapts regularization techniques, such as label smoothing and contrastive learning to the noisy labels setting [112, 116, 192, 103]. We also examine how biases in instance-dependent label noise have been tackled in the fairness literature. Finally, we review how past work has leveraged additional information in cases where both ground truth and observed labels are available for a small subset of the data.

### 4.2.1 Identifying Mislabeled Data

Approaches that learn to identify mislabeled instances fall into two sub-categories: 1) filtering approaches and 2) relabeling approaches. Filtering approaches use heuristics to identify mislabeled instances (e.g., MentorNet [70], Co-teaching [54], FINE [83]). Many are based on the idea that correctly labeled instances are easier to classify than mislabeled instances (i.e., the memorization effect) [7]. For example, mislabeled instances could be those that the model incorrectly classifies [170, 82, 164, 26], have a high loss value [194], or significantly increase the complexity of the model [46]. Given the identified mislabeled instances, these approaches either ignore them during training [201] or treat them as ‘unlabeled’ and apply techniques from semi-supervised learning (e.g., DivideMix [99], SELF [120]).

For binary classification, some approaches ‘correct’ (i.e., switch) the observed label for such instances [53, 203]. Building on this idea, others use a transition function to estimate the probability of the observed label being correct. Model predictions can then be adjusted by applying the transition function to the classifier’s predictions for each class. Some manually construct the transition function from expert knowledge [129], while others learn it [189, 191, 195, 187, 202, 15, 71, 10, 27, 100]. However, such approaches often make assumptions on the form of the noise distribution, and results can be sensitive to the choice of distribution [92].

To date, much of the work described above assumes instance-independent label noise (i.e., mislabeling is independent of the features). However, when this assumption is violated, the model may overfit to the noise [107]. From an emerging body of work in instance-dependent label noise [29, 187, 177, 205], current approaches remain limited in that they still rely on filtering heuristics. This introduces the limitation that disproportionately more instances

from subsets of the data where the noise rate is higher can be ignored during training.

### 4.2.2 Noise-Robust Loss Functions

Prior work also examines how regularization techniques can be adapted to the noisy labels setting, addressing issues related to overfitting on noisy data [116, 107, 38]. Label smoothing, and in some cases negative label smoothing, were found to improve accuracy on both correctly labeled and mislabeled data [107, 178]. With this approach, the observed labels are perturbed by a small, pre-determined value, with all labels receiving the same perturbation at every training epoch. Follow-up work found that, instead of applying the same perturbation at each epoch, adding a small amount of Gaussian stochastic label noise (SLN) at each epoch resulted in further improvements, as it helped to escape from local optima [25]. However, these approaches were most beneficial in the context of augmenting existing methods that identify mislabeled instances (e.g., stochastic label noise is applied to instances that are identified as correctly labeled by filtering approaches), and thus, potentially suffer from the same limitations. Alternatively, recent work has also proposed perturbing the features instead of the labels to encourage consistency in the model’s predictions [38], though mainly in the context of instance-independent label noise. Others have proposed noise-robust variations of cross entropy loss [41, 173] but generally relied on assumptions like the memorization effect.

### 4.2.3 Label Noise and Bias

More recently, the issue of bias and overall performance among subsets of the data in the presence of label noise has been addressed within the fairness literature. For example, when the frequencies at which groups (usually defined by a sensitive attribute) appear are different within a dataset, past work has shown that common approaches for addressing label noise can increase the prediction error for minority groups (i.e., rarer groups) [105]. To reduce bias in model performance resulting from label noise, past work proposed to re-weight instances from groups during training where model performance is poorer [68] in the instance-independent setting. Others have used peer loss [106] within groups [174] in cases where the noise depends only on the sensitive attribute. Recently, [184] examined the more general instance-dependent case in which the noise can depend on features other than the sensitive attribute. However, their approach required instances features to be grouped into distinct categories, such as root features and low-level features, which may not be well defined or available in all settings. Here, root features refer to features that are known not to be affected by group membership, and low-level features are those that are known to be affected by the root features, group membership, as well as any unobserved latent features.

#### 4.2.4 Anchor Points for Addressing Label Noise

In some noisy labels settings, a small subset of data with both observed and ground truth labels is available. For example, consider the example of automated EHR labeling. In order to evaluate the performance of a labeling tool, one must compare the labels from the automated labeling tool to a subset of data with ground truth labels. In these settings, such a subset is usually called an anchor set, and instances whose ground truth and observed labels are known are called anchor points. Anchor points are generally used to learn a transition function [188, 187, 15] or for label correction directly [185].

#### 4.2.5 How We Differ

We build from previous work in label noise and address key limitations. Generally, many state-of-the-art approaches [160] are limited in that they do not consider instance-dependent noise, do not consider the potential consequences of bias in label noise, or do not leverage the information our setting provides. We tackle these limitations by accounting for differences in noise rates among subsets of the data and taking advantage of additional information that can be found in our setting.

### 4.3 Methods

We introduce a two-stage approach for learning with instance-dependent label noise that leverages a small set of alignment points for which we have both observed and ground truth labels. Here, we define our notation and approach.

| Notation                                                       | Description                            |
|----------------------------------------------------------------|----------------------------------------|
| $g$                                                            | number of groups                       |
| $\hat{y} \in [0, 1]$                                           | predicted class probabilities          |
| $\tilde{y} \in \{-1, 1\}$                                      | observed label                         |
| $y \in \{-1, 1\}$                                              | ground truth label                     |
| $\hat{\beta} = P(y = \tilde{y}   \tilde{y}, \mathbf{x}; \phi)$ | prediction of label correctness        |
| $A$                                                            | alignment set, has $a$ instances       |
| $\bar{A}$                                                      | non-alignment set, $\bar{a}$ instances |
| $\theta$                                                       | main model parameters                  |
| $\phi$                                                         | auxiliary model parameters             |

Table 4.1: Notation. A summary of notation used throughout. Superscripts in parentheses specify instances (e.g.,  $\mathbf{x}^{(i)}$ ). Subscripts specify indexes into a vector (e.g.,  $\mathbf{x}_i$ )

### 4.3.1 Notation and Setting

Our notation is summarized in **Table 4.1**, with additional notation defined throughout as needed. Our dataset,  $D = A \cup \bar{A}$  consists of instances in  $A = \{\mathbf{x}^{(j)}, \tilde{y}^{(j)}, y^{(j)}\}_{j=1}^a$  and  $\bar{A} = \{\mathbf{x}^{(i)}, \tilde{y}^{(i)}\}_{i=1}^{\bar{a}}$ .  $A$  is the set of alignment points (i.e., the alignment set), where both  $\tilde{y}^{(j)}$  and  $y^{(j)}$  are known, and we assume that it includes instances where  $\tilde{y}^{(i)} \neq y^{(i)}$ . Alignment points are a special case of anchor points [104], where points that do *and* do not have matching observed and ground truth labels are both required.  $\bar{A}$  is the non-alignment set and contains instances for which we do not know the ground truth labels. In the presence of noisy labels, we assume that whether  $\tilde{y} == y$  is dependent on  $\mathbf{x}$  (i.e.,  $P(\tilde{y} == y) \neq P(\tilde{y} == y|\mathbf{x})$ ). Given this dataset, we aim to train a model to learn  $f : \mathbb{R}^d \rightarrow [0, 1]$  (i.e. the function used to predict the ground truth labels), so that we can map unseen instances into one of two classes based on their feature vectors. Our learned model parameters,  $\theta$ , are such that the output of the corresponding model represents the predicted class probabilities, (i.e.,  $\hat{y}$ ). Although we focus on binary classification, our setup can be applied to multiclass classification.

### 4.3.2 Justification and Desired Properties

Our setting is inspired by the use of pragmatic labeling tools in healthcare. Such tools are often based on various components of the EHR, and they are applied to identify cohorts or outcomes of interest [168, 121, 167, 197, 66]. However, while practical, such definitions are not always reflective of the ground truth, and thus, require validation through manual chart review. This is often done on a randomly chosen subset of individuals, which can be constructed to represent the target population and account for known heterogeneity. As a result,  $f$  is the function that predicts whether the condition is actually present, and the alignment set is the chart reviewed subset used to help learn  $f$ .

Through our approach, we aim to achieve: 1) robustness to the overall noise rate and 2) robustness to differences in noise rates between groups (i.e., the noise disparity). Revisiting our motivating example with EHR-based labeling tools, previous work has shown that labeling tools for rarer conditions such as drug-induced liver injury and dementia are more likely to be less reliable than those for common conditions [85]. Similar to how different noise rates can arise in practice, differences in noise rates between subgroups can also vary in practice [87]. As a result, achieving these properties can potentially make our approach generalize to a wide variety of settings.

### 4.3.3 Proposed Approach

Here, we describe the proposed network and training procedure.

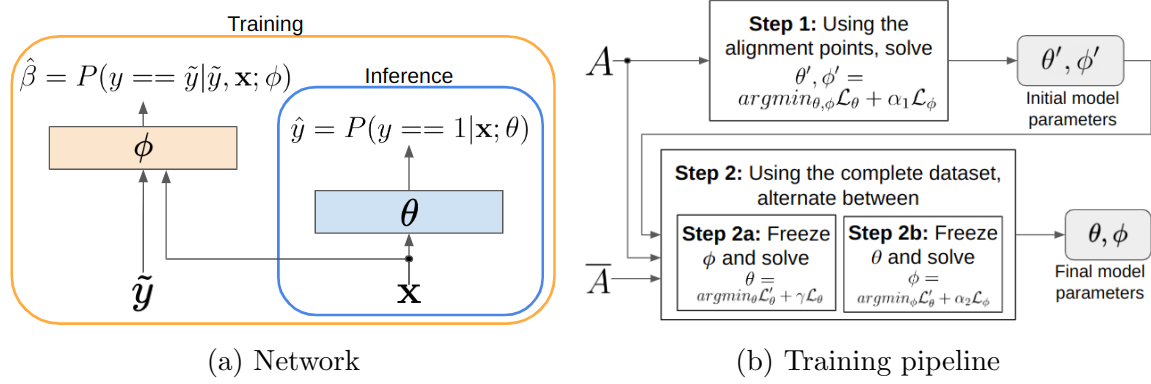


Figure 4.1: Our Approach. a) The model predicts,  $\hat{y}$ , at training and inference time using  $\theta$ . At training time, it also predicts whether the observed label is correct using  $\phi$ .  $\theta$  and  $\phi$  are pre-trained using  $A$  and then fine tuned with the complete dataset. b) We pretrain the model using the alignment points, then train on the noisy data.  $\mathcal{L}_\theta$ ,  $\mathcal{L}_\phi$ , and  $\mathcal{L}'_\theta$  are the objectives for the alignment points, label confidence score ( $\hat{\beta}_\phi$ ), and noisy data, respectively.  $\alpha_1$ ,  $\alpha_2$ ,  $\gamma$  are scalar hyperparameters.

**Proposed Network.** Our proposed network (**Figure 4.1a**) consists of two components. The first, parameterized by  $\theta$ , is a feed-forward network that uses feature vector  $\mathbf{x}$  to predict the class probability,  $\hat{y} = P(y = 1 | \mathbf{x}; \theta)$ . The second component, parameterized by  $\phi$ , is an auxiliary feed-forward network that uses observed label  $\tilde{y}$  and features  $\mathbf{x}$  to compute  $\hat{\beta} = P(y == \tilde{y} | \tilde{y}, \mathbf{x}; \phi)$ , an instance-dependent prediction for whether the observed label is correct based on  $\mathbf{x}$  and  $\tilde{y}$ .  $\hat{\beta}$  can be considered as a confidence score for the observed label, with higher values indicating higher confidence. Learning  $\hat{\beta}$  models the underlying pattern of label noise by forcing the model to learn which instances are correctly labeled. We use  $\hat{\beta}$  to reweight the objective function during the second step of training, as described below. By including the observed label as input to  $\phi$ , our approach also applies to instance-independent label noise because it accounts for the case when the underlying pattern of label noise does not depend on the features. In order to learn  $\hat{\beta}$ , we assume that the label noise pattern can be represented as some function, though the specific form of this function (e.g., linear) does not need to be known. During training, we compute the loss using the outputs from both networks. At inference time (i.e., in practical use after training), we compute the class predictions from the network parameterized by  $\theta$  only since  $\tilde{y}$  is unavailable.

**Training Procedure.** Our training procedure is summarized in **Figure 4.1b** and **Algorithm 2**. In Step 1, we pre-train both networks using the alignment points,  $A$ , minimizing

---

**Algorithm 2** Proposed approach. Here, ‘stopping criteria’ may refer to any stopping criteria, such as early stopping. The Freeze() function takes as input model parameters and freezes them, and Unfreeze() takes as input model parameters and unfreezes them.

---

**Input:**  $\{\mathbf{x}^{(i)}, \tilde{y}^{(i)}, y^{(i)}\}_{i \in A}, \{\mathbf{x}^{(i)}, \tilde{y}^{(i)}\}_{i \notin A}, \theta', \phi'$

**Output:**  $\hat{\mathbf{p}}_k$ , estimated distribution of occurrence at the original time scale

**Hyperparameters:** Scalars  $\alpha_1, \alpha_2, \gamma$

Train  $(\{\mathbf{x}^{(i)}, \tilde{y}^{(i)}, y^{(i)}\}_{i \in A}, \{\mathbf{x}^{(i)}, \tilde{y}^{(i)}\}_{i \notin A}, \theta', \phi')$

1. While  $\neg(\text{stopping criteria})$  (Step 1)

(a)  $\hat{y} = \theta'(\mathbf{x})$  (Predict label)

(b)  $\hat{\beta}_\phi = \phi'(\mathbf{x}, \tilde{y})$  (Predict label confidence)

(c)  $\mathcal{L}_\theta = \frac{-1}{|A|} \sum_{i \in A} \sum_{j=1}^c \mathbb{I}(y^{(i)} == j) \log(\hat{y}_j^{(i)})$

(d)  $\mathcal{L}_\phi = \frac{-1}{|A|} \sum_{i \in A} \mathbb{I}(\tilde{y}^{(i)} == y^{(i)}) \log(\hat{\beta}_\phi^{(i)}) + \mathbb{I}(\tilde{y}^{(i)} \neq y^{(i)}) \log(1 - \hat{\beta}_\phi^{(i)})$

(e) Loss =  $\mathcal{L}_\theta + \alpha_1 \mathcal{L}_\phi$

(f) Update model parameters and then compute stopping criteria

2.  $\theta, \phi \leftarrow \theta', \phi'$  then Freeze( $\phi$ )

3. While  $\neg(\text{stopping criteria})$  (Step 2)

(a)  $\hat{y} = \theta'(\mathbf{x})$  and  $\hat{\beta}_\phi = \phi'(\mathbf{x}, \tilde{y})$

(b)  $\mathcal{L}_\theta = \frac{-1}{|A|} \sum_{i \in A} \sum_{j=1}^c \mathbb{I}(y^{(i)} == j) \log(\hat{y}_j^{(i)})$

(c)  $\mathcal{L}_\phi = \frac{-1}{|A|} \sum_{i \in A} \mathbb{I}(\tilde{y}^{(i)} == y^{(i)}) \log(\hat{\beta}_\phi^{(i)}) + \mathbb{I}(\tilde{y}^{(i)} \neq y^{(i)}) \log(1 - \hat{\beta}_\phi^{(i)})$

(d)  $\mathcal{L}'_\theta = \frac{-1}{|A|} \sum_{k=1}^G \frac{1}{1-\hat{r}_k} \sum_{i \in \bar{A} \cap g_k} \sum_{j=1}^c \hat{\beta}_\phi^{(i)} \mathbb{I}(\tilde{y}^{(i)} == j) \log(\hat{y}_j^{(i)})$  (Weighted loss)

(e) If  $\phi$  is frozen (Step 2a)

i. Loss =  $\mathcal{L}_{\theta'} + \gamma \mathcal{L}_\theta$

ii. Unfreeze( $\phi$ ) then Freeze( $\theta$ )

(f) Else (Step 2b)

i. Loss =  $\mathcal{L}_{\theta'} + \alpha_2 \mathcal{L}_\phi$

ii. Unfreeze( $\theta$ ) then Freeze( $\phi$ )

(g) Update model parameters and then compute stopping criteria

4. Return  $\theta, \phi$  (Final model parameters)

---

an objective function based on cross entropy:  $\theta', \phi' = \operatorname{argmin}_{\theta, \phi} \mathcal{L}_\theta + \alpha_1 \mathcal{L}_\phi$ .  $\alpha_1 \in \mathbb{R}^+$  is a scalar hyperparameter;  $\theta'$  and  $\phi'$  are parameters that represent the initial values of  $\theta$  and  $\phi$ .  $\mathcal{L}_\theta$  is the cross-entropy loss between the class predictions and ground truth labels. It aids in learning the parameter values for  $\theta$ , and thus, the model's decision boundary.  $\mathbb{I}$  is an indicator function.

$$\mathcal{L}_\theta = \frac{-1}{|A|} \sum_{j \in A} \mathbb{I}(y^{(j)} == 1) \log(\hat{y}^{(j)}) + \mathbb{I}(y^{(j)} == -1) \log(1 - \hat{y}^{(j)})$$

$\mathcal{L}_\phi$  is the cross-entropy loss between the predicted confidence score  $\hat{\beta}^{(j)}$  and the actual agreement between  $\tilde{y}^{(j)}$  and  $y^{(j)}$ . It aids in learning the weights for  $\phi$ , and thus, the underlying label noise pattern.

$$\mathcal{L}_\phi = \frac{-1}{|A|} \sum_{j \in A} \mathbb{I}(\tilde{y}^{(j)} == y^{(j)}) \log(\hat{\beta}^{(j)}) + \mathbb{I}(\tilde{y}^{(j)} \neq y^{(j)}) \log(1 - \hat{\beta}^{(j)})$$

In Step 2, we initialize  $\theta$  and  $\phi$  as  $\theta'$  and  $\phi'$  and fine tune using the complete dataset. Step 2 consists of two parts, Step 2a and Step 2b. Each part aims to improve a specific component of the network (e.g.,  $\theta$ ) using another component of the network (e.g.,  $\phi$ ). We begin with Step 2a, move to Step 2b, and continue to alternate between Step 2a and Step 2b in a manner similar to expectation maximization so that we continually improve both  $\theta$  and  $\phi$ . In Step 2a, we freeze  $\phi$  and find  $\theta$  that minimizes the objective  $\mathcal{L}'_\theta + \gamma \mathcal{L}_\theta$ .  $\gamma \in \mathbb{R}^+$  is a scalar hyperparameter. In Step 2b, we freeze  $\theta$  and find  $\phi$  that minimizes the objective  $\mathcal{L}'_\theta + \alpha_2 \mathcal{L}_\phi$ .  $\alpha_2 \in \mathbb{R}^+$  is a scalar hyperparameter.  $\mathcal{L}'_\theta$  computes the cross-entropy loss over the potentially noisy, non-alignment points. Each instance is weighted by the model's confidence in whether the observed label is correct via  $\hat{\beta}^{(i)}$ , taking advantage of the model's learned noise pattern. Our approach aims to mitigate bias by up-weighting groups,  $k = 1, 2, \dots, g$  with a higher estimated noise rate,  $\hat{r}_k$ , so that they are not dominated by/ignored compared to groups with a lower estimated noise rate.

$$\mathcal{L}'_\theta = \frac{-1}{|A|} \sum_{k=1}^g \frac{1}{1 - \hat{r}_k} \sum_{i \in \bar{A} \cap G_k} \sum_{j \in \{-1, 1\}} \hat{\beta}_\phi^{(i)} \mathbb{I}(\tilde{y}^{(i)} == j) \log(\hat{y}_j^{(i)})$$

We calculate  $1 - \hat{r}_k$  as follows. We introduce sets  $G_k$  for  $k = 1, 2, \dots, g$  to represent disjoint subgroups of interest in the data, which are assumed to be known in advance.  $G_a \cap G_b = \emptyset$  for all  $a = 1, 2, \dots, g$ ,  $b = 1, 2, \dots, g$  with  $a \neq b$  and  $\cup_{k=1}^g G_k = D$ . Each group  $G_k$  is then associated with estimated noise rate  $\hat{r}_k = \frac{1}{|G_k|} \sum_{i \in G_k} 1 - \hat{\beta}^{(i)}$ . Although weighting each instance by  $\hat{\beta}$  is a form of soft filtering, weighting each group by the inverse of its overall

‘clean’ rate avoids the effect of de-emphasizing groups with higher predicted noise rates. As a result, the expected value of  $\mathcal{L}'_\theta$  with respect to  $\hat{\beta}$  is equal to the cross-entropy loss between the model’s predictions and ground truth labels (shown in **Appendix B**). However, this assumes accurate estimates of  $\hat{\beta}$ . Thus, we expect that the proposed approach will perform best when the alignment set is representative of the target population. In scenarios where the alignment set is biased (e.g., some groups are underrepresented), if the learned noise function does not transfer to the underrepresented group, then the proposed approach may not be beneficial. In our experiments, we test this.

During Step 2a,  $\mathcal{L}'_\theta$  is used to train  $\theta$  by learning to predict  $\hat{y}$  such that it matches observed label  $\tilde{y}$  on instances that are predicted to be correctly labeled. During Step 2b,  $\mathcal{L}'_\theta$  is used to train  $\phi$ . Here, since  $\theta$  is frozen and  $\phi$  is not, the network learns to predict the optimal  $\hat{\beta}$ . Based on  $\mathcal{L}'_\theta$  alone, there are two possible options to learn  $\hat{\beta}$ : 1) consistently make  $\hat{\beta}$  close to 0, and 2) predict  $\hat{\beta}$  such that it is close to 1 when  $\hat{y}$  matches  $\tilde{y}$  and close to 0 when  $\hat{y}$  does not match  $\tilde{y}$ . Since  $\tilde{y}$  is used as a proxy for  $y$  in this step, the second option aligns with what we want  $\hat{\beta}$  to represent. To encourage this over the first option (i.e., consistently predicting 0 for  $\hat{\beta}$ ), we include  $\mathcal{L}_\phi$  in Step 2b, which is *not* minimized by consistently predicting 0 for  $\hat{\beta}$ . Note that, in Step 2b, we rely on the cluster assumption [157] from semi-supervised learning, which broadly states that labeled data fall into clusters and that unlabeled data aid in defining these clusters. In the context of Step 2b, ‘labeled’ and ‘unlabeled’ are analogous to whether we know if the ground truth and observed labels match (i.e., alignment point versus non-alignment point), rather than the actual class labels. As a result, we also rely on the alignment set being representative of the target population here to avoid dataset shift.

In contrast to previous filtering approaches, our approach utilizes all data during training. Moreover, it does not require a specialized architecture beyond the auxiliary network to compute  $\hat{\beta}$ . Thus, it can be used to augment existing architectures.

## 4.4 Experimental Setup

We empirically explore the performance of our proposed approach relative to state-of-the-art baselines on five benchmark prediction tasks with two different label noise settings. We aim to test 1) the extent to which our desired properties hold, 2) the extent to which the proposed approach is robust to changes in the composition of the alignment set, and 3) which components of the proposed approach contribute the most.

### 4.4.1 Datasets

We consider five different binary prediction tasks on four datasets from several domains with synthetic and real datasets. Though inspired by healthcare, we also consider domains outside of healthcare to show the broader applicability of our approach in areas where harmful biases can arise (e.g., predicting recidivism and income). Throughout our experiments, we start by assuming the labels in the dataset are noise free, and we inject varying amounts of synthetic label noise. In this subsection, we describe the tasks, features, and ‘ground truth’ labels we use. The next subsection will describe how we introduce synthetic label noise.

#### 4.4.1.1 Synthetic

We generate a dataset containing 5,000 instances according to the generative process shown below. The positive rates for the majority and minority groups are 37.5% and 32.3%, respectively. Note that the  $Percentile(p, \{z\})$  function outputs the  $p^{th}$  percentile over all values in  $\{z\}$ . We defined the feature at index 0 to be a synthetic sensitive attribute. Instances with values below the 20<sup>th</sup> percentile for this feature were considered as the ‘minority’, and the rest were considered as the ‘majority’. Features 10-19 for the majority instances and features 20-29 for the minority instances were set to 0 to provide more contrast between the two groups. For individual  $i$ ,

$$\begin{aligned} d &= 30, \mathbf{x}^{(i)} \sim N(0, 1)^{30} \quad \mathbf{w} \sim N(0, 1)^{30}, z^{(i)} = \mathbf{x}^{(i)} \cdot \mathbf{w} \\ y^{(i)} &= 1 \text{ if } z^{(i)} > Percentile(50, \{z^{(j)}\}_{j=1}^{5000}) \text{ else } 0 \\ x_j^{(i)} &= 0 \text{ for } j = 10, 11, \dots, 19 \text{ if } x_0^{(i)} > Percentile(20, \{x_0^{(j)}\}_{j=1}^{5000}) \\ x_j^{(i)} &= 0 \text{ for } j = 20, 21, \dots, 29 \text{ if } x_0^{(i)} < Percentile(20, \{x_0^{(j)}\}_{j=1}^{5000}) \end{aligned}$$

#### 4.4.1.2 MIMIC-III

Within the healthcare domain, we leverage a publicly available dataset of electronic health record data [75]. We consider two separate prediction tasks: onset of 1) acute respiratory failure (ARF) and 2) shock in the ICU [124] using the four hour prediction setup for both tasks as described by [162], resulting in 15,873 and 19,342 ICU encounters, respectively. We used the same features as [162] for our tasks, where more information can be found at <https://physionet.org/content/mimic-eicu-fiddle-feature/1.0.0/>. Each encounter had 16,278 and 18,186 features for each task respectively. We consider groups based on race, with about 70% of patients being white (positive rate 4.5% [ARF], 4.1% [shock]) and 30% being non-white (positive rate 4.4% [ARF], 3.7% [shock]).

#### 4.4.1.3 Adult

Outside of healthcare, we leverage a publicly available dataset of census data [36]. We consider the task of predicting whether an individual’s income is over \$50,000. This dataset includes data pertaining to age, education, work type, work sector, race, sex, marital status, and country. Its training and test sets contain 32,561 and 16,281 individuals, respectively. We use a pre-processed version of this dataset<sup>1</sup> and randomly select 1,000 individuals out of 32,561 for training. We also only include features pertaining to age, education, work type, marital status, work sector, and sex to make the task more difficult (see below). After preprocessing, each individual was associated with 56 features, and all features had a range of 0-1. We consider groups based on sex, with 67.5% of individuals being male (positive rate 30.9%) and 32.5% being female (positive rate 11.3%).

**Preprocessing.** Although, we used a pre-processed version of this dataset, we omitted features pertaining to education, work type, and work sector to make the task more difficult. More specifically, in the file ‘headers.txt’ at the repository mentioned in Footnote 1, we kept all features beginning with ‘age’, ‘workclass’, ‘education’, ‘marital status’, and ‘occupation’. We also kept the ‘Sex.Female’ feature. The remaining features were excluded to make the task more difficult. Values were normalized for each feature to have a range of 0-1 by subtracting by the minimum value observed among all individuals and dividing by the range. During training, we only used 1,000 randomly selected individuals from the provided dataset to make the task more difficult, since there would be fewer samples from which to learn. We made the task more difficult for this dataset to further highlight the differences in performance between the approaches.

#### 4.4.1.4 COMPAS

Like the Adult dataset, we leverage publicly available dataset outside of healthcare collected by ProPublica from Broward County, Florida, USA [4]. We consider the task of predicting recidivism within two years, i.e., whether a criminal defendant is likely to re-offend. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) includes data pertaining to age, race, sex, and criminal history. We use a pre-processed version of this dataset<sup>2</sup> and also normalize each feature to have a range of 0-1 (see below). After preprocessing, the dataset included 6,172 individuals with 11 features per individual. We consider groups based on race, with 65.8% of individuals being white (positive rate 39.1%) and 34.2% being non-white (positive rate 44.5%).

---

<sup>1</sup><https://github.com/AissatouPaye/Fairness-in-Classification-and-Representation-Learning>

<sup>2</sup><https://www.kaggle.com/danofer/compass>

**Preprocessing.** Although, we used a pre-processed version of this dataset, we omitted the feature ‘score\_factor’ (i.e., the risk score for recidivism from the ProPublica model) to make the task more difficult. Values were normalized for each feature to have a range of 0-1 by subtracting by the minimum value observed among all individuals and dividing by the range of values among all individuals.

#### 4.4.2 Adding Label Noise

To test the robustness of our approach in different settings of label noise, we introduce synthetic instance-dependent label noise to our datasets. Like past work [160], our setup is limited for the real datasets because our added noise is synthetic and we use the labels provided in the dataset as ground truth, since we do not have access to actual ground truth labels on these public datasets.

To introduce instance-dependent noise, mislabeling was a function of the features. Let  $\mathbf{w}_m \sim N(0, 0.33)^D$  and  $z_m = \sigma(\mathbf{x} \cdot \mathbf{w}_m)$ , where  $\sigma$  is the sigmoid function, denote the coefficients describing the contribution of each feature to mislabeling and the risk of mislabeling, respectively. Whether an instance was mislabeled was based on  $z_m$  and the desired noise rate. For example, for a noise rate of 30%, instances whose value for  $z_m$  was above the 70<sup>th</sup> percentile had their labels flipped. This allowed us to vary the noise rate within subgroups in a straightforward manner. Across datasets, we focused on cases where the noise rate in the ‘minority’ population was always greater than or equal to that of the ‘majority’ group since this is more likely to occur [161].

#### 4.4.3 Evaluation Metrics

We evaluate our proposed approach in terms of discriminative performance and model bias. For discriminative performance, we evaluate using the AUROC (higher is better).

With respect to model bias, while there exist many different measures, we focus on EO [55], since it is commonly used in the context of healthcare [132, 190, 196], when similar performance across groups is desired [141, 133]. Because EO focuses on the difference between the true and false positive rates among groups, it is applicable to many settings in healthcare since the consequences of failing to treat a patient in need [135, 18], or giving an inappropriate treatment [17, 118] can be serious. More specifically, we measure the AUEOC [34] (higher is better). For classification threshold  $\tau$ , we calculate the EO (denoted  $\text{EO}(\tau)$ ) between the groups. The AUEOC is obtained by plotting  $\text{EO}(\tau)$  against all possible values of  $\tau$  and calculating the area under the curve.

We compute the harmonic mean (HM) between the AUROC and AUEOC to highlight how

the different approaches simultaneously maintain discriminative performance and mitigate bias. In the HM the worse performing metric dominates. For example, if a classifier has AUROC=0.5 and AUEOC=1.0, the HM will emphasize the poor discriminative performance.

#### 4.4.4 Baselines

We evaluate our proposed approach with several baselines to test different hypotheses.

**Standard** does not account for label noise and assumes that  $\tilde{y} == y$  is always true.

**SLN + Filter** [25] combines filtering [7] and SLN [25] and was shown to outperform state-of-the-art approaches like Co-Teaching [54] and DivideMix [99]. It relies on filtering heuristics, which indirectly rely on uniform random label noise to maintain discriminative performance and mitigate bias.

**JS (Jensen-Shannon) Loss** [38] builds on semi-supervised learning and encourages model consistency when predicting on perturbations of the input features. It was shown to be competitive with other state-of-the-art noise-robust loss functions [108]. It was proposed for instance-independent label noise.

**Transition** [187] learns to correct for noisy labels by learning a transition function and was shown to outperform state-of-the-art approaches such as MentorNet [70]. It applies to instance-dependent label noise, but it assumes that the contributions of each feature to mislabeling and input reconstruction are identical.

**CSIDN** (confidence-scored instance-dependent noise) [15] also learns a transition function and was shown to outperform state-of-the-art approaches such as forward correction [129]. Like our approach, CSIDN uses the concept of ‘confidence’ in the observed label to help with training. Unlike our approach, CSIDN uses the model’s class predictions directly as confidence scores (instead predicting them via an auxiliary network) and uses them to learn the transition function (as opposed to re-weighting the loss).

**Fair GPL** [174] builds on work addressing uniform random label noise [68] and uses peer loss (i.e., data augmentation that reduces the correlation between the observed label and model’s predictions) within subgroups [174]. It assumes that label noise only depends on group membership.

We also train a model using the ground truth labels (called **Clean Labels**) as an empirical upper bound for discriminative performance.

#### 4.4.5 Implementation Details

For each dataset, we randomly split the data into 80/20% training/test, ensuring that data from the same individual did not appear across splits. For the Adult dataset, we used

the test set provided and randomly selected 1,000 individuals from the training set. We then randomly selected 10% of the training data for all datasets except MIMIC-III from each subgroup to be alignment points, thereby ensuring that they were representative of the overall population. For MIMIC-III, 2% from each subgroup were selected as alignment points due to the larger size of the dataset. Alignment points were selected randomly to simulate our setting of focus, where we have a proxy labeling function and then randomly select a subset of the data to chart review in order to evaluate the proxy function. Then, for all datasets, half of the alignment points were then set aside as a validation set to use during training for early stopping and hyperparameter selection, while the other half remained in the training set. Later, in our experiments, we evaluated when the alignment set size varied and when the alignment set was biased. All approaches (i.e., baselines and proposed) were given the ground truth labels for data in the alignment set (i.e., no noise added to alignment points) during training so that some approaches did not have an unfair advantage.

All models were trained in Python3.7 and Pytorch1.7.1 [128], using Adam [84]. Hyperparameters, including the learning rate, L2 regularization constant, and objective function scalars (e.g.,  $\alpha$ ), were tuned using random search, with a budget of 20. We used early stopping (patience=10) based on validation set performance, which we measured with the HM. We report results on the held-out test set, showing the mean and standard deviation over 10 replications. More details about the implementation are here: <https://github.com/MLD3/Instance-Dependent-Label-Noise>.

## 4.5 Results and Discussion

We describe the results from experiments with instance-dependent noise. For each plot, we show 1) the AUROC, 2) AUEOC, and 3) the HM of the AUROC and AUEOC to assess general performance with respect to both metrics.

### 4.5.1 Robustness to Noise Rate

Here, we investigated how robust the proposed approach and baselines were to varying amounts of instance-dependent label noise (**Figures 4.2, 4.3**). Since noise was synthetically introduced and not dataset specific, we conducted two experiments on the synthetic dataset. In the first, we varied the overall noise rate from 10-60% in the majority group. For the minority group, we considered noise rates that were consistently 20% higher than that of the majority group, to keep the noise disparity level (i.e., the difference in noise rates between subgroups) constant. In the second, we varied the minority noise rate from 20-90% with a

majority noise rate fixed at 20% throughout (i.e., from 0-70% disparity).

**Part 1: Overall Noise Rate.** Overall, our proposed approach demonstrated robustness to a variety of noise rates within a realistic range (**Figure 4.2**). At low minority noise rates (i.e., below 40%), the proposed approach and baselines, with the exception of JS Loss, were competitive. As the noise rate increased, many of the baselines experienced noticeable degradation in performance. The proposed approach and Transition showed more robustness, with the proposed approach being the most robust until a minority noise rate of 80%, which represents an extreme case of label noise. More specifically, with respect to discriminative performance, the proposed approach is able to consistently outperform the baselines until a minority noise rate of 80%. This observation is similar to what we observed with the HM. With respect to bias mitigation, the proposed approach is consistent the baselines up to a minority noise rate of 60%. At a minority noise rate above 60%, our approach experienced the least degradation compared to the baseline approaches. This is in line with our expectations since our approach explicitly accounts for differences in noise rates among groups during training.

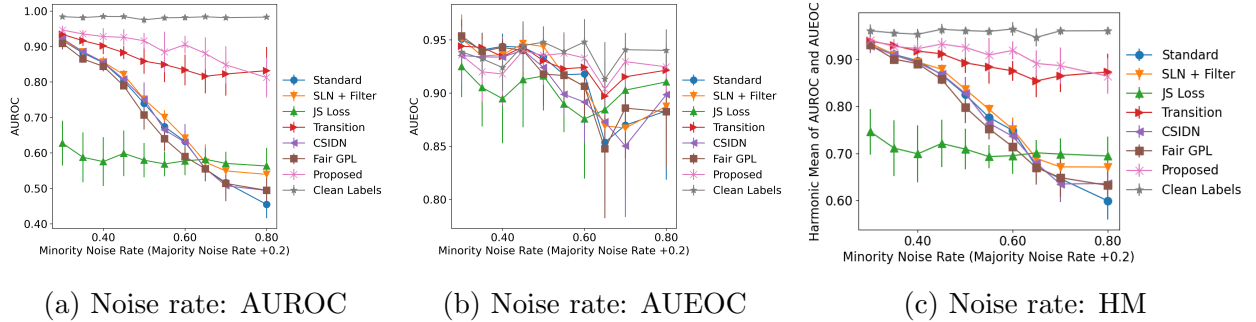


Figure 4.2: Robustness to noise rate in an instance dependent setting. We plot the mean and standard deviation for 10 random seeds. As the noise rate increases, the proposed approach generally shows the least degradation up to a minority noise rate of 80%.

**Part 2: Noise Disparity.** Like the previous experiment, the proposed approach was robust over a variety of noise disparities (**Figure 4.3**). This is likely because the objective function  $\mathcal{L}'_\theta$  from Step 2 of training accounts for disparities by scaling each instance-specific loss term with the reciprocal of its estimated group clean rate (i.e.,  $1 - \text{the estimated group noise rate}$ ). Similar to the previous experiment, at a minority noise rate of 80% and above, the proposed approach was no longer the most robust, though this setting is unlikely to occur in practice. More specifically, with respect to discriminative performance, we have similar observations to the previous experiment in that the proposed approach is able to consistently outperform the baselines until a minority noise rate of 80%. With respect to bias mitigation,

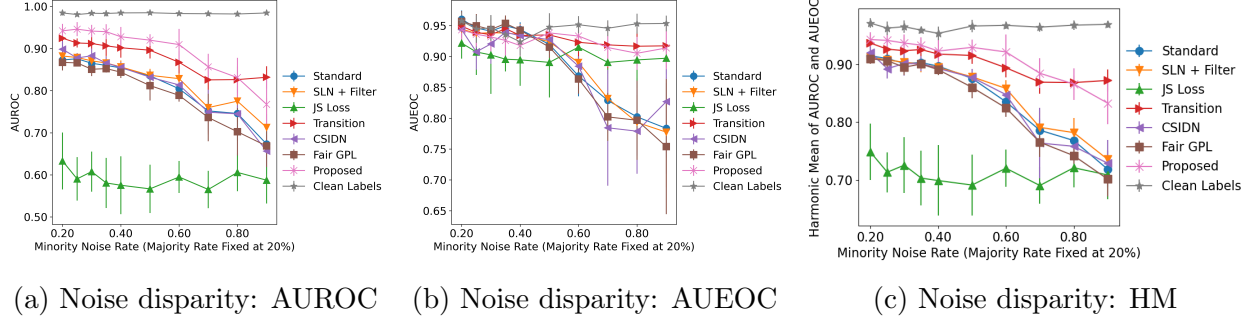


Figure 4.3: Robustness to noise disparity in an instance dependent setting. We plot the mean and standard deviation for 10 random seeds. As the noise disparity increases, the proposed approach generally shows the least degradation up to a minority noise rate of 80%.

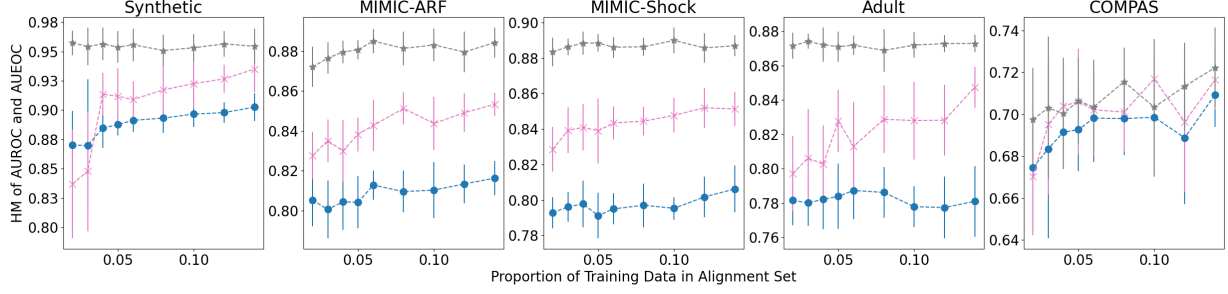
the proposed approach is comparable to the baselines up to a minority noise rate of 40%. At a minority noise rate above 40%, our approach experienced the least degradation compared to most of the other baseline approaches and was comparable to the Transition baseline.

#### 4.5.2 Sensitivity to Alignment Set Composition

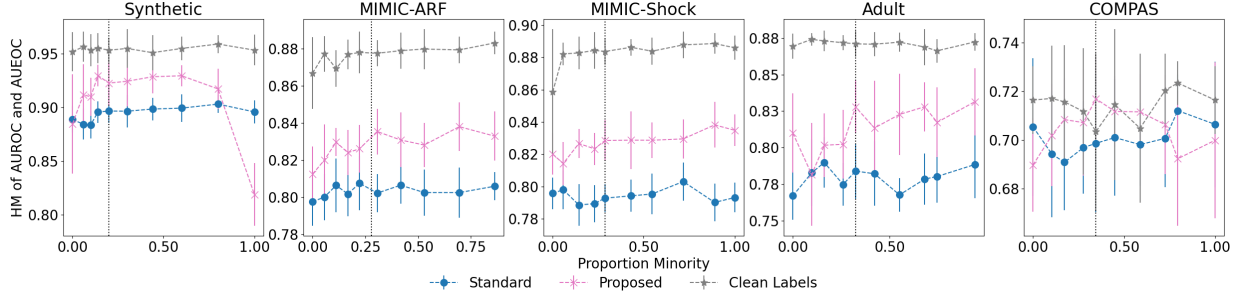
Our next set of experiments tested the proposed approach in settings where we relax key settings about the alignment set. We considered all datasets with instance-dependent noise. The majority/minority noise rates were 20%/40%, respectively. Here we show performance with respect to the proposed approach, Standard, and Clean Labels. Results showing the AUROC and AUEOC separately are in **Appendix B**.

Part 1: Alignment set size. We varied the size of the alignment set, from 1% and 15% of the training set, with the alignment set being representative of the test set (**Figure 4.4a**). The proposed approach was robust to a wide range of alignment set sizes, only showing noticeable degradation at alignment set sizes of 3% or lower. As the size of the alignment set grew, performance improved, likely since having a larger alignment set provided access to a larger set of ground truth labels at training time. Although the minimum number of points required in the alignment set is likely to vary depending on the task, our results are promising in that they show that our approach is effective on a variety of real life tasks, even when the alignment set is small (i.e., as little as 3% of the data).

Part 2: Biased alignment set. Here, we test how the proposed approach performs when the alignment set is not representative of the population. We varied the amount of bias in the alignment set by changing the proportion at which the subgroups were present. We kept the size of the alignment set constant at 10% of the training data (2% for MIMIC-III on both



(a) As we decrease the alignment set size (proportion of training data) performance decreases. Still, at an alignment set size of 3%, the proposed approach is robust.



(b) We varied the alignment set bias (minority proportion). The proposed approach is generally robust to variations in bias. The dashed vertical black line shows an unbiased alignment set.

Figure 4.4: Robustness to varying alignment set size and bias. Mean HM and standard deviation for 10 random seeds.

tasks). We observed that the proposed approach was robust over a wide range of conditions, i.e., when the minority proportion is 20%-80% (**Figure 4.4b**). We hypothesize that this is because the learned relationship between the features and noise can generalize across groups to an extent. In scenarios where performance of the proposed approach degraded, one subgroup heavily dominated the alignment set. This is shown in **Figure 4.4b** on the extremes of the x-axis of some datasets, which correspond to an alignment set that is heavily over-represented for one subgroup and heavily under-represented for the other. Our approach relies, in part, on having a relatively unbiased alignment set for estimating  $\hat{\beta}$  in order to avoid introducing dataset shift between the two steps of our training pipeline. Thus, these results are in line with our expectations and highlight a limitation of our approach. However, despite this reliance, we observe that our approach is still robust in some scenarios where the alignment set is biased.

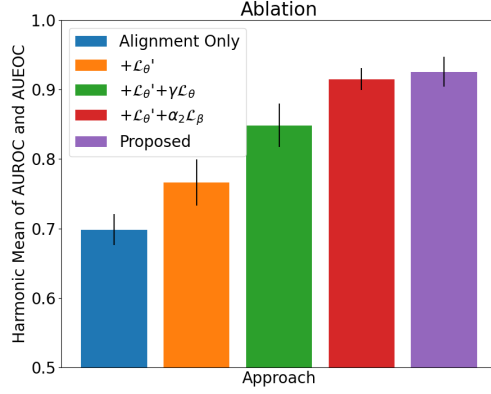


Figure 4.5: Ablation study of proposed approach.

### 4.5.3 Ablation Study and Hyperparameter Sensitivity

We examined our approach more closely by conducting an ablation study and a hyperparameter sensitivity analysis on the synthetic data. We used the synthetic dataset since our noise was synthetically introduced and not dataset specific. In our ablation study (**Figure 4.5**), we began with training on only the alignment points (i.e., Step 1 only), which achieved the worst performance. We then introduced Step 2 and added the remaining training data (i.e., non-alignment points) but only trained using  $\mathcal{L}_{\theta'}$ . This led to an improvement in performance, but not to the level of the full approach. The next two ablations build on the previous one. In the first one, we added continued supervision on the alignment points with  $\mathcal{L}_{\theta}$ , and observed an improvement in performance, likely due to the retention of high quality data in this step. In the second one, we added continued supervision on the alignment points using  $\mathcal{L}_{\phi}$ , and observed an even larger improvement. This is likely because including  $\mathcal{L}_{\phi}$  prevented the model from learning a solution where  $\hat{\beta}$  was small for all instances, as previously discussed. Finally, we end with our full proposed approach, which performed noticeably better than each of the ablations, showing the importance of each component.

In our hyperparameter sensitivity analysis on the synthetic data (**Figure 4.6**), we tested how performance of the (full) proposed approach varied to changes in the hyperparameters  $\alpha_1$ ,  $\alpha_2$ , and  $\gamma$ . For each of these hyperparameters, we measured performance at values between 0.01 and 100 on a logarithmic scale while keeping the other two values constant at 1. We found that  $\alpha_1$  and  $\gamma$  were the most robust to changes in the value. We found that  $\alpha_2$  was more sensitive, with values between 0.1 and 10 generally working best.

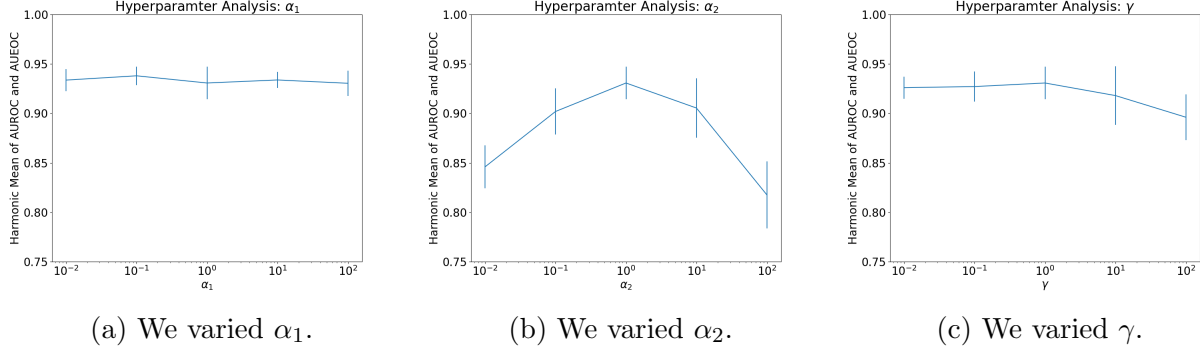


Figure 4.6: Sensitivity analysis of proposed approach to objective function hyperparameters.

## 4.6 Conclusion

We introduce a novel approach for learning with instance-dependent label noise. Our two-stage approach uses the complete dataset and learns the relationship between the features and label noise using a small set of alignment points. On several datasets, we show that the proposed approach leads to improvements over state-of-the-art baselines in maintaining discriminative performance and mitigating bias. Our approach is not without limitations. We empirically demonstrated that the success of the approach depends, in part, on the representativeness in the alignment set by evaluating performance when we changed the bias within the alignment set. Our experiments were also on pseudo-synthetic data in which we injected noise; this assumes we start from a noise free dataset. Finally, we only examined one form of bias in a specific case of instance-dependent label noise. Nonetheless, our case frequently arises in healthcare, especially when pragmatic (e.g., automated) labeling tools are used on large datasets, and chart review on the entire dataset is infeasible.

## CHAPTER 5

# Survival Analysis with Multiple Noisy Labels

### 5.1 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 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 can correspond to the time of disease onset. However, some TTEs are difficult to identify without manual chart review by a clinical expert (e.g., sepsis [154, 156, 101] and Alzheimer’s disease [167]), 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 [126, 121, 168, 197]. For example, one could label sepsis onset as 1) when the CDC (Center for Disease Control and Prevention) definition is met [145] or 2) when the Sepsis3 definition is met [153]. 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, one can potentially learn a more accurate survival model by combining noisy proxies from multiple different labelers. We refer to this setting as survival analysis with multiple noisy labels.

**Current Gaps.** In survival analysis, work addressing noisy labels is largely limited to the single labeler case [123]. Work studying multiple noisy labels focuses almost entirely on classification, where approaches generally aggregate the noisy labels from dataset at 1) preprocessing time [127, 142], or 2) inference time [33, 143, 98, 30]. 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). 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. In survival analysis with multiple noisy labels, where noisy labels correspond to noisy TTEs, one could 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) and do not necessarily produce accurate survival curves (**Figure 5.1**).

**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 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

**Organization.** The rest of this chapter is as follows. First, we present some background information on how previous work in 1) multiple noisy labels from classification, and 2) noisy labels in survival analysis. Then, we describe how we learn survival models from multiple noisy labels and reference labels. Finally, we describe experiments on synthetic and real data in which we evaluate our approach to adaptations of multiple noisy label approaches from the classification literature.

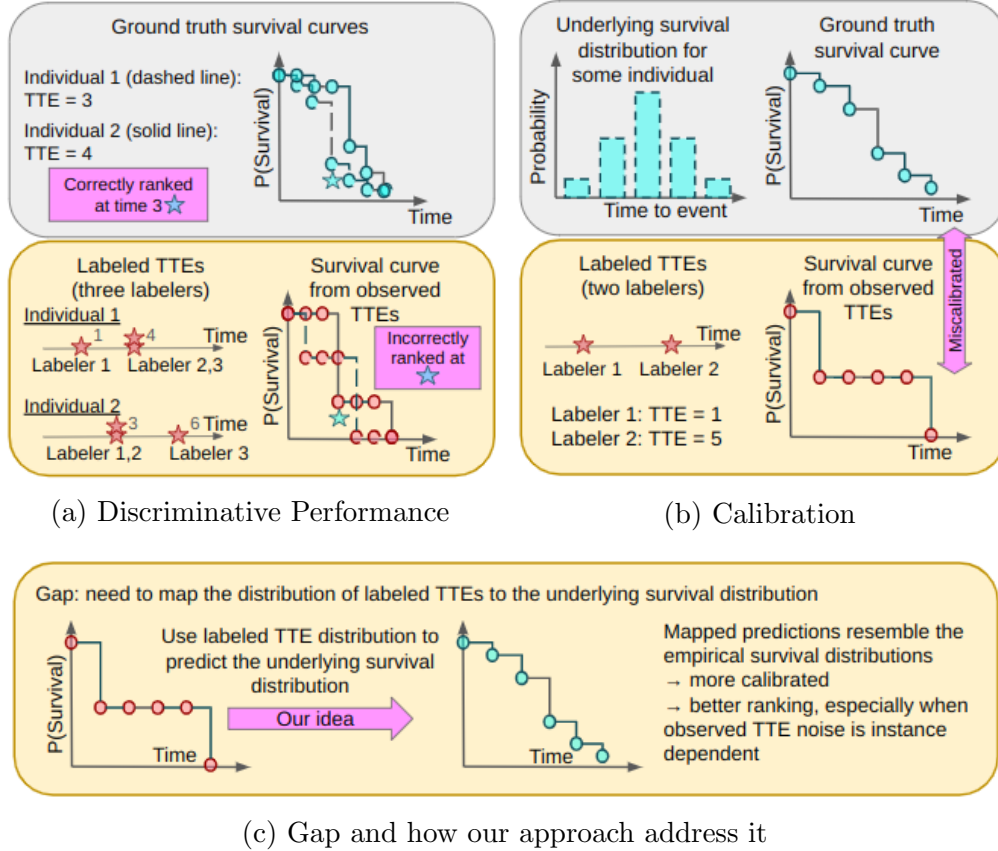


Figure 5.1: Comparison of distributions. We show examples of mismatches between the underlying survival distribution and distribution of noisy TTEs, and how these can hurt discriminative performance (a) and calibration (b). We then show how we address this (c).

## 5.2 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) the use of pseudo-labels, respectively.

### 5.2.1 Classification

Past work addressing multiple noisy labels tends to focus on the setting of multiple annotators from crowdsourcing [155] and mainly builds on the Dawid and Skene approach [33], which models each labeler individually by learning a confusion matrix for each labeler, and then aggregating at inference time. For example, follow up work has proposed to estimate specific parameters for individual labelers, such as sensitivity and specificity [143], and bias and variance [181]. Other lines of work propose to learn a mapping from the ground truth label

to the noisy label for each labeler [148, 62, 65] or to learn a model for each labeler separately based on the observed labels [50]. 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 [69]. This approach is most similar to ours in that we also model the distribution of observed labels. However, we do so in the context of survival analysis, where there are multiple layers of noise (i.e., noise with respect to event occurrence and noise with respect to the TTE), requiring our approach to have an additional error prediction component. In addition, the distribution of observed labels does not always align with the distribution from which the ground truth label was drawn. Thus, learning to model the observed labels alone without considering the underlying survival distribution limits model calibration.

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 [80, 165, 142]. Similarly, in regression-like tasks, past work has proposed weighted averages of the labels based on estimated labeler reliability [127, 73]. Estimated labeler reliability is often based on heuristics, such as by measuring the consistency with other labelers [127, 73, 142] and finding the weights that maximize the margin between the aggregated label set (i.e., estimated ground truth labels) and other possible estimated ground truth label sets [165]. Due to the presence of censorship in survival data, our approach did not aggregate at preprocessing time since it is not straightforward how to aggregate a potentially incomplete set of labels (e.g., partially censored individuals). In addition, even on uncensored individuals, doing so would have required us to make assumptions on how the observed and ground truth TTEs were related (e.g., the ground truth TTE is equal to the average of the noisy TTEs).

### 5.2.2 Survival Analysis

Recent work on survival analysis with noisy labels focuses on noise from a labeler but has begun to consider a scenario where, for some individuals, reference TTEs can be obtained via chart review with an expert. This subset is then used to help learn the relationship between the noisy and ground truth TTE [109]. The relationship between the ground truth and noisy TTEs can be learned in several ways, such as learning a joint distribution [109], learning the sensitivity and specificities of the noisy TTEs for various time points [63], learning a mapping between the noisy and ground truth TTEs [123], and learning which features are most predictive of the ground truth TTE [122]. This class of approaches is the most similar to ours. However, the multiple noisy labels setting brings novel challenges in that individuals

can be partially censored, or the distribution of labeled TTEs cannot be directly translated to a survival distribution without sacrificing calibration.

Another related field in survival analysis that tackles a similar problem to ours is interval censoring [138, 102]. 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 the ground truth TTE and that at least one noisy TTE occurs after. That is, we may have cases where all of the labeled TTEs occur before or after the ground truth TTE.

### 5.2.3 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 [97, 134], where one aims to learn from a dataset where some instances are labeled and some are not. A classifier is learned from the labeled instances and then pseudo-labels from that classifier are applied to the unlabeled instances to improve training.

## 5.3 Methods

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

### 5.3.1 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,  $\ell$  (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 [172]. Given an event of interest, we aim to predict: 1) whether the event occurs within  $T$  time steps (i.e.,  $P(e \in \{1, 2, \dots, T\}|\mathbf{x})$ ) and 2) the probability of the event occurring at each time point  $1, 2, \dots, T$  given that it occurred within  $T$  time steps (i.e.,  $P(e = t|\mathbf{x}, e \in \{1, 2, \dots, T\})$  for  $t = 1, 2, \dots, T$ ). Using these predictions, the corresponding survival function is  $\hat{S}(t|\mathbf{x}) = 1 - \sum_{u=1}^t \hat{P}(e = u|\mathbf{x}) = 1 - \sum_{u=1}^t \hat{P}(e = u|\mathbf{x}, e \in \{1, 2, \dots, T\})P(e \in \{1, 2, \dots, T\}|\mathbf{x})$ , and the predicted TTE is the median value of  $\hat{S}(t|\mathbf{x})$  (i.e.,  $\hat{e} = \operatorname{argmin}_t \hat{S}(t|\mathbf{x}) \leq 0.5$ ) [137], where

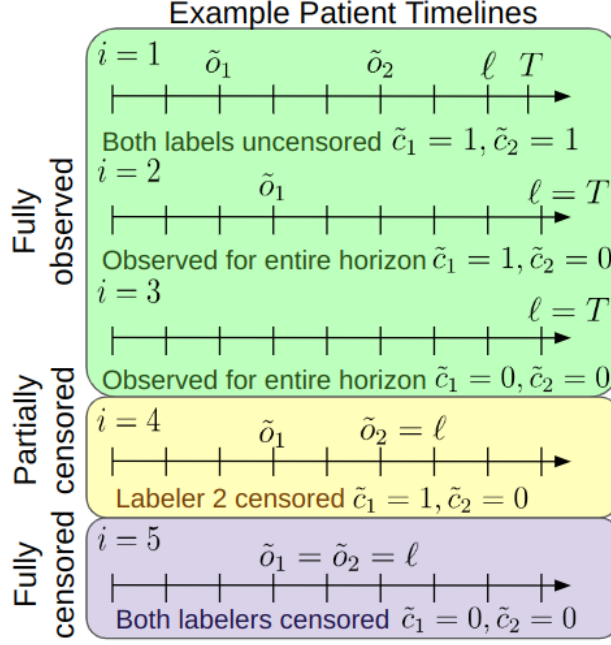


Figure 5.2: Demonstration with example timelines. Instead of observing ground truth time-to-events (TTEs) in our dataset, we observe  $m > 1$  noisy proxies. In this figure we have  $m = 2$  labelers for five example individuals. Notation)  $\ell$ : last time observed,  $\tilde{c}_j$ : indicator for whether an event was recorded for labeler  $j$   $\tilde{o}_j$ : observed time for labeler  $j$ , which is  $\ell$  if  $\tilde{c}_j == 0$  or noisy TTE  $\tilde{e}_j$  otherwise.

$\hat{\cdot}$  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  $\tilde{\mathbf{o}} \in \{1, 2, \dots, T\}^m$ , a vector of observed times for each labeler. Each entry  $\tilde{o}_j$  corresponds to the observed time for labeler  $j \in \{1, 2, \dots, m\}$ , which can be a noisy TTE or  $\ell$ . 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,  $\tilde{\mathbf{o}} = (\tilde{o}_1, \tilde{o}_2, \dots, \tilde{o}_m)$ . 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 5.2**). 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 throughout the entire prediction horizon (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). 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 [176], 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  $\ell$  is uncorrelated with  $e$ .

### 5.3.2 Proposed Approach

Our approach (**Figure 5.3**) 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. 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., a 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. 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.

**We now provide more details for each step.**

#### 5.3.2.1 Step 1a: Noisy TTE Prediction

Here, we learn to predict the distribution of noisy TTEs as the output of the model parameterized by  $\theta$ . Let  $\tilde{e}$  denote a noisy TTE sampled from  $\{\tilde{e}_1, \tilde{e}_2, \dots, \tilde{e}_m\}$  with uniform probability given  $\mathbf{x}$ . Note that the  $\tilde{e}_j$ 's 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  $[P(\tilde{e} = 1 | \mathbf{x}), P(\tilde{e} = 2 | \mathbf{x}), \dots, P(\tilde{e} = T | \mathbf{x}), P(\tilde{e} \notin \{1, 2, \dots, T\} | \mathbf{x})] \in [0, 1]^{T+1}$ .

For labeler  $j$ , censored and uncensored labels are handled in ways similar to past work [96, 144, 78, 166]. 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

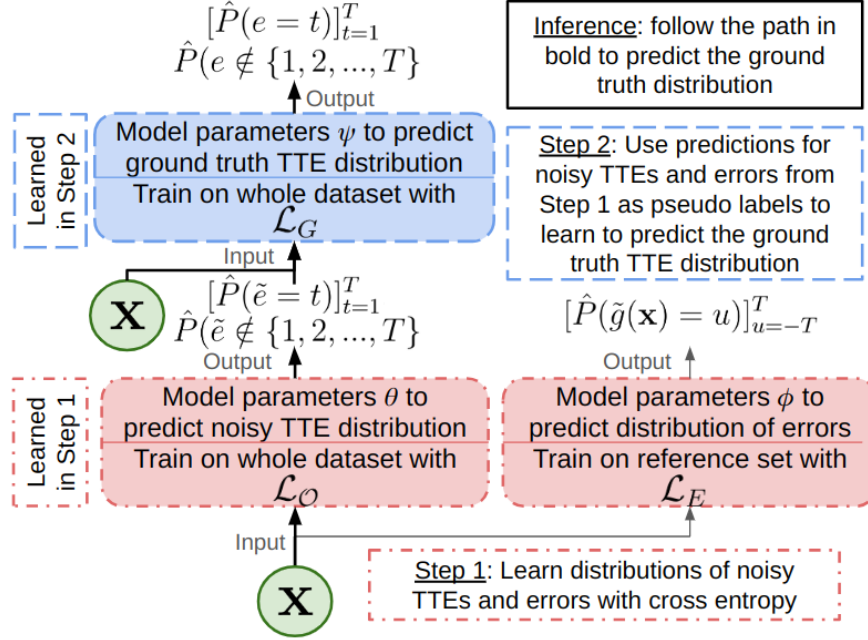


Figure 5.3: Overview of proposed approach. Based on  $\mathbf{x}$ , we first predict the noisy TTEs and their errors as intermediate outputs. We then concatenate the noisy TTE prediction with  $\mathbf{x}$  and map this to a prediction of the ground truth TTE as the final output of the approach. The 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.

event occurrence before the time of follow up using  $\mathcal{L}_C$  since we have no other information.

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

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

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

$$\mathcal{L}_O = \sum_{i=0}^n \sum_{j=1}^m \mathbb{I}(\tilde{c}_j^{(i)} = 1) \mathcal{L}_U(\mathbf{x}^{(i)}, j) + \mathbb{I}(\tilde{c}_j^{(i)} = 0) \mathcal{L}_C(\mathbf{x}^{(i)}, j)$$

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

### 5.3.2.2 Step 1b: Error Prediction

Here, we learn to predict the distribution of errors as the output of the model parameterized by  $\phi$ . 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  $\phi$  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  $P(e \notin \{1, 2, \dots, T\})$ . Similar to  $\tilde{e}$ , let  $\tilde{g}(\mathbf{x})$  denote a labeled TTE error sampled from  $\{\tilde{g}_1(\mathbf{x}), \tilde{g}_2(\mathbf{x}), \dots, \tilde{g}_m(\mathbf{x})\}$  with uniform probability. Like the  $\tilde{e}_j$ 's, the  $\tilde{g}_j(\mathbf{x})$ 's are not necessarily drawn from the same distribution. Predicting the distribution of TTE errors then becomes predicting  $[P(\tilde{g}(\mathbf{x}^{(i)}) = -T|\mathbf{x}), P(\tilde{g}(\mathbf{x}^{(i)}) = -T + 1|\mathbf{x}), \dots, P(\tilde{g}(\mathbf{x}^{(i)}) = T|\mathbf{x})] \in [0, 1]^{2T+1}$ .

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

### 5.3.2.3 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  $\psi$ , 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.,  $[P(e = 1|\mathbf{x}), P(e = 2|\mathbf{x}), \dots, P(e = T|\mathbf{x}), P(e \notin \{1, 2, \dots, T\}|\mathbf{x})] \in [0, 1]^{T+1}$ ). 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. Using these pseudo-labels, we use a cross-entropy loss to learn the ground truth survival distribution. We use cross entropy since it does not require any assumptions on the ground truth survival distribution other than that it can be approximated by the empirical survival distribution. This is represented in the first term of  $\mathcal{L}_G$  below.

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

Here,  $\hat{e}_{avg} = \lfloor \sum_t t \hat{P}(\tilde{e} = t | \tilde{e} \leq T, \mathbf{x}) + \sum_e \epsilon \hat{P}(\tilde{g}(\mathbf{x}) = \epsilon | \mathbf{x}) \rfloor$  is the mean predicted TTE [61], offset by the error prediction, whose value is clipped to 1 or  $T$  if needed.  $\hat{P}(e \leq T)$  is shorthand for  $\sum_{t=1}^T \hat{P}(e = t)$ , the probability of the event occurring within the prediction horizon, and  $\hat{P}(e = t | e \leq T) = \hat{P}(e = t) / \hat{P}(e \leq T)$  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  $\theta$ ,  $\phi$ , and  $\psi$  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 [125]. Thus, Step 2 serves as a re-calibration step.

**In contrast to approaches from classification, 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.

### 5.3.3 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. We present our main results below and include more detail in **Appendix C**.

#### 5.3.3.1 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 distribution of labeled TTEs and labeled 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 all labelers). We can

express the quantity  $P(|e - \hat{e}| \geq \varepsilon)$  for positive individuals with some error tolerance value  $\varepsilon$  as follows. Note that we only consider positive individuals because TTEs are not defined for negative individuals

$$\begin{aligned} &P(|e - \hat{e}| \geq \varepsilon) \text{ (here, we consider positive individuals)} \\ &\leq \frac{1}{\varepsilon} \alpha T \end{aligned}$$

where  $\alpha = \prod_{j=1}^m \alpha_j$  is the overall false negative rate, and  $\alpha_j$  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 lower false negative rates will lower this probability. As a result, **if we aim to train a model with target values  $\varepsilon$  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  $\varepsilon$  and  $p$  are realistic.**

### 5.3.3.2 Prediction Error for Event Occurrence

Similarly, for all individuals, we aim to minimize  $|P(e \notin \{1, 2, \dots, T\}) - \hat{P}(e \notin \{1, 2, \dots, T\})|$ . We learn this probability implicitly during Step 1 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/positive individuals. Therefore, we can express  $P(|P(e \notin \{1, 2, \dots, T\}) - \hat{P}(e \notin \{1, 2, \dots, T\})| \geq \delta)$  for error tolerance value  $\delta$  as

$$\begin{aligned} &P(|P(e \notin T) - \hat{P}(e \notin T)| \geq \delta) \\ &\leq \frac{1}{\delta} (\rho \alpha + (1 - \rho) \beta) \end{aligned}$$

where  $\beta$  is the false positive rate analog of  $\alpha$ , and  $\rho$  is the positive rate. In other words,  $P(|P(e \notin \{1, 2, \dots, T\}) - \hat{P}(e \notin \{1, 2, \dots, T\})| \geq \delta)$  *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  $\delta$  and  $0 < q < 1$  such that  $P(|P(e \notin \{1, 2, \dots, T\}) - \hat{P}(e \notin \{1, 2, \dots, T\})| \geq \delta) \leq q$ , then checking whether  $\rho \alpha + (1 - \rho) \beta \leq \delta q$  could serve as a guideline for whether  $\delta$  and  $q$  are realistic.**

## 5.4 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.

## 5.4.1 Datasets

### 5.4.1.1 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  $f(\mathbf{x})$ .  $I$  is the identity matrix;  $\sigma()$  is a sigmoid. Data generation is summarized below.

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

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  $\epsilon$  so that the distribution of noisy labels does not match the ground truth TTE distribution.

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

By this step, 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:  $\tilde{o}_j = T$ ,  $\tilde{c}_j = 0$

Individuals were chosen at random to be in  $\mathcal{C}$  and  $\mathcal{P}$  and the values of  $\tilde{o}$  such that  $\tilde{o}_j < \ell$  are set to  $\ell$  and the corresponding  $\tilde{e}_j$  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 with bounds  $[a, b]$ .

#### 5.4.1.2 Real Dataset

Within the healthcare domain, we leveraged MIMIC-III, a publicly available dataset of electronic health record data [75] 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 [156].

Like past work [117], 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 [162], 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 [76, 169, 145, 147]. The composite time to sepsis onset was defined like past work [77], using the CDC time if available, the CMS definition if the CDC time was not available, or no time was recorded if neither were available. As our noisy labels, we considered the Sepsis1 definition [19, 74], Sepsis3 definition [153, 117], and a composite definition made based on Sepsis3 and Sepsis1. All sepsis definitions were implemented as defined by Sean et al. [150]. Individual components, such as SIRS and SOFA, were identified based on past work<sup>12</sup>. Our inclusion/exclusion criteria were the same as Moor et al. [117], and also excluded admissions with TTEs within the first seven hours.

After preprocessing (see below), 486 (2.45%) admissions met the CMS definition, 122 (0.61%) met the CDC definition, 549 (2.76%) met the CMS/CDC 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 is in **Table 5.1**. For patients with noisy TTEs, 62.44% were

<sup>1</sup><https://github.com/MIT-LCP/mimic-code/tree/main>

<sup>2</sup><https://github.com/BorgwardtLab/mgp-tcn>

Table 5.1: 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)          |

administratively censored for at least labeler.

**Preprocessing.** Data 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), [162], 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, marital 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 ( $\text{theta}_1 = 0.001$ ,  $\text{theta}_2 = 0.001$ , and  $\text{theta\_freq} = 1.0$ ) 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.

## 5.4.2 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 [127, 142], or 2) inference time [33, 143, 30, 98]. The first two aggregate at preprocessing. The last aggregates at inference. All assume a known aggregation function (i.e., averaging).

**Naïve Average.** Noisy TTEs are averaged into a single label during preprocessing and are used to train a model with  $\mathcal{L}_O$  [127]. That is, for individuals in  $\mathcal{U}$  and  $\mathcal{P}$ , we aggregated  $\tilde{\mathbf{c}}$  into  $\mathbb{I}(\exists j \tilde{c}_j^{(i)} = 1)$ , and  $\tilde{\mathbf{o}}$  into the average of all uncensored TTEs (i.e.,  $\frac{1}{|\{j|\tilde{c}_j^{(i)}=1\}|} \sum_{j=1}^m \mathbb{I}(\tilde{c}_j^{(i)} = 1)\tilde{o}_j^{(i)}$ ). 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.

**Voting Average.** Noisy TTEs are aggregated via an averaging strategy that builds on *Naïve Average* [127, 142]. 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  $\ell$ . That is, if  $\frac{1}{m}|\{j|\tilde{c}_j^{(i)} = 1\}| \geq 0.5$ , 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 [143, 30] 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  $j = 1, 2, \dots, m$  and time points  $t = 1, 2, \dots, T$  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.

### 5.4.3 Evaluation

We evaluated all approaches with three metrics [137]: 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 [5]. For the remainder of this paper, we refer to the time-dependent C-index as the C-index for brevity. 3) We measured calibration on the synthetic dataset with the IMSE between the predicted and ground truth survival curves. For the MIMIC-III dataset, we measured calibration using the distributional divergence for calibration (DDC) [78] since we did not have access to the ground truth survival curves. For discriminative performance and calibration, 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, lower is better.

### 5.4.4 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

Table 5.2: 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       |

were initialized with 123456789+x where x is the replication number. The runtime for the results in **Figures 5.4** and **5.5** was 48 hours for each plot, and 30 minutes for **Table 5.3**.

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 5.2**. For the MIMIC-III dataset, where there was class 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  $\mathcal{L}_C$  during training.

For our model architectures, we trained feed forward neural networks whose architecture was similar to DeepHit [96, 166]. 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 outputted the probability of event occurrence at each time point. The proposed approach was implemented as three feed forward neural networks whose structure mirrored **Figure 5.3**.  $\theta$  was a feed forward network.  $\phi$  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.  $\psi$  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 [91]). All models were trained in Python3.7 and Pytorch1.7.1, using Adam.

## 5.5 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 with 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 robust the noise rate with respect to event occurrence through the second term of  $\mathcal{L}_G$

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 event occurrence. Then, we evaluated performance of the approaches on sepsis onset prediction in MIMIC-III.

### 5.5.1 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 all experiments on the synthetic data, ground truth labels in the test set were uncensored for evaluation.

#### 5.5.1.1 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

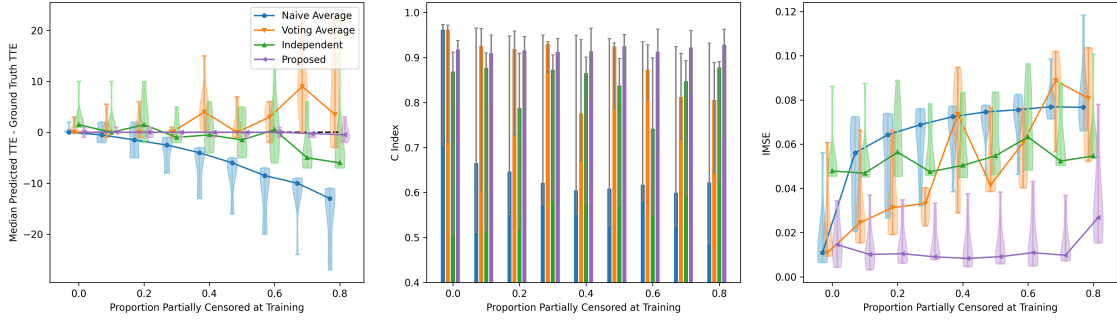


Figure 5.4: We varied the rate of partial censorship while keeping the rate of full censorship at 10% during training. The proposed approach degraded the least as the rate of partial censorship increased. At each point along the x-axis, we plot the median, and distribution of values for each approach 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.

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  $\theta$  and  $\psi$  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 5.4**). 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 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

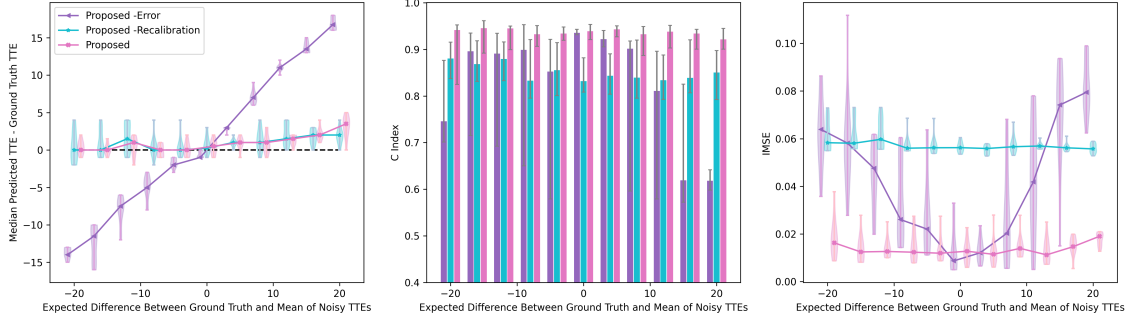


Figure 5.5: We varied the TTE noise mean. Our approach maintained good performance across all metrics, while its ablations did not. At each point along the x-axis, we plot the median, and distribution of values for each approach across all replications. We show the C-index results as a bar plot for clarity with error bars representing the range in values across runs. The other plots have jitter along the x-axis for clarity.

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 was the average of the noisy TTEs. 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].

### 5.5.1.2 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 (shown below), while *Proposed -Error* implemented Steps 1a and 2 and did not learn the error pattern among labelers.

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

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).

*Proposed* and *Proposed -Recalibration* were able to accurately predict the TTE across

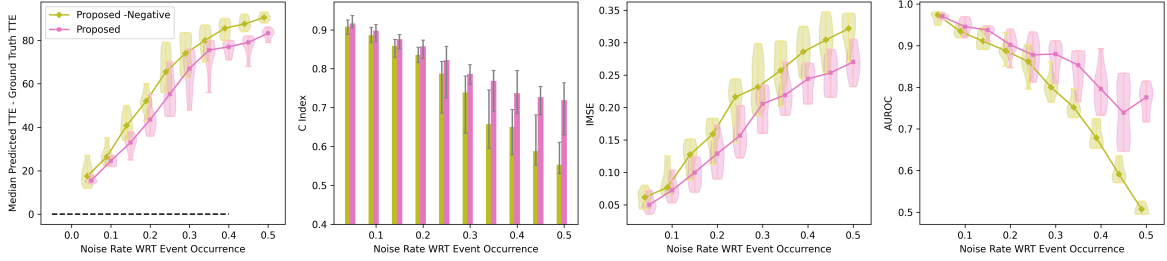


Figure 5.6: We varied the noise rate at training with respect to event occurrence. The proposed approach with explicit supervision over negative individuals degraded the least as the noise rate increased. At each point along the x-axis, we plot the median, and distribution of values for each approach 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 other plots have jitter along the x-axis for clarity.

a variety of noise means, while *Proposed -Error* was not (**Figure 5.5**). While TTE error degraded for *Proposed -Error*, we note that it was more stable across replications than the other metrics since it depended less on the variability of the predictions 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 settings. *Proposed -Recalibration* suffered with respect to both discriminative performance and calibration, 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.

### 5.5.1.3 Robustness to Event Occurrence 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 5.6**, 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

Table 5.3: 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 ( $\uparrow$ )  | DDC ( $\downarrow$ )    | AUROC ( $\uparrow$ )    |
|----------------|----------------------------|-------------------------|-------------------------|-------------------------|
| Naïve Average  | 17.50 [16.25-18.00]        | 0.48 [0.45-0.52]        | 1.14 [1.08-1.29]        | 0.49 [0.46-0.51]        |
| Voting Average | 17.50 [16.25-18.00]        | 0.45 [0.42-0.48]        | 1.13 [1.08-1.23]        | 0.48 [0.43-0.50]        |
| Independent    | 18.50 [17.25-19.00]        | 0.48 [0.46-0.50]        | 0.91 [0.84-1.12]        | 0.47 [0.44-0.50]        |
| Proposed       | <b>14.50</b> [13.25-15.75] | <b>0.59</b> [0.54-0.62] | <b>0.86</b> [0.82-1.23] | <b>0.63</b> [0.55-0.66] |

the second term of  $\mathcal{L}_G$  that provides explicit supervision over negative individuals in the reference set. As shown in the results, we observed modest improvements in both the TTE error and calibration, 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. We include additional results on discriminative performance between those who did and did not experience the event. Having negative individuals introduces an extra dimension of evaluation for survival models. That is, can the model distinguish who does and does not experience the event? This is similar to measuring the the C-index, but at a coarser granularity and can be done by using the AUROC from classification. The overall trends were consistent with those for the C-index.

### 5.5.2 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 5.3**. 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 and the limited set of features that we used. We chose our features to align with past work [117] and exclude those that might lead to label leakage [77, 182]. 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 discriminative performance between those who did/did not experience sepsis. Like the synthetic dataset, we also observe similar trends to the C-index when we examine the AUROC.

## 5.6 Conclusion

We tackled the problem of multiple noisy labels in survival analysis and introduced a new approach. Our approach first predicts the distribution of noisy TTEs and their errors and then maps that to a prediction of the ground truth TTE. Through our experiments, we showed where adaptations of approaches from classification fail and demonstrated how our approach builds on them by addressing some of the challenges unique to survival analysis. Our approach is not without limitation. For example, we rely on a small, randomly sampled subset of clinician-labeled TTEs or are required to assume a known aggregation function in the absence of such a subset. Although assuming a known aggregation is unlikely to hold in practice, even when labels are uncensored, there are many settings in which we can obtain ground truth labels from a subject expert for a small subset of the data. 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. Nonetheless, we bring together concepts from the noisy labels literature in classification and survival analysis to highlight an important problem.

## CHAPTER 6

### Conclusion

In this thesis, we examined issues around multiple events and noisy labels in the context of predicting clinical outcomes using survival analysis. For many outcomes in healthcare, mining EHR data has become increasingly popular [193] due to their widespread adoption and potential to contain data representative of routine clinical care. However, working with EHR data is difficult since EHR data are not carefully curated. Patients may enter and leave the healthcare system as they please, leaving inconclusive outcomes for many conditions, giving rise to censored labels. Instead of documenting clinical diagnoses, EHR diagnoses often represent billing codes, which are not always accurate [168, 121, 167]. This gives rise to noisy labels, both at the class level (e.g., whether or not a patient was diagnosed with AD) and the TTE level (e.g., when was a patient diagnosed with AD). Learning to train machine learning (ML) models with censored and noisy labels using survival analysis then becomes an important problem to study in order to effectively leverage EHR data.

#### 6.1 Summary of Contributions

To address these issues, we built from work in survival analysis, focusing on 1) multi-event survival analysis and 2) noisy label learning with respect to both whether and when the event occurs. We summarize the specific challenges we address and our contributions below.

In survival analysis, we aim to predict TTEs while incorporating both censored and uncensored individuals during training. Early work in the multi-event setting relies on restrictive assumptions like the TTEs among events being independent [94]. Recent work has addressed this limitation in part by using a multitask framework [96]. However, in doing so, the model may still not be able to efficiently learn these dependencies since temporal relationships among TTEs at coarse granularities are not explicitly leveraged. For example, if the TTEs for two different events are often close together, then it may be easier to first learn that they both occur in the same coarse grained time bin (i.e., a lower variance task), as

opposed to immediately learning that they occur at two separate time points. In **Chapter 3**, we proposed a hierarchical approach for survival analysis that leverages the relatedness among time points to improve survival predictions. Our approach breaks down the main task into a series of coarser grained, more simple subtasks that build on each other. Predictions from the subtasks can then be combined to obtain a prediction for the main task. We hypothesized that our approach improved predictions by more tightly constraining the upper bound on the absolute error of the survival curve, and we showed empirically that our approach improved performance over well-accepted baselines.

The field of noisy label learning aims to learn accurate class predictions, even when the training data can contain incorrect labels. We first focused on label noise with respect to whether the event occurs, which can be considered as a binary classification task. Of the growing interest in instance-dependent label noise, one limitation affecting approaches was that many did not consider the effects of bias in label noise and how it would impact downstream performance. For example, if female patients are underdiagnosed for heart disease compared to male patients and therefore have a higher noise rate compared to male patients, current approaches could potentially generalize better to male patients. Even among approaches that have begun to consider these biases, many are limited in that they assume that label noise depends only on group membership (e.g., male and female) [106, 174] or that individual features can be broken down into distinct categories [184]. In **Chapter 4**, we proposed an approach for learning with instance-dependent label noise that uses a small subset for which we have ground truth labels to learn the noise function. Based on the learned noise function, our approach re-weights the objective function such that it is unbiased in expectation with respect to group membership. We hypothesized that this would allow our approach to improve its discriminative performance without sacrificing bias mitigation, and we show empirically that our approach is effective in doing so compared to state-of-the-art baselines.

Although the majority of work in noisy label learning focuses on classification, label noise with respect to when the event occurs has also received attention in the survival analysis literature. Past work on noisy labels in survival analysis focuses on a setting similar to the one we explored in the previous contribution, where there is a small subset of the data for which we have both ground truth and noisy labels [109, 63, 122, 123]. These works aim to learn the relationship between the noisy and ground truth TTEs in the context of a single noisy labeler. In **Chapter 5**, we proposed an approach for learning with multiple noisy labelers in survival analysis. Unlike the single labeler case, multiple labels have not yet been considered in survival analysis. Because of this, additional challenges arose that were not present in the single labeler setting. For example, learning to predict the survival function by

learning to predict the distribution of noisy TTEs can result in survival curves that have poor discriminative performance or are not calibrated. Our approach addresses these challenges by predicting the distributions of noisy TTEs and the respective errors separately and then combining these predictions to use as pseudo-labels to train the model to predict the ground truth survival function. We also provided practical guidelines for how model accuracy can be affected by the overall noise of the labelers based on theoretical analyses and empirically showed that our approach improved on adaptations of approaches from classification.

## 6.2 What Next?

There are several opportunities to build on our contributions. We describe some below.

In our hierarchical approach, we considered different settings beyond single event analysis, such as multiple events, competing risks, and semi-competing risks. However, each event was only allowed to occur once. Among other specialized settings within survival analysis, an interesting extension for this approach may be to consider recurrent events. Recurrent events describes a setting in survival analysis where an event can happen more than once [163]. For example, a patient could be diagnosed with mild cognitive impairment (i.e., a state of cognitive decline that often precedes AD and dementia in general), revert to being cognitively healthy, and then convert back to having mild cognitive impairment [86]. Here, the patient experiences the onset of mild cognitive impairment twice. The most straightforward way to extend the approach could be to model each occurrence as a separate event. However, this is limited in that it requires the number of times an event is allowed to occur to be predefined, and the approach is not guaranteed to produce consistent estimates. For example, the TTE prediction for the first event occurrence may occur after the prediction of the second TTE. In addition to recurrent events, others may want to consider extending the approach to use a recurrent network architecture, to further leverage the relatedness between time points. Recurrent architectures have been adopted by past work [144], but only in the context of predicting the main task without considering hierarchically nested subtasks.

In our approach for learning with instance-dependent class labels, we relied on having access to a small subset of data for which both ground truth and observed labels were available. Our results showed that the approach was sometimes sensitive to the size and composition of this subset. In settings where the alignment set is too small or too biased, it could be interesting to examine how to pick individuals from the training set to give to an expert who can then assign ground truth labels via manual review. This problem is currently being explored in the field of active learning [152], a setting from semi-supervised learning where the majority of data are unlabeled and we would like to know for which

instances, if a ground truth label was obtained, would model performance improve the most. In our setting, criteria for instance selection would likely be based on a combination of discriminative performance and bias mitigation. Developing criteria to pick instances for labeling and quantifying the expected improvement in performance are potential avenues for future work. By extension, in the absence of an alignment set at training time, active learning could also be used to construct one. Another assumption that our approach relied on was that the groups of interest were known in advance. Others may want to consider ways to extend the approach to settings where the groups are not known at training time.

Finally, in our approach for survival analysis with multiple noisy labels, we noticed that, although our approach applied to a broader range of settings than the baselines and ablations, it was not always the most sample efficient. For example, when the noise was 0-mean and censorship was low, it was more effective to use *Naive Average*. In addition, when the label noise was 0-mean, an ablation of the proposed approach that did not have the error prediction component achieved better calibration than the proposed approach. This shows that the error prediction and ground truth mapping components were data-hungry and likely relied on the size of the reference set. Similar to the alignment set from our approach for noisy class labels, using active learning to select which individuals to label may be interesting to explore. Unlike the classification setting, however, there are potentially more aspects of evaluation to consider in survival analysis, and combining them may be less straightforward due to the different ranges that evaluation metrics may take. In addition, we did not consider bias mitigation in this contribution since we focused on challenges specific to survival analysis. However, like any other instance-dependent label noise problem, extensions of the approach that address bias in label noise could also be worth considering in survival analysis.

## 6.3 Final Conclusion

Our main contributions are: 1) we improve the accuracy of survival predictions through our hierarchical approach, where we break down the main task into a series of subtasks that build on each other; 2) we propose an approach to learn with instance-dependent label noise in classification and consider settings in which bias in label noise can hurt performance at inference time for subsets of the data; and 3) we propose an approach to learn with multiple noisy labels in survival analysis such that discriminative performance and calibration are preserved while maintaining accurate TTE predictions. Though motivated by EHR datasets from healthcare, the issues addressed in this thesis can be found in other fields, and thus, our contributions can play a role in effectively learning from datasets with censored and noisy labels in the broader ML community.

## APPENDIX A

### Supplement for Chapter 3

#### A.1 Extended Derivation for Claim 2

Within each coarse grained bin  $\tau_z$ , the error can be upper bounded for the coarse grained (i.e., guiding) and direct (i.e., original) tasks by

$$\begin{aligned}
& \frac{1}{b} \sum_{t=zb+1}^{z(b+1)} |\hat{S}_G(t|\mathbf{x}) - S(t|\mathbf{x})| \\
& \leq \frac{1}{b} \sum_{t=zb+1}^{z(b+1)} |\max(\hat{S}_G(zb+1|\mathbf{x}), S(zb+1|\mathbf{x})) - \min(\hat{S}_G(z(b+1)|\mathbf{x}), S(z(b+1)|\mathbf{x}))| \\
& \leq \frac{1}{b} \sum_{t=zb+1}^{z(b+1)} \left| |\hat{S}_G(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| + |S(zb+1|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \right. \\
& \quad \left. + |\hat{S}_G(z(b+1)|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \right| \\
& = |\hat{S}_G(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| + |S(zb+1|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \\
& \quad + |\hat{S}_G(z(b+1)|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \text{ (Guiding task)}
\end{aligned}$$

and

$$\begin{aligned}
& \frac{1}{b} \sum_{t=zb+1}^{z(b+1)} |\hat{S}_O(t|\mathbf{x}) - S(t|\mathbf{x})| \\
& \leq \frac{1}{b} \sum_{t=zb+1}^{z(b+1)} |\max(\hat{S}_O(zb+1|\mathbf{x}), S(zb+1|\mathbf{x})) - \min(\hat{S}_O(z(b+1)|\mathbf{x}), S(z(b+1)|\mathbf{x}))| \\
& \leq \frac{1}{b} \sum_{t=zb+1}^{z(b+1)} \left| |\hat{S}_O(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| + |S(zb+1|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \right. \\
& \quad \left. + |\hat{S}_O(z(b+1)|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \right| \\
& = |\hat{S}_O(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| + |S(zb+1|\mathbf{x}) - S(z(b+1)|\mathbf{x})|
\end{aligned}$$

$$+ |\hat{S}_O(z(b+1)|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \text{ (Original task)}$$

The inequality

$$\begin{aligned} & \frac{1}{T/b} \sum_{z=1}^{T/b} |\hat{S}_G(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| + |\hat{S}_G(z(b+1)|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \\ & < \frac{1}{T/b} \sum_{z=1}^{T/b} |\hat{S}_O(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| + |\hat{S}_O(z(b+1)|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \end{aligned}$$

is true since

$$\begin{aligned} & \frac{1}{T/b} \sum_{z=1}^{T/b} |\hat{S}_G(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| + |\hat{S}_G(z(b+1)|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \\ & = \frac{2}{T/b} \sum_{z=1}^{T/b} |\hat{S}_G(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| \\ & < \frac{2}{T/b} \sum_{z=1}^{T/b} |\hat{S}_O(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| \\ & = \frac{1}{T/b} \sum_{z=1}^{T/b} |\hat{S}_O(zb+1|\mathbf{x}) - S(zb+1|\mathbf{x})| + |\hat{S}_O(z(b+1)|\mathbf{x}) - S(z(b+1)|\mathbf{x})| \end{aligned}$$

The second and fourth lines are true because the first and last terms of the expanded sum are 0 since both the predicted and ground truth survival curves start at 1 and end at 0, and the remaining terms are repeats. The third line is true since we assume that the guiding task is more accurate at the bin edges. As a result, the maximum error on the survival curve is smaller if we can more accurately predict the coarser grained task.  $\square$

## APPENDIX B

### Supplement for Chapter 4

#### B.1 Proposed and Clean Label Loss

Minimizing the proposed loss  $\mathcal{L}'_\theta$  from Step 2 of the proposed method is equal to minimizing cross entropy on the clean labels in expectation (i.e.,  $\mathcal{L}'_\theta$  is unbiased in expectation).

$$\begin{aligned}
\mathcal{L}'_\theta &= \frac{-1}{|\bar{A}|} \sum_{k=1}^g \sum_{i \in \bar{A} \cap G_k} \frac{1}{1 - \hat{r}_k} \sum_{j=1}^c \hat{\beta}_\phi^{(i)} \mathbb{I}(\tilde{y}^{(i)} = j) \log(\hat{y}_j^{(i)}) \\
&\Rightarrow \mathbb{E} \left[ \sum_{k=1}^g \sum_{i \in \bar{A} \cap G_k} \frac{1}{1 - \hat{r}_k} \sum_{j=1}^c \hat{\beta}_\phi^{(i)} \mathbb{I}(\tilde{y}^{(i)} = j) \log(\hat{y}_j^{(i)}) \right] \\
&= \sum_{k=1}^g \sum_{i \in \bar{A} \cap G_k} \frac{1}{1 - \hat{r}_k} \sum_{j=1}^c \mathbb{E} \left[ \hat{\beta}_\phi^{(i)} \mathbb{I}(\tilde{y}^{(i)} = j) \log(\hat{y}_j^{(i)}) \right] \\
&= \sum_{k=1}^g \sum_{i \in \bar{A} \cap G_k} \frac{1}{1 - \hat{r}_k} \sum_{j=1}^c (1 - \hat{r}_k) \mathbb{I}(y^{(i)} = j) \log(\hat{y}_j^{(i)}) \\
&= \sum_{k=1}^g \sum_{i \in \bar{A} \cap G_k} \sum_{j=1}^c \mathbb{I}(y^{(i)} = j) \log(\hat{y}_j^{(i)})
\end{aligned}$$

As a reminder, each group  $G_k$  is then associated with estimated noise rate  $\hat{r}_k = \frac{1}{|g_k|} \sum_{i \in G_k} 1 - \hat{\beta}_\phi^{(i)}$  and estimated clean (i.e., correct) rate  $1 - \hat{r}_k = \frac{1}{|G_k|} \sum_{i \in G_k} \hat{\beta}_\phi^{(i)}$ . We can express the noise and clean rates in terms of  $\hat{\beta}_\phi^{(i)}$  since

$$\begin{aligned}
1 - r_k &= \frac{1}{|G_k|} \sum_{i \in G_k} \mathbb{I}(\tilde{y}^{(i)} = y^{(i)}) \\
&= P(y = \tilde{y} | \tilde{y}, \mathbf{x}) \text{ for a random instance in } G_k \\
&= \frac{1}{|G_k|} \sum_{i \in G_k} P(y^{(i)} = \tilde{y}^{(i)} | \tilde{y}^{(i)}, \mathbf{x}^{(i)})
\end{aligned}$$

where  $r_k$  and  $1 - r_k$  are the actual noise and clean rates within group  $k$ , respectively. Therefore, since  $\hat{\beta}_\phi$  is trained to predict  $P(y = \tilde{y}|\tilde{y}, \mathbf{x})$ , we estimate the noise rates using  $\hat{\beta}_\phi$ .

## B.2 Extended Results for Sensitivity to Alignment Set

For our experiments varying the size and composition of the alignment set, we show the AUROC and AUEOC separately below.

When considering discriminative performance, we found that the proposed approach was generally robust to changes in both alignment set size and bias across datasets (**Figure B.1**). The exception appears to be with the COMPAS dataset, where performance across all approaches had more variability due to the limited dataset size. We also note that discriminative performance of the proposed approach worsened considerably on the synthetic dataset when the alignment set was completely biased to either group. This may be due to the way the dataset was constructed, where the differences in the features between the groups was exaggerated for demonstrative purposes.

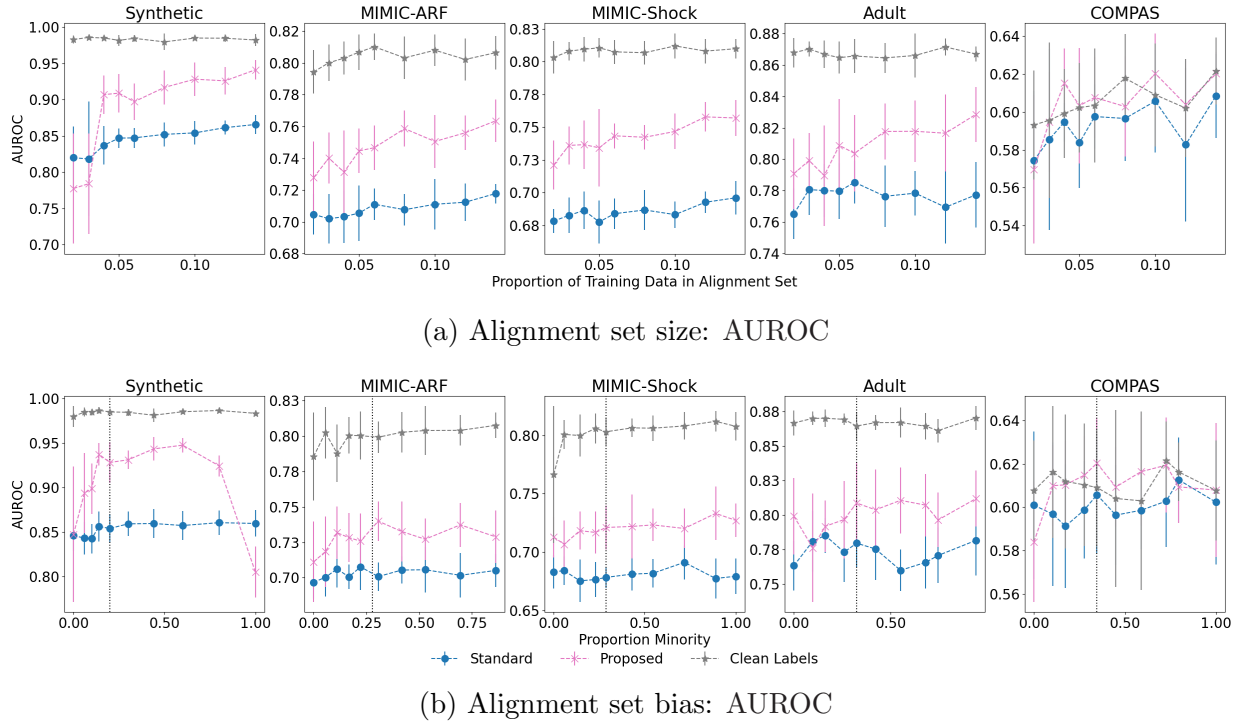
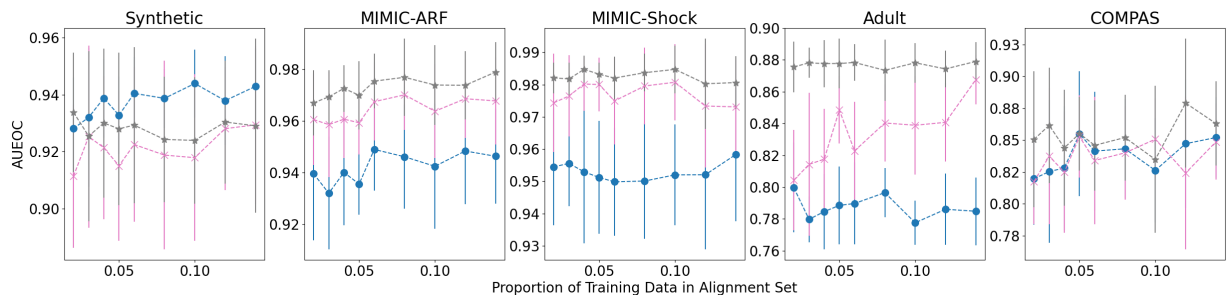


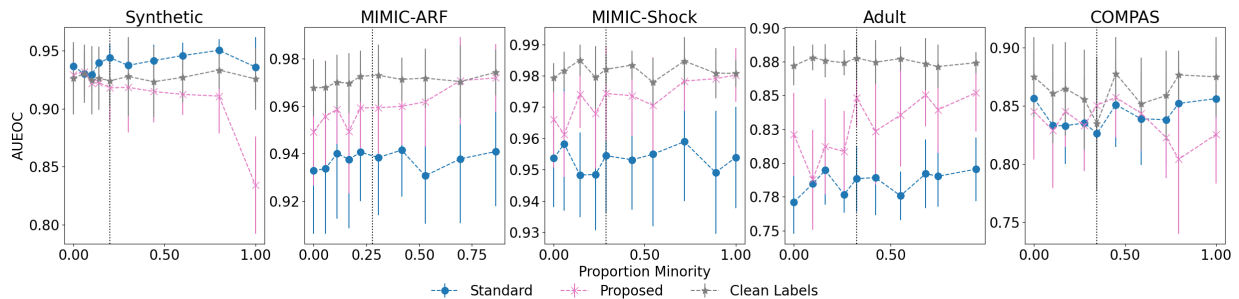
Figure B.1: Robustness with respect to AUROC when varying alignment set size and bias. We plot the mean and standard deviation for 10 random seeds.

When considering bias mitigation, we observed similar results to discriminative performance (**Figure B.2**). In contrast to the previous set of results, the standard approach

appeared to perform the best with respect with respect to the AUEOC on the synthetic dataset across different levels of alignment set bias. This may be because both the outcome and pattern of label noise were constructed to be ‘easy’ to learn. In doing so, the proposed approach may have learned an underlying noise pattern that was near-deterministic, leading it move away from soft filtering to a harder form of filtering. Therefore, the limitations of filtering approaches from past work with respect to bias mitigation would also apply here. However, for the other datasets, we do not observe this pattern, which is promising since they represent a more realistic setting.



(a) Alignment set size: AUEOC



(b) Alignment set bias: AUEOC

Figure B.2: Robustness with respect to AUEOC when varying alignment set size and bias. We plot the mean and standard deviation for 10 random seeds.

## APPENDIX C

### Supplement for Chapter 5

#### C.1 Extended Derivation for Practical Guidelines Based on Theoretical Noise Sensitivity

Here we present the extended derivation for the section ‘Practical Guidelines Based on Theoretical Noise Sensitivity’ in **Chapter 5**.

##### C.1.1 TTE Prediction Error in Positive Individuals

Using Markov’s inequality, we can express the quantity  $P(|e - \hat{e}| \geq \varepsilon)$  for positive individuals with some error tolerance value  $\varepsilon$  as follows.

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

where  $\alpha = \prod_{j=1}^m \alpha_j$  is the overall false negative rate, and  $\alpha_j$  is the false negative rate for labeler  $j$ . 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 labeled TTEs and labeled TTE errors for true positive individuals, so the expected error for  $e - \hat{e}$  will mainly depend on false negative individuals.

### C.1.2 Prediction Error for Event Occurrence

Similarly, we can express  $P(|P(e \notin \{1, 2, \dots, T\}) - \hat{P}(e \notin \{1, 2, \dots, T\})| \geq \delta)$  for error tolerance value  $\delta$  as

$$\begin{aligned}
& P(|P(e \notin T) - \hat{P}(e \notin T)| \geq \delta) \\
& \leq \frac{1}{\delta} \mathbb{E}_P[|P(e \notin T) - \hat{P}(e \notin T)|] \text{ (Markov)} \\
& = \frac{1}{\delta} \left( \rho(1 - \alpha) \mathbb{E}_{TP}[|P(e \notin T) - \hat{P}(e \notin T)|] \right. \\
& \quad + \rho\alpha \mathbb{E}_{FN}[|P(e \notin T) - \hat{P}(e \notin T)|] \\
& \quad + (1 - \rho)(1 - \beta) \mathbb{E}_{TN}[|P(e \notin T) - \hat{P}(e \notin T)|] \\
& \quad \left. + (1 - \rho)\beta \mathbb{E}_{FP}[|P(e \notin T) - \hat{P}(e \notin T)|] \right) \text{ (breakdown)} \\
& \approx \frac{1}{\delta} \left( \rho\alpha \mathbb{E}_{FN}[|P(e \notin T) - \hat{P}(e \notin T)|] \right. \\
& \quad \left. + (1 - \rho)\beta \mathbb{E}_{FP}[|P(e \notin T) - \hat{P}(e \notin T)|] \right) \text{ (if we assume (**))} \\
& \leq \frac{1}{\delta} (\rho\alpha + (1 - \rho)\beta) \text{ (since the absolute error is at most 1)}
\end{aligned}$$

where  $\beta$  is the false positive rate analog of  $\alpha$ , and  $\rho$  is the positive rate. Similarly, TN: true negative, FP: false positive,  $\beta$ : false positive rate,  $\rho$ : positive rate. (\*\*) We assume that we can accurately learn from true positive/negative individuals, so the expected error for  $P(e \notin T) - \hat{P}(e \notin T)$  will mainly depend on false negatives/positives.

## BIBLIOGRAPHY

- [1] José Alemán and David D Yang. A duration analysis of democratic transitions and authoritarian backslides. *Comparative Political Studies*, 44(9):1123–1151, 2011.
- [2] Per Kragh Andersen and Richard D Gill. Cox’s regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.
- [3] Per Kragh Andersen and Niels Keiding. Multi-state models for event history analysis. *Statistical methods in medical research*, 11(2):91–115, 2002.
- [4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias, 2016.
- [5] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.
- [6] Gregory T Armstrong, Toana Kawashima, Wendy Leisenring, Kayla Stratton, Marilyn Stovall, Melissa M Hudson, Charles A Sklar, Leslie L Robison, and Kevin C Oeffinger. Aging and risk of severe, disabling, life-threatening, and fatal events in the childhood cancer survivor study. *Journal of clinical oncology*, 32(12):1218, 2014.
- [7] Devansh Arpit, Stanislaw K Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron C Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *ICML*, 2017.
- [8] Alzheimer’s Association. 2014 alzheimer’s disease facts and figures. *Alzheimer’s & dementia*, 20(5), 2024.
- [9] Anand Avati, Tony Duan, Sharon Zhou, Kenneth Jung, Nigam H Shah, and Andrew Y Ng. Countdown regression: sharp and calibrated survival predictions. In *Uncertainty in Artificial Intelligence*, pages 145–155. PMLR, 2020.
- [10] HeeSun Bae, Seungjae Shin, Byeonghu Na, JoonHo Jang, Kyungwoo Song, and Il-Chul Moon. From noisy prediction to true label: Noisy prediction calibration via generative model. In *International Conference on Machine Learning*, pages 1277–1297. PMLR, 2022.
- [11] Noam Barda, Noa Dagan, Cyrille Cohen, Miguel A Hernán, Marc Lipsitch, Isaac S Kohane, Ben Y Reis, and Ran D Balicer. Effectiveness of a third dose of the bnt162b2 mrna covid-19 vaccine for preventing severe outcomes in israel: an observational study. *The Lancet*, 398(10316):2093–2100, 2021.

- [12] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.
- [13] Omer Ben-Aharon, Racheli Magnezi, Moshe Leshno, and Daniel A Goldstein. Median survival or mean survival: which measure is the most appropriate for patients, physicians, and policymakers? *The Oncologist*, 24(11):1469–1478, 2019.
- [14] Ralf Bender and Ulrich Grouven. Ordinal logistic regression in medical research. *Journal of the Royal College of physicians of London*, 31(5):546, 1997.
- [15] Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *International Conference on Machine Learning*, pages 825–836. PMLR, 2021.
- [16] J Martin Bland and Douglas G Altman. The logrank test. *Bmj*, 328(7447):1073, 2004.
- [17] Frank Bogun, Daejoon Anh, Gautham Kalahasty, Erik Wissner, Chadi Bou Serhal, Rabih Bazzi, W Douglas Weaver, and Claudio Schuger. Misdiagnosis of atrial fibrillation and its clinical consequences. *The American journal of medicine*, 117(9):636–642, 2004.
- [18] Roger C Bone. Sepsis and its complications: the clinical problem. *Critical care medicine*, 22(7):S8–11, 1994.
- [19] Roger C Bone, Robert A Balk, Frank B Cerra, R Phillip Dellinger, Alan M Fein, William A Knaus, Roland MH Schein, and William J Sibbald. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, 101(6):1644–1655, 1992.
- [20] NE Breslow. Discussion of the paper by dr cox cited below. *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972.
- [21] Cavell Brownie, James E Hines, and James D Nichols. Constant-parameter capture-recapture models. *Biometrics*, pages 561–574, 1986.
- [22] Jaime S Cardoso, Joaquim F Pinto da Costa, and Maria J Cardoso. Modelling ordinal relations with svms: An application to objective aesthetic evaluation of breast cancer conservative treatment. *Neural Networks*, 18(5-6):808–817, 2005.
- [23] Javier Castañeda and Bart Gerritse. Appraisal of several methods to model time to multiple events per subject: modelling time to hospitalizations and death. *Revista Colombiana de Estadística*, 33(1):43–61, 2010.
- [24] Trenton Chang, Michael W Sjoding, and Jenna Wiens. Disparate censorship & under-testing: A source of label bias in clinical machine learning. In *MLHC*, 2022.
- [25] Pengfei Chen, Guangyong Chen, Junjie Ye, Jingwei Zhao, and Pheng-Ann Heng. Noise against noise: stochastic label noise helps combat inherent label noise. In *International Conference on Learning Representations*, 2021.

- [26] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *International Conference on Machine Learning*, pages 1062–1070, 2019.
- [27] De Cheng, Yixiong Ning, Nannan Wang, Xinbo Gao, Heng Yang, Yuxuan Du, Bo Han, and Tongliang Liu. Class-dependent label-noise learning with cycle-consistency regularization. In *Advances in Neural Information Processing Systems*, 2022.
- [28] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *International Conference on Learning Representations*, 2020.
- [29] Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. In *International Conference on Machine Learning*, pages 1789–1799. PMLR, 2020.
- [30] Zhendong Chu, Jing Ma, and Hongning Wang. Learning from crowds by modeling common confusions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5832–5840, 2021.
- [31] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.
- [32] David R Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.
- [33] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.
- [34] Tiago de Freitas Pereira and Sébastien Marcel. Fairness in biometrics: a figure of merit to assess biometric verification systems. *arXiv e-prints*, pages arXiv–2011, 2020.
- [35] Yuichiro Doki, Jaffer A Ajani, Ken Kato, Jianming Xu, Lucjan Wyrwicz, Satoru Motoyama, Takashi Ogata, Hisato Kawakami, Chih-Hung Hsu, Antoine Adenis, et al. Nivolumab combination therapy in advanced esophageal squamous-cell carcinoma. *New England Journal of Medicine*, 386(5):449–462, 2022.
- [36] Dheeru Dua and Casey Graff. Uci machine learning repository. <http://archive.ics.uci.edu/ml>, 2017.
- [37] Matthew Engelhard, Samuel Berchuck, Joshua D’Arcy, and Ricardo Henao. Neural conditional event time models. In *MLHC*, 2020.
- [38] Erik Engleson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34, 2021.

- [39] Jonathan T Evans, Robert W Walker, Jonathan P Evans, Ashley W Blom, Adrian Sayers, and Michael R Whitehouse. How long does a knee replacement last? a systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up. *The Lancet*, 393(10172):655–663, 2019.
- [40] G Michael Felker, Kevin J Anstrom, Kirkwood F Adams, Justin A Ezekowitz, Mona Fiuzat, Nancy Houston-Miller, James L Januzzi, Daniel B Mark, Ileana L Piña, Gayle Passmore, et al. Effect of natriuretic peptide-guided therapy on hospitalization or cardiovascular mortality in high-risk patients with heart failure and reduced ejection fraction: a randomized clinical trial. *JAMA*, 318(8):713–720, 2017.
- [41] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *International Joint Conferences on Artificial Intelligence*, pages 2206–2212, 2020.
- [42] Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509, 1999.
- [43] Jason P Fine, Hongyu Jiang, and Rick Chappell. On semi-competing risks data. *Biometrika*, 88(4):907–919, 2001.
- [44] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *Machine Learning: ECML 2001: 12th European Conference on Machine Learning Freiburg, Germany, September 5–7, 2001 Proceedings 12*, pages 145–156. Springer, 2001.
- [45] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.
- [46] Dragan Gamberger, Nada Lavrač, and Sašo Džeroski. Noise elimination in inductive concept learning: A case study in medical diagnosis. In *International Workshop on Algorithmic Learning Theory*, pages 199–212. Springer, 1996.
- [47] Guadalupe Gomez, Olga Julià, Frederic Utzet, and Melvin L Moeschberger. Survival analysis for left censored data. In *Survival analysis: State of the art*, pages 269–288. Springer, 1992.
- [48] Samuel R Gross, Barbara O’Brien, Chen Hu, and Edward H Kennedy. Rate of false conviction of criminal defendants who are sentenced to death. *Proceedings of the National Academy of Sciences*, 111(20):7230–7235, 2014.
- [49] Shulamith T Gross and Tze Leung Lai. Nonparametric estimation and regression analysis with left-truncated and right-censored data. *Journal of the American Statistical Association*, 91(435):1166–1180, 1996.
- [50] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

- [51] Pedro Antonio Gutiérrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervás-Martínez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2015.
- [52] Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *The Journal of Machine Learning Research*, 21(1):3289–3351, 2020.
- [53] Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor Tsang, and Masashi Sugiyama. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*, pages 4006–4016. PMLR, 2020.
- [54] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pages 8527–8537, 2018.
- [55] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.
- [56] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.
- [57] Edward F Harrington. Online ranking/collaborative filtering using the perceptron algorithm. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 250–257, 2003.
- [58] Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.
- [59] James J Heckman and Burton Singer. Econometric duration analysis. *Journal of econometrics*, 24(1-2):63–132, 1984.
- [60] Jin-Jian Hsieh and Jian-Lin Wang. Quantile residual life regression based on semi-competing risks data. *Journal of Applied Statistics*, 45(10):1770–1780, 2018.
- [61] Shi Hu, Egill Fridgeirsson, Guido van Wingen, and Max Welling. Transformer-based deep survival analysis. In *Survival Prediction-Algorithms, Challenges and Applications*, pages 132–148. PMLR, 2021.
- [62] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Mingyue Niu, and Minghao Yang. Deep learning for continuous multiple time series annotations. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 91–98, 2018.
- [63] Sally Hunsberger, Paul S Albert, and Lori Dodd. Analysis of progression-free survival data using a discrete time survival model that incorporates measurements with and without diagnostic error. *Clinical Trials*, 7(6):634–642, 2010.

- [64] Panagiotis G Ipeirotis, Foster Provost, Victor S Sheng, and Jing Wang. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*, 28:402–441, 2014.
- [65] Joseph Jacob, Olga Ciccarelli, Frederik Barkhof, and Daniel C Alexander. Disentangling human error from the ground truth in segmentation of medical images. ACL, 2021.
- [66] Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A Young, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar. Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 105–115, 2021.
- [67] Fei Jiang and Sebastien Haneuse. A semi-parametric transformation frailty model for semi-competing risks survival data. *Scandinavian Journal of Statistics*, 44(1):112–129, 2017.
- [68] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020.
- [69] Liangxiao Jiang, Hao Zhang, Fangna Tao, and Chaoqun Li. Learning from crowds with multiple noisy label distribution propagation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11):6558–6568, 2021.
- [70] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pages 2304–2313. PMLR, 2018.
- [71] Zhimeng Jiang, Kaixiong Zhou, Zirui Liu, Li Li, Rui Chen, Soo-Hyun Choi, and Xia Hu. An information fusion approach to learning with instance-dependent label noise. In *International Conference on Learning Representations*, 2022.
- [72] Marshall M Joffe. Administrative and artificial censoring in censored regression models. *Statistics in medicine*, 20(15):2287–2304, 2001.
- [73] Manas Joglekar, Hector Garcia-Molina, and Aditya Parameswaran. Evaluating the crowd with confidence. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 686–694, 2013.
- [74] Alistair E W Johnson, David J Stone, Leo A Celi, and Tom J Pollard. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39, 2018.
- [75] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

- [76] Annahieta Kalantari, Haney Mallemat, and Scott D Weingart. Sepsis definitions: the search for gold and what cms got wrong. *Western Journal of Emergency Medicine*, 18(5):951, 2017.
- [77] Fahad Kamran, Donna Tjandra, Andrew Heiler, Jessica Virzi, Karandeep Singh, Jessie E King, Thomas S Valley, and Jenna Wiens. Evaluation of sepsis prediction models before onset of treatment. *NEJM AI*, 1(3), 2024.
- [78] Fahad Kamran and Jenna Wiens. Estimating calibrated individualized survival curves with deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 240–248, 2021.
- [79] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.
- [80] David Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowd-sourcing systems. *Advances in neural information processing systems*, 24, 2011.
- [81] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.
- [82] Taghi M Khoshgoftaar and Pierre Rebours. Generating multiple noise elimination filters with the ensemble-partitioning filter. In *Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, 2004. IRI 2004.*, pages 369–375. IEEE, 2004.
- [83] Taehyeon Kim, Jongwoo Ko, JinHwan Choi, Se-Young Yun, et al. Fine samples for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34, 2021.
- [84] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [85] Jacqueline C Kirby, Peter Speltz, Luke V Rasmussen, Melissa Basford, Omri Gottesman, Peggy L Peissig, Jennifer A Pacheco, Gerard Tromp, Jyotishman Pathak, David S Carrell, et al. Phekb: a catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association*, 23(6):1046–1052, 2016.
- [86] Thomas D Koepsell and Sarah E Monsell. Reversion from mild cognitive impairment to normal or near-normal cognition: risk factors and prognosis. *Neurology*, 79(15):1591–1598, 2012.
- [87] Olga Kostopoulou, Brendan C Delaney, and Craig W Munro. Diagnostic difficulty and error in primary care—a systematic review. *Family practice*, 25(6):400–413, 2008.

- [88] Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. Hdltext: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 364–371. IEEE, 2017.
- [89] Stefan Kramer, Gerhard Widmer, Bernhard Pfahringer, and Michael De Groeve. Prediction of ordinal classes using regression trees. *Fundamenta Informaticae*, 47(1-2):1–13, 2001.
- [90] Richard L Kravitz, Naihua Duan, and Joel Braslow. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly*, 82(4):661–687, 2004.
- [91] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of machine learning research*, 20(129):1–30, 2019.
- [92] Martin Ladouceur, Elham Rahme, Christian A Pineau, and Lawrence Joseph. Robustness of prevalence estimates derived from misclassified data from administrative databases. *Biometrics*, 63(1):272–279, 2007.
- [93] Ezio Lanza, Riccardo Muglia, Isabella Bolengo, Dario Poretti, Felice D’Antuono, Roberto Ceriani, Guido Torzilli, and Vittorio Pedicini. Survival analysis of 230 patients with unresectable hepatocellular carcinoma treated with bland transarterial embolization. *PloS one*, 15(1):e0227711, 2020.
- [94] Martin G Larson. Covariate analysis of competing-risks data with log-linear models. *Biometrics*, pages 459–469, 1984.
- [95] Changhee Lee, Jinsung Yoon, and Mihaela Van Der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2019.
- [96] Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [97] Dong-Hyun 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*, volume 3, page 896. Atlanta, 2013.
- [98] Jingzheng Li, Hailong Sun, and Jiyi Li. Beyond confusion matrix: learning from multiple annotators with awareness of instance features. *Machine Learning*, 112(3):1053–1075, 2023.
- [99] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

- [100] Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, and Tongliang Liu. Estimating noise transition matrix with label correlations for noisy multi-label learning. In *Advances in Neural Information Processing Systems*, 2022.
- [101] Holger A Lindner, Shigehiko Schamoni, Thomas Kirschning, Corinna Worm, Bianka Hahn, Franz-Simon Centner, Jochen J Schoettler, Michael Hagmann, Jörg Krebs, Dennis Mangold, et al. Ground truth labels challenge the validity of sepsis consensus definitions in critical illness. *Journal of Translational Medicine*, 20(1):27, 2022.
- [102] Jane C Lindsey and Louise M Ryan. Methods for interval-censored data. *Statistics in medicine*, 17(2):219–238, 1998.
- [103] Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. In *International Conference on Machine Learning*. PMLR, 2022.
- [104] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.
- [105] Yang Liu. Understanding instance-level label noise: Disparate impacts and treatments. In *International Conference on Machine Learning*, pages 6725–6735. PMLR, 2021.
- [106] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*, pages 6226–6236. PMLR, 2020.
- [107] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*. PMLR, 2020.
- [108] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International Conference on Machine Learning*, pages 6543–6553. PMLR, 2020.
- [109] Amalia S Magaret. Incorporating validation subsets into discrete proportional hazards models for mismeasured outcomes. *Statistics in Medicine*, 27(26):5456–5470, 2008.
- [110] Salim Malakouti and Milos Hauskrecht. Hierarchical adaptive multi-task learning framework for patient diagnoses and diagnostic category classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 701–706. IEEE, 2019.
- [111] Nathan Mantel et al. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50(3):163–170, 1966.
- [112] Naresh Manwani and PS Sastry. Noise tolerance under risk minimization. *IEEE transactions on cybernetics*, 43(3):1146–1151, 2013.

- [113] Nancy N Maserejian, Carol L Link, Karen L Lutfey, Lisa D Marceau, and John B McKinlay. Disparities in physicians’ interpretations of heart disease symptoms by patient gender: results of a video vignette factorial experiment. *Journal of women’s health*, 18(10):1661–1667, 2009.
- [114] Amalia S Meier, Barbra A Richardson, and James P Hughes. Discrete proportional hazards models for mismeasured outcomes. *Biometrics*, 59(4):947–954, 2003.
- [115] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised hierarchical text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6826–6833, 2019.
- [116] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*, 2019.
- [117] Michael Moor, Max Horn, Bastian Rieck, Damian Roqueiro, and Karsten 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, pages 2–26. PMLR, 2019.
- [118] Henry A Nasrallah. Consequences of misdiagnosis: inaccurate treatment and poor patient outcomes in bipolar disorder. *The Journal of clinical psychiatry*, 76(10):27608, 2015.
- [119] Wayne Nelson. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1):27–52, 1969.
- [120] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. In *International Conference on Learning Representations*, 2020.
- [121] Jenna M Norton, Kaltun Ali, Claudine T Jurkovitz, Krzysztof Kiryluk, Meyeon Park, Kensaku Kawamoto, Ning Shang, Sankar D Navaneethan, Andrew S Narva, and Paul Drawz. Development and validation of a pragmatic electronic phenotype for ckd. *Clinical Journal of the American Society of Nephrology*, 14(9):1306–1314, 2019.
- [122] Eric J Oh, Bryan E Shepherd, Thomas Lumley, and Pamela A Shaw. Improved generalized raking estimators to address dependent covariate and failure-time outcome error. *Biometrical Journal*, 63(5):1006–1027, 2021.
- [123] Eric J Oh, Bryan E Shepherd, Thomas Lumley, and Pamela A Shaw. Raking and regression calibration: Methods to address bias from correlated covariate and time-to-event error. *Statistics in medicine*, 40(3):631–649, 2021.
- [124] Jeeheh Oh, Jiaxuan Wang, Shengpu Tang, Michael W. Sjoding, and Jenna Wiens. Relaxed parameter sharing: Effectively modeling time-varying relationships in clinical time-series. In *MLHC*, 2019.

- [125] Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639, 2005.
- [126] Romina Pace, Tricia Peters, Elham Rahme, and Kaberi Dasgupta. Validity of health administrative database definitions for hypertension: a systematic review. *Canadian Journal of Cardiology*, 33(8):1052–1059, 2017.
- [127] Natalie Parde and Rodney 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*, pages 1907–1912, 2017.
- [128] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [129] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.
- [130] Mengjiao Peng, Liming Xiang, and Shanshan Wang. Semiparametric regression analysis of clustered survival data with semi-competing risks. *Computational Statistics & Data Analysis*, 124:53–70, 2018.
- [131] M Pérez-Ortiz, Manuel Cruz-Ramírez, María Dolores Ayllón-Terán, N Heaton, Rubén Ciria, and César Hervás-Martínez. An organ allocation system for liver transplantation based on ordinal regression. *Applied Soft Computing*, 14:88–98, 2014.
- [132] Stephen R Pfohl, Tony Duan, Daisy Yi Ding, and Nigam H Shah. Counterfactual reasoning for fair clinical risk prediction. In *Machine Learning for Healthcare Conference*, pages 325–358. PMLR, 2019.
- [133] Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of biomedical informatics*, 113:103621, 2021.
- [134] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11557–11568, 2021.
- [135] Susan K Pingleton. Complications of acute respiratory failure. *Am Rev Respir Dis*, 137(6):1463–1493, 1988.
- [136] Kenneth H Pollock, Scott R Winterstein, Christine M Bunck, and Paul D Curtis. Survival analysis in telemetry studies: the staggered entry design. *The Journal of Wildlife Management*, pages 7–15, 1989.

- [137] Shi-ang Qi, Neeraj Kumar, Mahtab Farrokh, Weijie Sun, Li-Hao Kuan, Rajesh Ranganath, Ricardo Henao, and Russell Greiner. An effective meaningful way to evaluate survival models. In *ICML*, 2023.
- [138] Daniel Rabinowitz, Anastasios Tsiatis, and Jorge Aragon. Regression with interval-censored data. *Biometrika*, 82(3):501–513, 1995.
- [139] Amy H Racine-Poon and David G Hoel. Nonparametric estimation of the survival function when cause of death is uncertain. *Biometrics*, pages 1151–1158, 1984.
- [140] Rüdiger Rackwitz. Reliability analysis—a review and some perspectives. *Structural safety*, 23(4):365–395, 2001.
- [141] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.
- [142] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access, 2017.
- [143] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda 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*, pages 889–896, 2009.
- [144] Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. Deep recurrent survival analysis. In *Proc. AAAI*, pages 1–8, 2019.
- [145] Chanu Rhee, Raymund Barretto Dantes, Lauren Epstein, and Michael 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.
- [146] Chanu Rhee, Travis M Jones, Yasir Hamad, Anupam Pande, Jack Varon, Cara O’Brien, Deverick J Anderson, David K Warren, Raymund B Dantes, Lauren Epstein, et al. Prevalence, underlying causes, and preventability of sepsis-associated mortality in us acute care hospitals. *JAMA network open*, 2(2):e187571–e187571, 2019.
- [147] Chanu Rhee, Zilu Zhang, Sameer S Kadri, David J Murphy, Greg S Martin, Elizabeth Overton, Christopher W Seymour, Derek C Angus, Raymund Dantes, Lauren Epstein, et al. Sepsis surveillance using adult sepsis events simplified esofa criteria versus sepsis-3 sofa criteria. *Critical care medicine*, 47(3):307, 2019.
- [148] Filipe Rodrigues, Mariana Lourenco, Bernardete Ribeiro, and Francisco C Pereira. Learning supervised topic models for classification and regression from crowds. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2409–2422, 2017.

- [149] Matthew T Roe, Paul W Armstrong, Keith AA Fox, Harvey D White, Dorairaj Prabhakaran, Shaun G Goodman, Jan H Cornel, Deepak L Bhatt, Peter Clemmensen, Felipe Martinez, et al. Prasugrel versus clopidogrel for acute coronary syndromes without revascularization. *New England Journal of Medicine*, 367(14):1297–1309, 2012.
- [150] C Yu Sean, Kevin D Betthauser, Aditi Gupta, Patrick G Lyons, Albert M Lai, Marin H Kollef, Philip RO Payne, and Andrew P Michelson. Comparison of sepsis definitions as automated criteria. *Critical care medicine*, 49(4):e433–e443, 2021.
- [151] Paul Hongsuck Seo, Geeho Kim, and Bohyung Han. Combinatorial inference against label noise. In *Advances in Neural Information Processing Systems*, pages 1173–1183, 2019.
- [152] Burr Settles. Active learning literature survey. 2009.
- [153] Christopher W Seymour, Vincent X Liu, Theodore J Iwashyna, Frank M Brunkhorst, Thomas D Rea, André Scherag, Gordon Rubenfeld, Jeremy M Kahn, Manu Shankar-Hari, Mervyn Singer, et al. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):762–774, 2016.
- [154] Manu Shankar-Hari, Gary S Phillips, Mitchell L Levy, Christopher W Seymour, Vincent X Liu, Clifford S Deutschman, Derek C Angus, Gordon D Rubenfeld, Mervyn 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*, 315(8):775–787, 2016.
- [155] Victor S Sheng and Jing Zhang. Machine learning with crowdsourcing: A brief summary of the past research and future directions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9837–9843, 2019.
- [156] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.
- [157] Aarti Singh, Robert Nowak, and Jerry Zhu. Unlabeled data: Now it helps, now it doesn’t. *Advances in neural information processing systems*, 21, 2008.
- [158] Steven M Snapinn. Survival analysis with uncertain endpoints. *Biometrics*, pages 209–218, 1998.
- [159] Scott D Solomon, Adel R Rizkala, Jianjian Gong, Wenyan Wang, Inder S Anand, Junbo Ge, Carolyn SP Lam, Aldo P Maggioni, Felipe Martinez, Milton Packer, et al. Angiotensin receptor neprilysin inhibition in heart failure with preserved ejection fraction: rationale and design of the paragon-hf trial. *JACC: Heart Failure*, 5(7):471–482, 2017.

- [160] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [161] Derek H Suite, Robert La Bril, Annette Primm, and Phyllis Harrison-Ross. Beyond misdiagnosis, misunderstanding and mistrust: relevance of the historical perspective in the medical and mental health treatment of people of color. *Journal of the National Medical Association*, 99(8):879, 2007.
- [162] Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 2020.
- [163] Mani Thenmozhi, Visalakshi Jeyaseelan, Lakshmanan Jeyaseelan, Rita Isaac, and Rupa Vedantam. Survival analysis in longitudinal studies for recurrent events: applications and challenges. *Clinical Epidemiology and Global Health*, 7(2):253–260, 2019.
- [164] Jaree Thongkam, Guandong Xu, Yanchun Zhang, and Fuchun Huang. Support vector machine for outlier detection in breast cancer survivability prediction. In *Asia-Pacific Web Conference*, pages 99–109. Springer, 2008.
- [165] Tian Tian and Jun Zhu. Max-margin majority voting for learning from crowds. *Advances in neural information processing systems*, 28, 2015.
- [166] Donna Tjandra, Yifei He, and Jenna Wiens. A hierarchical approach to multi-event survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 591–599, 2021.
- [167] Donna Tjandra, Raymond Q Migrino, Bruno Giordani, and Jenna Wiens. Cohort discovery and risk stratification for alzheimer’s disease: an electronic health record-based approach. *Alzheimer’s & Dementia: Translational Research & Clinical Interventions*, 6(1):e12035, 2020.
- [168] Sudhi G Upadhyaya, Dennis H Murphree Jr, Che G Ngufor, Alison M Knight, Daniel J Cronk, Robert R Cima, Timothy B Curry, Jyotishman Pathak, Rickey E Carter, and Daryl J Kor. Automated diabetes case identification using electronic health record data at a tertiary care facility. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, 1(1):100–110, 2017.
- [169] Arjun K Venkatesh, Todd Slesinger, Jessica Whittle, Tiffany Osborn, Emily Aaronson, Craig Rothenberg, Nalani Tarrant, Pawan Goyal, Donald M Yealy, and Jeremiah D Schuur. Preliminary performance on the new cms sepsis-1 national quality measure: early insights from the emergency quality network (e-quality). *Annals of emergency medicine*, 71(1):10–15, 2018.
- [170] Sofie Verbaeten. Identifying mislabeled training examples in ilp classification problems. In *Proceedings of twelfth Belgian-Dutch conference on machine learning*, pages 1–8, 2002.

- [171] Jan Verwaeren, Willem Waegeman, and Bernard De Baets. Learning partial ordinal class memberships with kernel-based proportional odds models. *Computational Statistics & Data Analysis*, 56(4):928–942, 2012.
- [172] David M Vock, Julian Wolfson, Sunayan Bandyopadhyay, Gediminas Adomavicius, Paul E Johnson, Gabriela Vazquez-Benitez, and Patrick 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*, 61:119–131, 2016.
- [173] Deng-Bao Wang, Yong Wen, Lujia Pan, and Min-Ling Zhang. Learning from noisy labels with complementary loss functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10111–10119, 2021.
- [174] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 526–536, 2021.
- [175] Lu Wang, Yan Li, Jiayu Zhou, Dongxiao Zhu, and Jieping Ye. Multi-task survival analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 485–494. IEEE, 2017.
- [176] Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.
- [177] Qizhou Wang, Bo Han, Tongliang Liu, Gang Niu, Jian Yang, and Chen Gong. Tackling instance-dependent label noise via a universal probabilistic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10183–10191, 2021.
- [178] Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Yang Liu. To smooth or not? when label smoothing meets noisy labels. In *International Conference on Machine Learning*. PMLR, 2022.
- [179] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.
- [180] Lee-Jen Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992.
- [181] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23, 2010.
- [182] Andrew Wong, Erkin Otles, John P Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penzo, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8):1065–1070, 2021.

- [183] Shannon Wongvibulsin, Katherine C Wu, and Scott L Zeger. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (rf-slam) data analysis. *BMC Medical Research Methodology*, 20(1):1–14, 2020.
- [184] Songhua Wu, Mingming Gong, Bo Han, Yang Liu, and Tongliang Liu. Fair classification with instance-dependent label noise. In *Conference on Causal Learning and Reasoning*, pages 927–943. PMLR, 2022.
- [185] Yichen Wu, Jun Shu, Qi Xie, Qian Zhao, and Deyu Meng. Learning to purify noisy labels via meta soft label corrector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10388–10396, 2021.
- [186] Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels. In *International Conference on Learning Representations*, 2022.
- [187] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33, 2020.
- [188] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *Advances in neural information processing systems*, pages 6838–6849, 2019.
- [189] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.
- [190] Jie Xu, Yunyu Xiao, Wendy Hui Wang, Yue Ning, Elizabeth A Shenkman, Jiang Bian, and Fei Wang. Algorithmic fairness in computational medicine. *EBioMedicine*, 84:104250, 2022.
- [191] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L<sub>dmi</sub>: An information-theoretic noise-robust loss function. *Advances in Neural Information Processing Systems*, 32, 2019.
- [192] Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise. In *International Conference on Machine Learning*. PMLR, 2022.
- [193] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. Mining electronic health records (ehrs) a survey. *ACM Computing Surveys (CSUR)*, 50(6):1–40, 2018.
- [194] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Tin-Yau Kwok. Searching to exploit memorization effect in learning with noisy labels. In *International Conference on Machine Learning*, pages 10789–10798. PMLR, 2020.

- [195] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in Neural Information Processing Systems*, 33, 2020.
- [196] Vithya Yogarajan, Gillian Dobbie, Sharon Leitch, Te Taka Keegan, Joshua Bensemann, Michael Witbrock, Varsha Asrani, and David Reith. Data and model bias in artificial intelligence for healthcare applications in new zealand. 2023.
- [197] Jung Eun Yoo, Dong Wook Shin, Kyungdo Han, Dahye Kim, Seung-Pyo Lee, Su-Min Jeong, Jinkook Lee, and SangYun Kim. Blood pressure variability and the risk of dementia: a nationwide cohort study. *Hypertension*, 75(4):982–990, 2020.
- [198] Ji Won Yoon, Stephen J Roberts, Mathew Dyson, and John Q Gan. Bayesian inference for an adaptive ordered probit model: An application to brain computer interfacing. *Neural Networks*, 24(7):726–734, 2011.
- [199] Shelemyahu Zacks. *Introduction to reliability analysis: probability models and statistical methods*. Springer Science & Business Media, 2012.
- [200] Quan Zhang and Mingyuan Zhou. Nonparametric bayesian lomax delegate racing for survival analysis with competing risks. In *Advances in Neural Information Processing Systems*, pages 5002–5013, 2018.
- [201] Xuchao Zhang, Xian Wu, Fanglan Chen, Liang Zhao, and Chang-Tien Lu. Self-paced robust learning for leveraging clean labels in noisy data. In *AAAI*, pages 6853–6860, 2020.
- [202] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.
- [203] Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris Metaxas, and Chao Chen. Error-bounded correction of noisy labels. In *International Conference on Machine Learning*, pages 11447–11457. PMLR, 2020.
- [204] Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In *International Conference on Machine Learning*, pages 27412–27427. PMLR, 2022.
- [205] Zhaowei Zhu, Jialu Wang, and Yang Liu. Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. In *International Conference on Machine Learning*. PMLR, 2022.
- [206] Kathryn Ziegler-Graham, Ron Brookmeyer, Elizabeth Johnson, and H Michael Arrighi. Worldwide variation in the doubling time of alzheimer’s disease incidence rates. *Alzheimer’s & Dementia*, 4(5):316–323, 2008.