Deep Survival Modeling for Personalized Prognosis and Treatment Optimization

Mingzhu Liu

CMU-RI-TR-25-71

July 31, 2025

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania

Thesis Committee:
Dr. Artur Dubrawski, *chair*
Dr. George H. Chen
Angela Chen

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Robotics.

Abstract

Prognostic modeling from medical data holds the promise of informing personalized care and improving clinical decision-making. This thesis explores two applications of survival analysis and deep learning to estimate long-term patient risk and treatment benefit in high-impact cardiopulmonary settings. In the first study, we develop a deep multimodal time-to-event prediction framework that estimates patient-specific mortality risk using chest radiographs and demographic features. Unlike traditional binary classifiers, our approach, leveraging models such as Cox proportional hazards and deep survival machines, accounts for right-censoring and allows for risk estimation at arbitrary time horizons, offering greater flexibility and clinical utility. In the second study, we apply individualized treatment effect estimation to determine which patients with stable ischemic heart disease are most likely to benefit from coronary artery bypass grafting (CABG). Using a recently proposed machine learning algorithm, Cox Mixtures with Heterogeneous Effects (CMHE), we stratify patients based on their predicted survival gain from CABG versus optimal medical therapy alone and validate these predictions on an external surgical cohort. Together, these works demonstrate the potential of survival-based machine learning models to enhance personalized risk prediction and treatment optimization in real-world clinical scenarios.

Contents

When this dissertation is viewed as a PDF, the page header is a link to this Table of Contents.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Accurate prognostication and personalized treatment planning are critical goals in modern medicine. As healthcare increasingly adopts data-driven approaches, there is growing interest in using machine learning (ML) to support clinical decision-making, particularly in high-stakes domains such as cardiopulmonary health. Despite recent advances, existing predictive models often fall short in two important respects: they are typically limited to binary outcomes at fixed time points, and they rarely account for individualized treatment effects. These limitations hinder the ability to provide flexible, nuanced, and personalized guidance to clinicians and patients.

This thesis is motivated by two representative challenges in clinical prediction and decision-making. First, in risk prediction from medical imaging, deep learning models have shown impressive performance but are often trained to predict static binary labels, such as mortality within a certain number of years, without properly handling censoring (individuals being lost to follow-up) or time-varying covariates (measurements or observations changing over time). This rigid formulation fails to reflect the continuous nature of disease progression and patient trajectories in real-world settings. Furthermore, predictions from such models may not be reliable across different populations or time horizons, limiting their clinical utility. Second, in treatment decision-making, conventional risk scores such as the Society of Thoracic Surgeons (STS) scores are valuable but provide average population-level risk estimates.

They do not estimate the causal effect of a treatment for a particular patient. As a result, patients who could benefit substantially from a procedure such as coronary artery bypass grafting (CABG) may be overlooked due to the high perceived operative risk, even when long-term survival benefit is plausible.

To overcome these limitations, survival analysis provides a powerful framework by modeling time as a continuous random variable and explicitly accounting for censoring, which is common in real-world clinical data. Unlike traditional binary classifiers that predict whether an event will occur by a fixed time point, survival models estimate the probability of an event occurring over time, offering a more flexible and informative perspective on patient risk. One of the most well-known survival models is the Cox Proportional Hazards (CPH) model [12], which estimates the hazard, or instantaneous event rate, for individuals based on their covariates. However, the CPH model assumes that the hazard ratios between individuals are constant over time, an assumption that may not hold in many clinical scenarios.

To address this, a wide variety of alternative survival models have been proposed that do not rely on the proportional hazards assumption. Classical approaches such as the conditional Kaplan–Meier estimator [7] and Random Survival Forests [28] offer nonparametric, flexible alternatives for survival prediction. More recently, deep learning-based methods have emerged to model complex, nonlinear relationships in time-to-event data. These include the Neural Multi-Task Logistic Regression (N-MTLR) model [16], Nnet-survival [18], Cox-Time [33], and Deep Survival Machines [40]. Collectively, these models enable more personalized, dynamic, and clinically relevant predictions, making survival analysis a compelling alternative to fixed-horizon classifiers, particularly in high-stakes applications like risk stratification and treatment planning in cardiopulmonary care.

This thesis is motivated by the need to develop survival-based machine learning methods that model time-to-event outcomes, incorporate censoring, and provide personalized predictions for both prognosis and treatment response. Using rich multimodal data, such as chest radiographs and structured clinical variables, these models aim to enable more flexible, individualized, and actionable decision support tools that align with the principles of precision medicine.

1.2 Thesis Organization

The remainder of this thesis is structured as follows.

1. Chapter 2 introduces deep survival models for chest radiographs, detailing the methodology, key challenges, contributions, and experimental validation [35].
2. Chapter 3 presents deep survival models for identifying phenogroups, detailing key findings.
3. Chapter 4 concludes with a summary of contributions and future research opportunities.

Chapter 2

Deep Survival Models Can Improve Long-Term Mortality Risk Estimates from Chest Radiographs

2.1 Introduction

Chest radiography is a common diagnostic tool for lung disorders and offers prognostic value in terms of overall patient health and well-being. The ability to prognosticate patient susceptibility to an adverse outcome such as death or serious complications using such inexpensive medical testing is therefore of immense interest. Indeed, recent advances in deep learning [21] have demonstrated that, e.g., convolutional neural network (CNN) approaches perform well in diagnosing conditions from chest radiograms, in some cases even outperforming trained radiologists [6, 51, 52, 53, 58, 60].

The ability to consolidate the likelihood of adverse outcomes into a single scalar risk estimate is an exciting prospect as such a score could be utilized to stratify patients based on their risk levels and to prioritize subsequent treatment and interventions, helping improve overall longevity. However, deep learning models are commonly trained to predict multiple adverse diagnoses [5], and consolidating those individual diagnoses into a singular risk score or estimate can be a challenging task.

In studies that involve risk estimation from medical imaging data [10, 32, 36, 38,

48, 50], the time-to-event outcome is binarized as an indicator of whether a patient survived beyond a certain point in time (such as a 5-year risk), and imaging data from individuals who were lost to follow-up (censored) are discarded when training the model. However, long-term patient risk levels might vary over time depending on patient physiology, medical history, and demographics. This can happen as a result of time-varying covariates, where some measurements or observations change over time. For example, in studying cancer risk, smoking status can be a time-varying covariate [61]. As a result, modeling patients' risk at a fixed time in the future might be insufficient to obtain an overall sense of their prognosed health and well-being. Furthermore, in real-world observational studies, patients cannot be followed up indefinitely due to resource constraints and study design. A large number of patients thus might be lost to follow-up before an event of interest is observed in them, and typical binary classifiers ignore these observations. Moreover, a binary classifier is only able to make predictions at the fixed time horizon that is chosen to be the threshold for the binary targets, which makes it less flexible in that if a user wants the patient risk at a different time, the model needs to be trained again on the new binary target.

Deep survival analysis and time-to-event prediction aim to address those limitations by modeling time as a continuous random variable and incorporating censoring in the estimation of patient risk. Survival analysis is a class of statistical methods for estimating the time until an event occurs. As opposed to binary classification, the time-to-event outcome for a particular individual is considered to be a continuous random variable. In this study, we were interested in estimating the risk of a patient experiencing mortality within a time horizon, given a chest radiograph and a set of demographic covariates. We assumed that an observation can be right-censored, which means that the observation of a subject can be terminated before death occurs.

Cox proportional hazards (CPH) [12] is arguably one of the most popular approaches for analyzing time-to-event survival data. The CPH model makes a proportional hazards assumption, stating that the ratio between the hazard rates of two individuals is constant over all time horizons (please refer to Appendix A.2 for a more detailed description). In real-world application scenarios, however, the proportional hazards assumption may be overly restrictive or not hold at all. In response, a wide range of alternative survival models have been developed over the past decades

that do not rely on this assumption. Classical approaches, such as the conditional Kaplan-Meier estimator [7] and Random Survival Forests [28], offer nonparametric or ensemble-based alternatives. More recent deep learning models that can capture nonlinear elements from the data, such as the Neural Multi-Task Logistic Regression (N-MTLR) model [16], Nnet-survival [18], and Cox-Time [33], have further relaxed the proportional hazards constraint by allowing for time-varying effects or modeling time explicitly in the relative risk function. It is possible to cast survival analysis problems as classification problems through survival stacking [13]. We adopt the Deep Survival Machines (DSM) [40], which models the survival distribution as a mixture of learnable components (see Appendix A.3 for details).

In this paper, we examine a multimodal application of time-to-event data to predict long-term patient risk from chest radiographs and patient demographic data. In particular, we use models such as CPH and DSM as replacements for binary models. We aim to demonstrate that the proposed approach is able to generate predictions at any time horizon while achieving superior discriminative performance and calibration. The source code is available at https://github.com/autonlab/Deep_Chest_Survival.

2.2 Materials and Methods

The dataset used in this study is from the randomized controlled Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer screening trial [2]. The goal of the trial was to investigate whether screening has an effect on cancer-related mortality.

At 10 screening centers nationwide, 154,934 participants were enrolled between November 1993 and September 2001. One coordinating center performed data management and trial coordination. Each arm involved 37,000 females and 37,000 males aged 55–74 at entry.

In the control arm, participants received their usual medical care. In the intervention arm, women received chest X-rays, flexible sigmoidoscopy, CA125, and transvaginal ultrasound (TVU), and men received chest X-rays, flexible sigmoidoscopy, serum prostate-specific antigen (PSA), and digital rectal exams (DREs). PSA and CA125 were performed at entry, then annually for 5 years. DRE, TVU, and chest X-ray exams were performed at entry, then annually for 3 years. Only two annual

repeat X-ray exams were performed for non-smokers. Sigmoidoscopy was performed at entry, then at the 5-year point.

A radiologist interpreted each posteroanterior chest X-ray and radiologists' impressions were included in the dataset. Participants were followed for at least 13 years. Deaths during the trial were confirmed by mailed annual study update (ASU) and linkage to the National Death Index (NDI). Causes of death during the trial were also determined.

Data were collected on forms designed for an NCS OpScan 5 optical-mark reader, and then automatically loaded into a database. Information collected at screening centers went through a hub through modems using common carrier lines. Additional details about the PLCO dataset can be found in [23, 25, 49].

The dataset includes radiologists' impressions that indicate whether abnormalities exist. The survival curves based on the time-to-death data of each patient are shown in Figure 2.1. For demonstration, we selected the noted presence of bone/soft tissue lesions, cardiac abnormalities, COPD, granuloma, pleural fibrosis, pleural fluid, scarring, radiographic abnormalities, mass, pleural, nodule, atelect, hilar, infiltrate, and other abnormalities. The curves are stratified by whether the mentioned abnormalities and lung cancer were present at the beginning of the study.

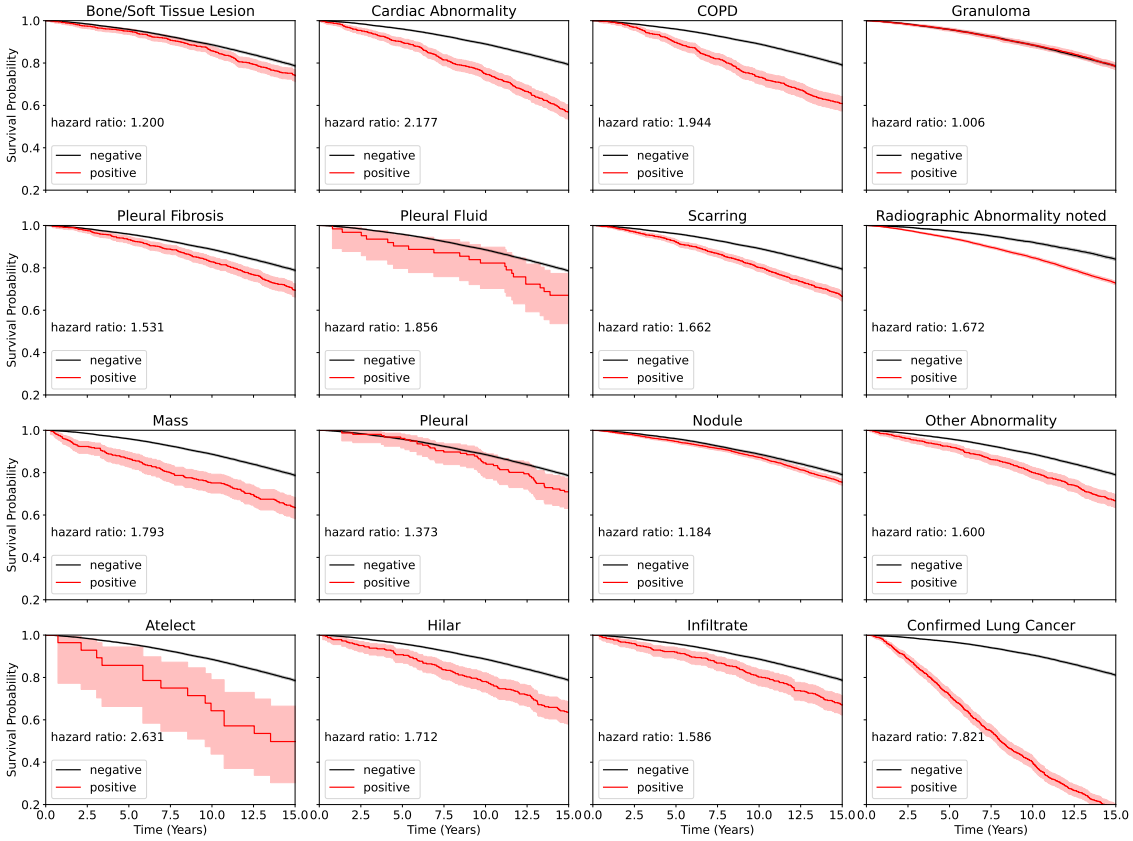


Figure 2.1: Survival curves plotted with Kaplan–Meier survival estimates for phenogroups, stratified by abnormalities.

2.2.1 Preprocessing

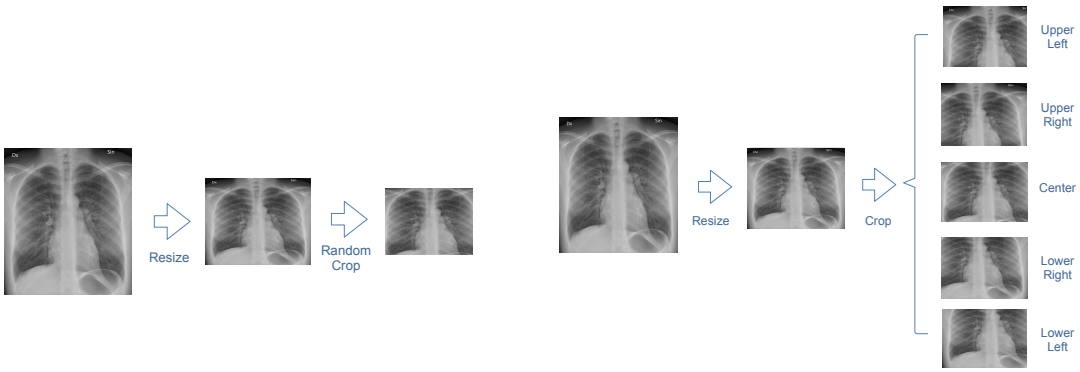
In this study, we used grayscale chest radiographs, demographic data, time-to-death, and censor indicators. Demographic data record sex, race, ethnicity, cigarette smoking status, and age. Age was defined to be the age at the time of screening. Time-to-event was defined as the number of days from the chest X-ray screen to the day of death or the last known day of life (censoring time). Censor indicators were binary variables, where 0 indicated that the outcome was censored, and 1 indicated that death was observed.

In total, 89,643 images were used; 60% were used for training the model, 20% were held out as the validation data for optimizing model hyperparameters, and 20% were held out for testing and final model evaluation. We ensured that images from the

same patient did not appear in more than one of the training, testing, or validation subsets of data.

An illustration of the image preprocessing procedure is shown in Figure 2.2. The size of each image is 2500 by 2100 pixels. In order to reduce computational costs, each chest radiograph was resized to a 256 by 256-pixel image. Images in the training dataset were cropped to 224 by 224 pixels at a random position. Images in the testing dataset were cropped at the four corners and the center. Predictions were obtained for each crop independently and averaged to obtain a final prediction for the entire image.

Statistics of data from the subset of the PLCO database used in this study are summarized in Table 2.1. Samples with negative times were discarded.



(a) Images in the training dataset were resized and cropped at a random position.

(b) Images in the test dataset were resized and cropped at the corners and the center.

Figure 2.2: An illustration of the resizing and cropping procedure.

Characteristic	Number
Number of patients	25,433
Number of samples (images)	89,643
Average number of samples (images) per patient	3.52
% Censored outcomes	67.40%
Age (25th–75th percentiles)	65.5 ± 7.5
% White patients	12.68%
% Patients who smoke	79.19%
% Patients surviving beyond 10 years	83.18%
% Patients with confirmed primary invasive lung cancer diagnosed during the trial	4.07%
Time-to-event (years)	15.60 ± 4.33

Table 2.1: Statistics of the subset of PLCO data used in this study.

2.2.2 Evaluation Metrics

We evaluated the performance of the considered models using the Brier score [8], the concordance index based on the inverse probability of censoring weights [56], the cumulative/dynamic area under the receiver operating characteristic curve (AUC) [27, 55], and the expected calibration error (ECE) [47] for 2-year, 5-year, and 10-year time horizons. Additional details on evaluation metrics can be found in Appendix A.4. Bootstrapping with 100 samples was performed to obtain 95% confidence intervals. Samples were drawn from test-set predictions with 100 replacements.

The Brier score measures the average squared distance between the actual survival probability and the predicted survival probability for a given time horizon. The higher the Brier score, the worse the discriminative performance of the model.

The concordance index measures discriminative performance in terms of the probability of assigning higher risk to subjects with shorter times-to-events by calculating the ratio of the number of correctly ordered pairs of patients, based on their actual

outcomes, to the total number of such pairs. The concordance index, based on the inverse probability of censoring weights, uses the Kaplan–Meier estimator to estimate the censoring distribution.

The receiver operating characteristic (ROC) curve is typically depicted with the true positive rate as a function of the false positive rate for varying sensitivity thresholds of a binary discrimination or detection model. The higher the AUC, the more accurately ranked the test data instances, i.e., the model gives the positive samples higher scores than the negatives.

The expected calibration error (ECE) is calculated by splitting data sorted by predicted probabilities into subsequent bins and averaging the per-bin differences between predicted probabilities and true probabilities. The lower the ECE, the better the model prediction scores correlate with true probabilities; hence there is less miscalibration.

2.2.3 Methods

In this work, we propose a novel model that extracts survival probabilities from chest X-rays, trained in an end-to-end fashion. Our model consists of an encoder and a survival model that are connected by a multilayer perceptron, as shown in Figure 2.3. The encoder is based on a DenseNet model [26] that takes the chest X-rays as inputs and outputs a numeric vector as the representation of the images. The encoder was adopted from the `TorchXRayVision` Python package [11] and pre-trained on other chest X-ray datasets, including CheXpert and MIMIC-CXR.

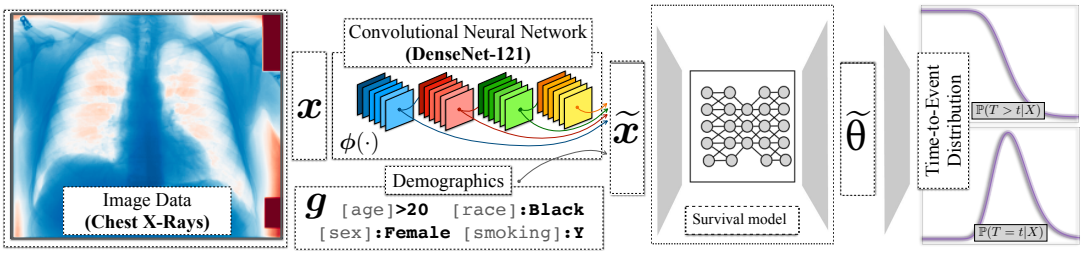


Figure 2.3: Model architecture. The proposed model takes chest radiographs and demographic information as inputs and outputs survival probabilities. The encoder (DenseNet-121) extracts information from the images, which is then combined with demographic information as inputs to the survival model.

The survival models are adopted from the `auton-survival` Python package [43]. A deep neural network version of Cox proportional hazards (DCPH) and DSM are used alternatively. They take the vector representation of chest X-rays and demographic data as inputs and output risk scores for certain time horizons. In this study, we predicted 2-year, 5-year, and 10-year mortality.

We compared the proposed models with a common approach from the literature, such as CXR-risk in Lu et al. [36], which is to train a model to predict mortality for a fixed time horizon. We denote such a model as Thresholded Binary Classification (TBC). In TBC, the time-to-event outcome is binarized to a label representing survival beyond t years from the time the X-ray was obtained, $\mathcal{Y} = \mathbf{1}\{T > t\}$. Specifically, cases where censoring or survival times are longer than t are classified as "survived", and non-censored cases where the times are shorter than or equal to t are classified as "dead". The rest, where the times are shorter than t and censored, are ignored. The number of samples in each category is summarized in Table 2.2.

In order to compare whether survival models were a better alternative than a binary classifier, TBC was built in the same way as Fig. 2.3, except that the survival model was replaced with a multilayer perceptron. The number of layers was tuned with a validation dataset, and additional details are presented in Appendix A.5.

Time (Years)	Survived	Dead	Ignored
2	88,376	907	0
5	85,245	4,398	0
10	77,350	12,293	0
20	4,458	26,560	58,615

Table 2.2: The number of samples in each category.

A hyperparameter search was performed by finding model configurations that minimized the mean Brier score for 2-, 5-, and 10-year time horizons in the validation dataset. Additional implementation details are presented in Appendix A.5.

2.3 Results

The number of patients at risk before different times, defined as those who have censoring or survival times greater than the corresponding time, is shown in Table 2.3.

Time (years)	0	2	5	10	15	20
Number of patients at risk	25,433	25,139	24,373	22,500	18,121	3,155

Table 2.3: Number of patients at risk before 0, 2, 5, 10, 15, and 20 years.

The performance of each model measured with the Brier score, concordance index, AUC, and ECE for 2-year, 5-year, and 10-year time horizons is summarized in Table 2.4. Two-sample t-tests were performed to compare the performance of TBC and DSM, and the performance of TBC and DCPH. Calibration curves plotted with linear and logarithmic scales are shown in Figure 2.4.

Model	**Brier Score**		
	2-year	5-year	10-year
TBC	0.0354 (0.0342, 0.0366)	0.0555 (0.0535, 0.0575)	0.1028 (0.0994, 0.1062)
DSM	**0.0132** (0.0117, 0.0148) 1.9309×10^{-40}	**0.0455** (0.0427, 0.0482) 4.1050×10^{-8}	0.1016 (0.0987, 0.1045) 0.2999
DCPH	**0.0132** (0.0116, 0.0147) 1.3310×10^{-40}	0.0456 (0.0429, 0.0484) 6.8380×10^{-8}	**0.1015** (0.0985, 0.1045) 0.2877
Model	**Concordance Index**		
	2-year	5-year	10-year
TBC	0.7746	0.7562	0.7541

	(0.7473, 0.8018)	(0.7403, 0.7721)	(0.7456, 0.7626)
DSM	**0.7959** (0.7684, 0.8234) 0.1412	**0.7681** (0.7540, 0.7822) 0.1375	0.7595 (0.7504, 0.7685) 0.2002
DCPH	0.7920 (0.7658, 0.8182) 0.1839	0.7658 (0.7522, 0.7794) 0.1853	**0.7604** (0.7511, 0.7697) 0.1647

Model	**AUC**		
	2-year	5-year	10-year
TBC	0.7766 (0.7491, 0.8040)	0.7623 (0.7461, 0.7785)	0.7724 (0.7630, 0.7819)
DSM	**0.7981** (0.7703, 0.8260) 0.1408	**0.7748** (0.7604, 0.7892) 0.1305	0.7782 (0.7684, 0.7880) 0.2049
DCPH	0.7943 (0.7679, 0.8208) 0.1811	0.7725 (0.7587, 0.7863) 0.1749	**0.7793** (0.7695, 0.7890) 0.1643

Model	**ECE**		
	2-year	5-year	10-year
TBC	0.1126 (0.1102, 0.1149)	0.0747 (0.0718, 0.0776)	0.0216 (0.0182, 0.0250)
DSM	0.0052 (0.0038, 0.0067) 5.4185×10^{-90}	0.0110 (0.0080, 0.0141) 2.4773×10^{-51}	0.0152 (0.0119, 0.0185) **0.0047**
DCPH	**0.0043** (0.0032, 0.0054) 5.3225×10^{-93}	**0.0089** (0.0066, 0.0113) 2.5831×10^{-57}	**0.0135** (0.0100, 0.0169) **0.0007**

Table 2.4: Brier score, concordance index, AUC, and ECE evaluated on test dataset with bootstrapping. Best performances are in bold. The 95% confidence intervals are shown in parentheses. TBC (baseline model) performed the worst. P-values for the two-sample t-tests between TBC and DSM and between TBC and DCPH are shown under confidence intervals. p-values lower than 0.05 are in bold.

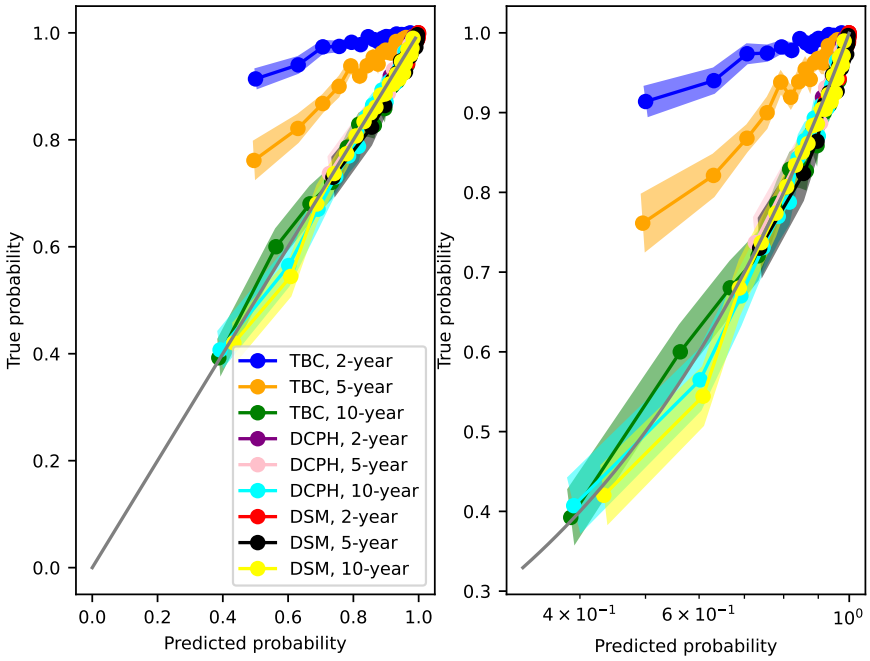


Figure 2.4: Calibration plots for TBC, DCPH, and DSM at 2-year, 5-year, and 10-year time horizons, with 95% confidence intervals (left: linear scale, right: logarithmic scale). TBC calibration at 2-year and 5-year time horizons is significantly worse than the proposed alternatives.

DSM and DCPH perform no worse than TBC with respect to the Brier score, concordance index, AUC, and ECE. In particular, they perform significantly better on shorter-horizon (2-year and 5-year) Brier scores and calibration. The 2-year and 5-year Brier scores of TBC are 0.0354 and 0.0555, whereas those of DSM are 0.0132 and 0.0455 and those of DCPH are 0.0132 and 0.0456. The 2-year, 5-year, and 10-year ECEs of TBC are 0.1126, 0.0747, and 0.0216, whereas those of DSM are 0.0052, 0.0110, and 0.0152 and those of DCPH are 0.0043, 0.0089, 0.0135. The

calibration curves for 2-year and 5-year TBC also significantly deviate from the perfectly calibrated line, compared to other curves. One reason why TBC performs significantly worse at 2-year and 5-year time horizons than at the 10-year time horizon might be that there are significantly more imbalances in the target, as shown in Table 2.2. Techniques for alleviating this problem are available, such as tuning the classification threshold and monotonically transforming predicted probabilities based on isotonic regression [44]. These results suggest that the proposed models offer improved discriminative performance and calibration when compared to a commonly used binary model.

Kaplan–Meier curves obtained using the test dataset are shown in Figure 2.5. The blue curve is obtained by using the ground-truth time-to-event. The rest shows the predicted survival probabilities averaged over the test dataset, using DSM, DCPH, and TBC. The models are evaluated at 1-, 2-, 3-, 5-, 10-, 15-, and 20-year time horizons. DSM and DCPH provide curves that match the ground-truth Kaplan–Meier curve more closely than TBC, especially at longer time horizons. The survival probabilities predicted by TBC deviate from the ground truth much more at the 20-year time horizon than at other time horizons. One reason might be that at the 20-year time horizon, most samples were ignored (Table 2.2).

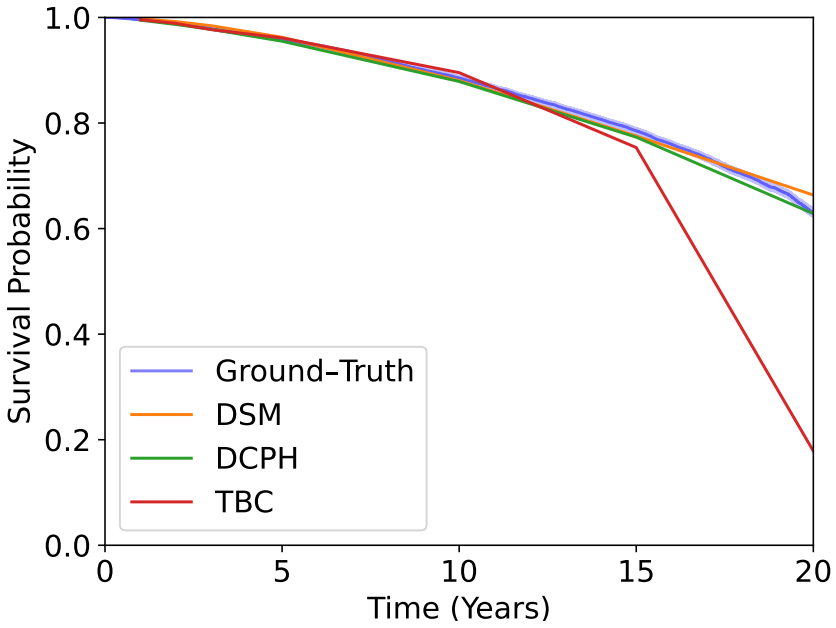


Figure 2.5: The Kaplan–Meier curve obtained from each patient in the test dataset, plotted with the ground–truth time-to-event, and average predicted survival probabilities at each time horizon using TBC, DSM, and DCPH. DSM and DCPH provide more accurate estimations.

DSM does not show significantly improved performance when compared to DCPH. This suggests that, with regard to the dataset we used, the assumption of proportional hazards does not substantially impact predictive power and suggests that there is no substantial prognostic value in non-proportional hazards.

2.4 Discussion

We demonstrated that deep learning survival models that use time-to-event and censor indicators can offer more flexibility and perform better at predicting long-term mortality than commonly used binary classifiers, in terms of Brier score and ECE, on multiple time horizons.

Patient risk prediction from imaging data using deep learning is traditionally performed with a binary mortality outcome variable, not accounting for censored outcomes. Lu et al. [36] demonstrated that CNNs have the prognostic capability to assess a

patient's risk of morality over long time horizons (12 years) in a large cohort of patients from the Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer screening trial [2]. They further validated their estimated risk score on an internal held-out set as well as on an external study, the National Lung Screening Trial (NLST). Other studies applied similar techniques to multiple contexts and various time horizons. [10, 32, 38, 48, 50]. For instance, Kolossváry et al. [32] used deep learning to predict 30-day all-cause mortality using chest radiographs. Raghu et al. [50] estimated postoperative mortality (30-day mortality and in-hospital mortality) after cardiac surgeries using preoperative chest X-rays.

As shown in the TBC study, a model needs to be trained on each of the time horizons of interest. It appears that these models do not necessarily reach the attainable performance, possibly due to not considering censoring information and the time-varying effects of covariates. Survival analysis (DSM and DCPH) can incorporate these factors natively.

CPH is a semiparametric method that requires no predefined probability function to represent survival times, but assumes proportional hazards. It estimates the hazard, which is defined as the instantaneous probability that an event will occur. DeepSurv [30] is a deep learning framework based on CPH. The hazard function in DeepSurv is determined by the parameters of a trained deep neural network.

DSM adds flexibility by modeling the survival distribution as a mixture of multiple component distributions. It assumes that each component distribution is parametric, and does not assume proportional hazards. The parameters of each component and mixture weights are learned from data using neural networks. The individual survival distribution is estimated as an average of the learned distributions weighted by the learned mixture weights.

We expected that DSM would perform better than DCPH, since it imposes fewer assumptions. However, in the specific case of the PLCO dataset, relaxing the assumption of proportional hazards does not appear to yield additional benefits, since the performances of DSM and DCPH are comparable.

Our study has limitations. For instance, as shown in Table 2.1, 50% of patients are between 58 and 73 years of age, and only 12.68% are non-White. Therefore, those who are younger and non-White may not be adequately represented in the training data, resulting in a bias toward those who are older and White, and potentially impacting

the models' effectiveness in promoting patient care uniformly across subpopulations. Further studies are needed to assess the performance of the proposed models when used on populations of different ages and races than those represented in the data used in the current study.

One of our future steps involves the application of the proposed models to counterfactual phenotyping, where groups of individuals belong to underlying clusters and demonstrate heterogeneous treatment effects [42]. This may provide insights into measuring the distinctive effects of interventions on certain subgroups of patients and enable personalized interventions that lead to optimal outcomes.

Chapter 3

Machine Learning Identifies Patients Who Derive Survival Benefit from Coronary Revascularization

3.1 Introduction

Coronary artery bypass grafting (CABG) has been a cornerstone in the treatment of patients with coronary artery disease for decades [17, 59]. However, controversy has arisen about the role of CABG in patients with stable ischemic heart disease [4, 17, 37, 54, 57]. Although CABG is associated with a short-term increased risk of adverse events, including death, a recent, modern meta-analysis suggests a potential survival benefit from CABG plus optimal medical therapy (OMT) versus OMT alone [17]. These results, along with notable advances in CABG patient care that allow low perioperative mortality rates [4, 15], suggest that more scrutiny is warranted to determine which patients are best suited for revascularization with CABG.

Machine learning (ML) is a rapidly emerging technology that can help determine the optimal treatment strategy for patients with coronary artery disease [45]. A recently developed ML algorithm, Cox Mixtures with Heterogeneous Effects (CMHE),

utilizes survival data to estimate treatment effects in individual patients [41]. In particular, the algorithm uses deep learning architectures to stratify patients according to the predicted survival benefit (or detriment) based on patient characteristics. However, to date, CMHE has not been used to optimize the revascularization strategy in patients with stable coronary artery disease. Using machine learning, our aim was to discover which patients are the most likely to benefit from coronary revascularization and test model predictions on a separate cohort of patients who underwent surgical revascularization at a single tertiary medical center.

3.2 Methods

3.2.1 Data

Data from the Bypass Angioplasty Revascularization Investigation in Type 2 Diabetes (BARI2D) randomized trial was included for training and validation of the CMHE model [24]. The BARI2D data includes baseline characteristics and survival data from 2,368 patients with type 2 diabetes and stable coronary artery disease who were randomly assigned to a prompt revascularization with OMT (R+OMT, n=1,176, 49.7%) or OMT alone (n=1,192, 50.3%). Prior to randomization, patients were stratified to be appropriate candidates for either percutaneous coronary intervention (PCI, n=1,605, 66.4%) or CABG (n=763, 33.6%) based on coronary angiography. The baseline characteristics of the BARI2D patients are included in Table 3.1. All data from the BARI2D study was used for CMHE model training and validation. A second CMHE model trained using only patients who underwent CABG could not be created due to too few patients in the CABG stratum in the BARI2D study.

External model testing was performed using a single-institutional cohort of patients with diabetes mellitus who underwent elective CABG. A total of 1,507 patients who met these criteria underwent CABG between 2010 and 2024. Of these, 162 (10.7%) were excluded due to missing survival data, resulting in a total cohort of 1,345 patients. The baseline and operative characteristics are included in Table 3.1. Among both databases, there were 14 features that were present within both the BARI2D and the institutional data. The CMHE models developed in this study were generated using these 14 features. The baseline feature values between the datasets are presented in

Table 3.1.

Feature	BARI2D Dataset	CABG Dataset	p-value
Congestive heart failure	6.6% [156/2348]	9.5% [128/1345]	0.002
Ejection fraction <50%	16.9 % [400/2368]	30.6% [411/1345]	< 0.001
Hispanic	12.5% [297/2368]	0.7% [9/1345]	< 0.001
Hyperlipidemia	81.9% [1913/2336]	95.7% [1287/1345]	< 0.001
Hypertension	82.5% [1929/2337]	95.5 % [1284/1345]	< 0.001
Myocardial infarction	32.0% [746/2328]	52.0% [699/1345]	< 0.001
Race (white)	70.4% [1666/2368]	94.3% [1269/1345]	< 0.001
Recent alcohol use	23.7% [561/2368]	54.7% [736/1345]	< 0.001
Recent smoker	12.5% [295/2368]	20.1% [270/1345]	< 0.001
Sex (female)	29.6% [702/2368]	26.9% [362/1345]	0.083
Age (years)	61.9 ± 8.8	64.5 ± 9.09	< 0.001
Body mass index	31.7 ± 5.8	31.8 ± 6.3	0.564
Creatinine (mg/dL)	1.0 ± 0.3	1.3 ± 1.1	< 0.001
Hemoglobin A1C (%)	7.6 ± 1.6	7.5 ± 1.5	0.002

Table 3.1: Feature comparisons between BARI2D clinical trial data and a single-institutional dataset of patients with diabetes mellitus undergoing coronary artery bypass grafting.

3.2.2 Model Training

Two outcome variables were evaluated: overall survival (OS) and major adverse cardiac and cerebrovascular events (MACCE). MACCE was defined according to the definition of the Society of Thoracic Surgeons (STS) as the composite of death, stroke, myocardial infarction, or ischemia-driven re-revascularization . The BARI2D model for predicting MACCE outcomes was created using the outcome of the major cardiovascular events (MCE) of BARI2D, which includes a combination of death, myocardial infarction or stroke. CMHE models censored the data from time to event by assuming that an individual's survival outcome depends on baseline survival rates and the treatment effect. Using deep learning, the model employs an encoder to learn representations of latent phenogroups based on observed covariates. It then estimates the individual-level survival curve under the intervention, therefore recovering subgroups of patients that respond differentially to the intervention.

Using BARI2D data, CMHE models were trained based on time-to-event for OS

or MACCE following revascularization with either PCI or CABG. The model output predicts whether patients are more likely to benefit from R+OMT or OMT alone.

3.2.3 Model Testing

Using the CMHE model, patients in the BARI2D and institutional datasets were stratified into phenogroups 1 and 2, based on the predicted response to revascularization, with phenogroup 1 likely to benefit from R+OMT, while phenogroup 2 was more likely to benefit from OMT alone. Kaplan-Meier survival curves were generated and compared using hazard ratios and rank likelihood tests. The demographic data of the patients between the phenogroups were compared using t-tests and chi-squared tests for continuous and categorical variables, respectively.

3.3 Results

The CMHE model trained on BARI2D overall survival data identified two distinct phenogroups based on response to revascularization. Phenogroup 1 ($n = 966, 40.8\%$) was more likely to receive benefit from revascularization (HR $= 0.60, p = 0.003$), whereas phenogroup 2 ($n = 1,402, 59.2\%$) was more likely to be harmed (HR $= 1.55, p = 0.007$). Patients in phenogroup 1 and phenogroup 2 in BARI2D who underwent revascularization experienced similar Kaplan-Meier survival ($p = 0.407$, Figure 3.1). Phenogroup 1 patients were older (64.0 ± 8.5 versus 60.5 ± 8.7 years, $p < 0.001$), more often female (44.2% [427/966] versus 19.6% [275/1,402], $p < 0.001$), and had higher HbA1C (8.1 ± 1.7 versus 7.3 ± 1.5, $p < 0.001$, Table 3.2).

When applied to a dataset of patients undergoing elective CABG, patients in phenogroup 1 ($n = 303, 39.8\%$) had similar Kaplan-Meier survival compared to phenogroup 2 ($n = 1,042, 60.2\%$, HR $= 1.00, p = 0.433$, Figure 3.1) and similar operative mortality compared to phenogroup 2 (1.7% [5/303] versus 1.2% [13/1,042], $p = 0.800$). Patients in phenogroup 1 also had a higher STS predicted risk of mortality ($2.0 \pm 2.1\%$ versus $1.2 \pm 1.2\%$, $p < 0.001$).

CMHE assumes that conditional on the latent group, individual time-to-event distributions obey proportional hazards. We note that for the BARI2D dataset, in phenogroup 1, proportional hazards does not hold for hyperlipidemia ($p = 0.024$)

and ejection fraction <50% ($p = 0.054$, and in phenogroup 2, it does not hold for age ($p = 0.021$). For the CABG dataset, in phenogroup 1, proportional hazards does not hold for hypertension ($p = 0.040$) and body mass index ($p = 0.019$), and in phenogroup 2, it does not hold for race ($p = 0.013$).

3.4 Discussion

Randomized controlled trials have long been the gold standard for evidence-based practice in medicine, including cardiovascular medicine as it relates to coronary revascularization [34]. However, clinical trials necessarily group diverse populations of patients with differing baseline characteristics to identify broad treatment recommendations, but within these populations, there may be subgroups of patients with differing response to treatment, a concept referred to as "heterogeneous treatment effect" and is defined in the Predictive Approaches to Treatment Effect Heterogeneity Statement (PATH) [1, 31]. Performing a separate randomized controlled trial in all possible subpopulations would be prohibitively costly and impractical. Thus, ML techniques can help to discover these phenogroups and thus allow for a more personalized treatment approach for individual patients[31, 41, 45, 46]. The CMHE ML algorithm was developed with the aim of discovering phenogroups with heterogeneous treatment effects in time-to-event data, making it ideal for applying personalized recommendations to prolong survival in patients who may be considered for coronary revascularization[41]. In the present study, a CMHE model was trained using data from the BARI2D randomized controlled trial, and the phenogroups were applied to a separate data set of patients who underwent CABG at a single large tertiary medical center between 2011-2024.

Note that the assumption of CMHE does not hold for some variables, and there are other algorithms for causal survival analysis that might be applicable. Causal Survival Forest [14] estimates the conditional average treatment effect on the survival time, by adapting the causal forest algorithm [3] to right-censored data. Bo et al. propose a causal meta-learning approach that estimates individualized treatment effect by extending T-learner and X-learner. Multiple Imputation for Survival Treatment Response [39] handles instrumental variables and heavy censoring using ecursively imputed survival trees. Chapfuwa et al. [9] propose a generative model that adjusts

for informative censoring and a nonparametric hazard ratio metric for evaluating average and individualized treatment effects.

3.5 Conclusion

ML can help elucidate which patients are likely to benefit from revascularization versus optimal medical therapy alone. Interestingly, patients in phenogroup 1, who are expected to experience a long-term survival benefit from revascularization based on the CMHE ML model, had a significantly higher STS predicted risk of operative mortality compared to phenogroup 2. Despite the increased perioperative risk, survival rates did not differ significantly between the two phenogroups. These results highlight the potential for this and other ML models to complement the STS risk prediction model, improving preoperative decision-making about optimal treatment selection for patients with coronary artery disease.

This study illustrates that valuable insights can be gained even from trials deemed negative or inconclusive by regulatory standards, such as the BARI2D study, which reported no significant difference in the rates of death between patients receiving prompt revascularization and those treated with OMT [24]. Importantly, machine learning models such as CMHE provide a more nuanced, individualized approach to treatment evaluation. Unlike conventional tools such as the STS risk score, which lacks transparency in its methodology, CMHE supports counterfactual reasoning, enabling personalized estimates of treatment benefit, and is available as an open-source tool, promoting reproducibility and further research.

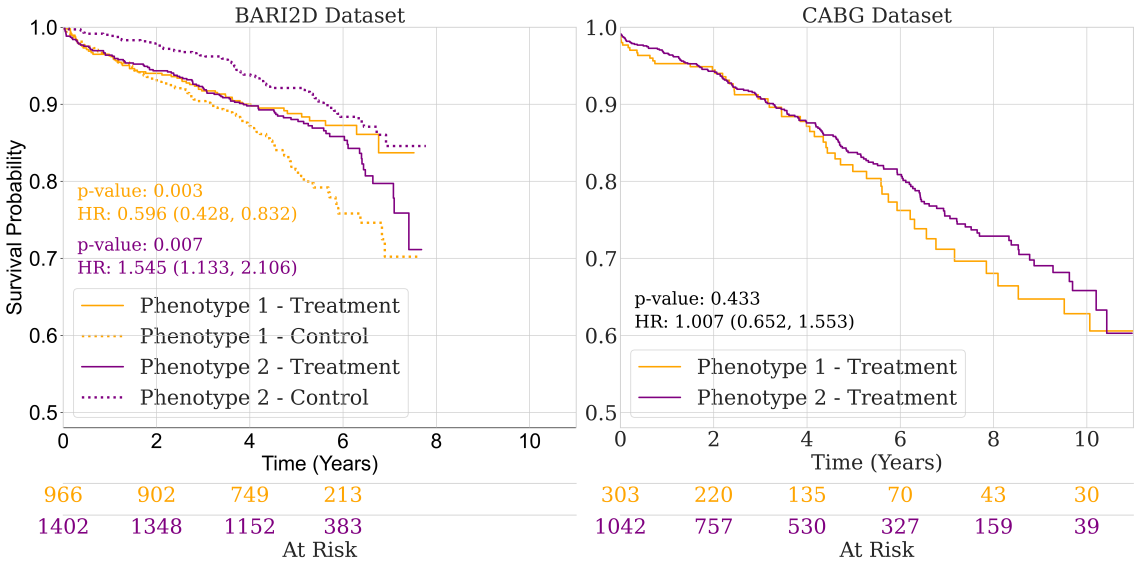


Figure 3.1: Kaplan-Meier survival curves. (Left) shows survival estimates for two phenogroups of BARI2D patients identified using the Cox Mixtures with Heterogeneous Effects machine learning (ML) model based on survival response to revascularization. (Right) shows this same model applied to an institutional database of patients who underwent coronary artery bypass grafting (CABG). This plot shows that long-term survival does not significantly differ between phenogroups, despite a higher baseline predicted risk of mortality in phenogroup 1. These results indicate phenogroup 1 patients may have derived particular benefit from revascularization with CABG, as predicted by the ML model.

Feature	Phenogroup 1	Phenogroup 2	p-value
	BARI2D Dataset		
Congestive heart failure	7.4% [71/957]	6.1% [85/1391]	0.243
Ejection fraction <50%	16.7% [157/939]	18.0% [243/1351]	0.466
Hispanic	19.7% [190/966]	7.6% [107/1402]	<0.001
Hyperlipidemia	82.0% [171/951]	81.8% [1133/1385]	0.938
Hypertension	84.5% [148/954]	81.2% [1123/1383]	0.045
Myocardial infarction	25.5% [251/948]	35.9% [495/1380]	<0.001
Race (white)	56.6% [547/966]	79.8% [1119/1402]	<0.001
Recent alcohol use	4.2% [40/956]	37.3% [521/1395]	<0.001
Recent smoker	2.7% [26/966]	19.2% [269/1402]	<0.001
Sex (female)	44.2% [427/966]	19.6% [275/1402]	<0.001
Age (years)	64.0 ± 8.6	60.5 ± 8.7	<0.001
Body mass index	28.3 ± 3.6	34.1 ± 5.8	<0.001
Creatinine (mg/dL)	1.1 ± 0.3	1.0 ± 0.3	0.230
Hemoglobin A1C (%)	8.1 ± 1.7	7.3 ± 1.4	<0.001
	CABG Dataset		
Congestive heart failure	27.1% [82/303]	19.3% [201/1042]	<0.001
Ejection fraction <50%	34.3% [104/303]	29.5% [307/1042]	0.122
Hispanic	1.3% [4/303]	0.5% [5/1037]	0.238
Hyperlipidemia	94.0% [285/303]	96.2% [1003/1042]	0.154
Hypertension	93.7% [284/303]	96.9% [1000/1042]	0.136
Myocardial infarction	48.2% [146/303]	53.1% [553/1042]	0.152
Race (white)	92.1% [270/303]	95.0% [990/1042]	0.071
Recent alcohol use	15.2% [46/303]	66.2% [690/1042]	<0.001
Recent smoker	3.0% [9/303]	25.0% [261/1042]	<0.001
Sex (female)	34.3% [133/303]	22.0% [229/1042]	<0.001
Age (years)	67.8 ± 9.1	63.6 ± 8.9	<0.001
Body mass index	27.0 ± 3.8	33.2 ± 6.2	<0.001
Creatinine (mg/dL)	1.6 ± 1.7	1.2 ± 0.8	<0.001
Hemoglobin A1C (%)	7.7 ± 1.6	7.4 ± 1.5	0.002
STS Predicted Mortality (%)	2.0 ± 2.1	1.2 ± 1.2	<0.001
Operative Mortality	1.7% [5/303]	1.2% [13/1029]	0.800

Table 3.2: Feature comparisons between phenogroups for BARI2D clinical trial data and a single-institutional dataset of patients with diabetes mellitus undergoing coronary artery bypass grafting.

Chapter 4

Conclusions and Future Work

4.1 Summary of Contributions

This thesis demonstrates the power and versatility of survival-based machine learning approaches in tackling two clinically important but distinct challenges: personalized risk prediction from chest radiographs and individualized treatment effect estimation for coronary revascularization. In the first study, we showed that time-to-event models, when combined with imaging and demographic data, can generate flexible, accurate, and well-calibrated mortality risk predictions, overcoming the limitations of fixed-horizon binary classifiers. In the second study, we extended the utility of survival analysis to treatment decision making by identifying subpopulations of patients with stable ischemic heart disease who derive the most benefit from coronary artery bypass grafting. Using a heterogeneous treatment effect model (CMHE), we moved beyond average treatment effects to support individualized, data-driven clinical choices.

Together, these works underscore the importance of modeling time explicitly and accounting for censoring in real-world health data. They also highlight the growing role of deep learning in integrating high-dimensional inputs such as medical imaging with survival analysis for both prognosis and treatment optimization. Future work could expand these models to larger, multi-center datasets; incorporate additional modalities such as electronic health records and genomic data; and explore deployment in clinical workflows to support dynamic, personalized decision-making. Ultimately, this thesis contributes to the broader goal of building intelligent systems that can

improve patient outcomes through precision medicine.

4.2 Future Research Opportunities

Building upon this work, several promising avenues can be explored to further advance survival prediction.

1. **Improving Fairness and Generalizability Across Populations**: the studies highlight potential demographic imbalances in training data, particularly with respect to age and race, which can limit the generalizability of the models to underrepresented sub-populations. Future work should explore model retraining and evaluation using datasets that are more demographically diverse, and integrate fairness-aware modeling techniques to reduce bias in risk and treatment predictions across subgroups.
2. **Integration of Time-Varying and Longitudinal Data**: Current models primarily use static snapshots of patient data (e.g., a single radiograph or baseline clinical characteristics). Future directions could include incorporating longitudinal imaging, laboratory values, or EHR-derived time series data to capture evolving patient risk and improve predictions over time using dynamic survival models or recurrent neural architectures.

Appendix A

Deep Survival Models Can Improve Long-Term Mortality Risk Estimates from Chest Radiographs

A.1 Censoring and Time-to-Event Predictions

We are interested in estimating $\mathbb{P}(T < t|X)$, the risk of a patient experiencing mortality within a time horizon $t \in \mathbb{R}^+$, given a chest radiograph $\boldsymbol{x} \in \mathbb{R}^{d \times d}$ and a set of demographic covariates $\boldsymbol{g} \in \mathbb{R}^m$. We assume that observations on T can be right-censored, which means that the observation is terminated before death occurs.

Censoring makes it challenging to learn an estimator of survival. As evidenced in Figure A.1, the censored patients need to be accounted for to obtain an unbiased estimate of patient survival. Standard survival models such as the Cox model make strong assumptions about the distribution of the event times. Hence, they cannot generalize to situations where the assumptions are violated (Figure A.2).

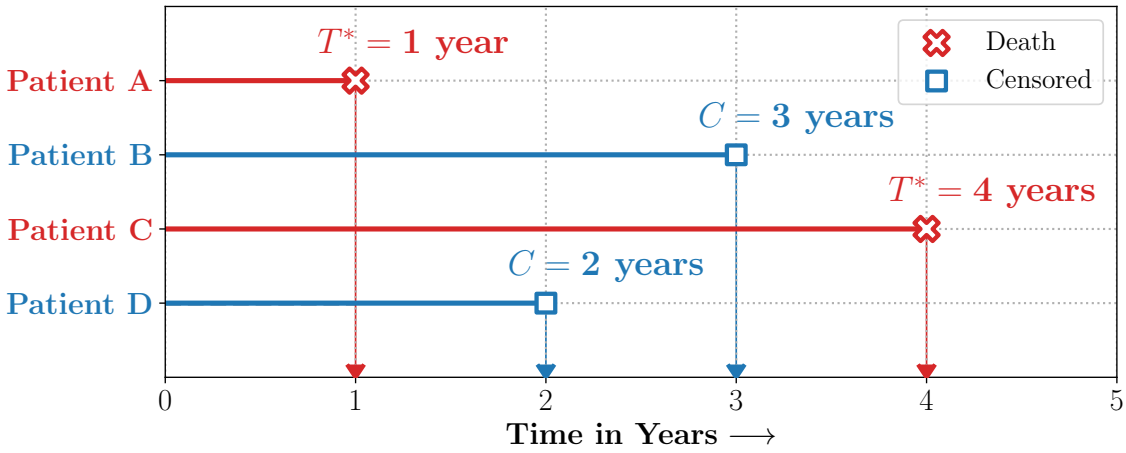


Figure A.1: Censoring and Time-to-Event Predictions. Patients A and C died 1 and 4 years from entry into the study, whereas Patients B and D exited the study without experiencing death (were lost to follow-up) at 2 and 3 years from entry into the study. Time-to-event and survival regression thus involve estimates that are adjusted for individuals whose outcomes were censored.
Source: Adapted from [43]

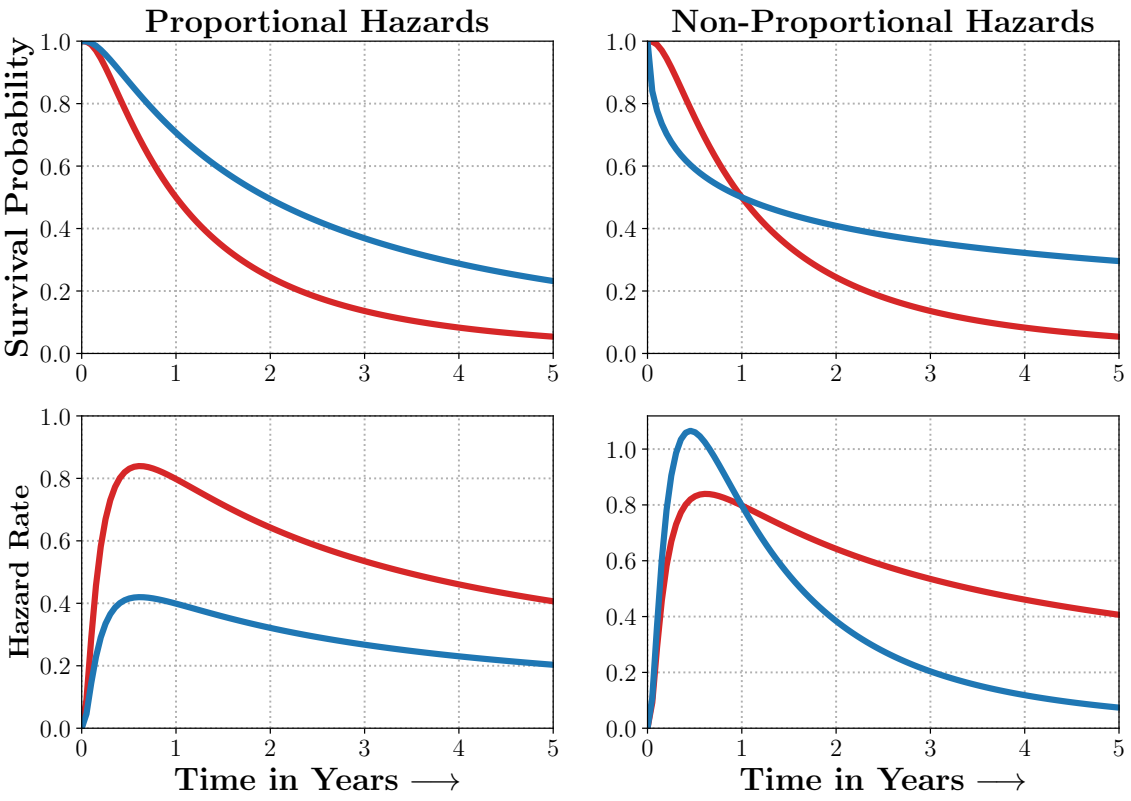


Figure A.2: Non-Proportional Hazards. When the proportional hazards assumptions are satisfied, the survival curves and their corresponding hazard rates dominate each other and do not intersect. In many real-world scenarios, however, the survival curves do. Models such as Deep Survival Machines include flexible estimators of times-to-events in the presence of non-proportional hazards.
Source: Adapted from [43]

A.2 Additional Details on Proportional Hazards Model

Cox proportional hazards (CPH) makes the proportional hazards assumption that the ratio between two hazards is constant over time [12]. The hazard $\lambda(t)$ as a function of some time t is defined as the probability that given that a patient survives t, the patient will not survive $t + \delta$, and δ approaches 0 (Equation (A.1)).

$$\lambda(t) = \lim_{\delta \to 0} \frac{P(t + \delta > T \geq t | T \geq t)}{\delta} \tag{A.1}$$

CPH models the hazard as Equation (A.2), given covariates $x_{ij}, i \in \{1, \ldots, n\}, j \in \{1, \ldots, p\}$, where p is the number of predictors, and n is the number of patients. The parameters of the model are β_j. $\lambda_0(t)$ is known as the baseline hazard function.

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \tag{A.2}$$

The proportional hazards assumption comes from the fact that the proportion of two hazard functions is fixed, since $\frac{\lambda_i(t)}{\lambda_j(t)} = \exp(\beta_1(x_{i1} - x_{j1}) + \ldots + \beta_p(x_{ip} - x_{jp}))$.

We estimated that $\beta = \{\beta_1, \ldots, \beta_p\}$ are the ones that maximize the partial likelihood
(Equation (A.3)), where $i : E_i = 1$ are patients whose outcomes are not censored, and $\mathscr{R}(T_i)$ is the set of patients who have not experienced the event at time T_i.

$$L(\beta) = \prod_{i:E_i=1} \frac{\exp(\beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{\sum_{j \in \mathscr{R}(T_i)} \exp(\beta_1 x_{j1} + \ldots + \beta_p x_{jp})} \tag{A.3}$$

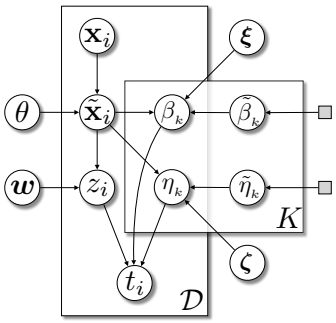
DeepSurv [30] is a deep neural network implementation of CPH. It estimates the risk function $\hat{h}_\theta(x)$ of an individual with covariates x, parametrized by the weights θ of a multi-layer perceptron. The loss function is the negative log partial likelihood (Equation (A.4)).

$$L_{DCPH}(\theta) = -\sum_{i:E_i=1} \left(\hat{h}_\theta(x_i) - \log \sum_{j \in \mathscr{R}(T_i)} e^{\hat{h}_\theta(x_j)} \right) \tag{A.4}$$

A.3 Additional Details on Deep Survival Machines

In this section, we describe Deep Survival Machines, which models the time-to-event distribution. The DSM model involves modeling the time-to-event distribution as a fixed-size mixture of K parametric survival distributions. This allows the model to generalize to situations where the survival curves intersect (a clear violation of the

proportional hazards assumption.). The final time-to-event outcome is determined by averaging over the latent $Z = k$. We provide DSM in plate notation in Figure A.3.



The Generative Story

1. $\mathbf{x}_i \sim \mathcal{D}$
   We draw the covariates of the individual, $\mathbf{x}_i$.
2. $\boldsymbol{w}, \boldsymbol{\zeta}, \boldsymbol{\xi} \sim \mathcal{N}(0, 1/\lambda)$
   The parameters of the model are drawn from a zero mean Gaussian distribution.
3. $z_i \sim \text{Discrete}\big(\texttt{softmax}(\Phi_\theta(\boldsymbol{x}_i)^\top \boldsymbol{w})\big)$
   Conditioned on the covariates, $\boldsymbol{x}_i$, and the parameters, $\boldsymbol{w}$, we draw the latent z_i.
4. $\log \tilde{\beta}_k \sim \mathcal{N}(\beta_0, 1/\lambda)$, $\log \tilde{\eta}_k \sim \mathcal{N}(\eta_0, 1/\lambda)$
   The set of parameters $\{\tilde{\beta}_k\}_{k=1}^K$ and $\{\tilde{\eta}_k\}_{k=1}^K$ are drawn from the prior β_0 and η_0.
5. $\boldsymbol{t}_i \sim \text{Distribution}\big(\beta_k, \eta_k\big)$
   where $\beta_k = \tilde{\beta}_k + \texttt{act}(\Phi_\theta(\boldsymbol{x}_i)^\top \boldsymbol{\zeta})$
   $\eta_k = \tilde{\eta}_k + \texttt{act}(\Phi_\theta(\boldsymbol{x}_i)^\top \boldsymbol{\xi})$
   Finally, the event time t_i is drawn, conditioned on β_{z_i} and η_{z_i}.

Figure A.3: Deep Survival Machines in plate notation.
Source: Adapted from [40]

A.4 Censoring Adjusted Evaluation Metrics

Brier Score (BS): The Brier score involves computing the mean squared error of the estimated survival probabilities, $\widehat{\mathbb{P}}(T > t|X = \boldsymbol{x}) = f(\boldsymbol{x}, t)$, at a specified time horizon, t. As a proper scoring rule, the Brier score gives a sense of both discrimination and calibration.

Under the assumption that the censoring distribution is independent of the time-to-event, we can obtain an unbiased, censoring adjusted estimate of the Brier

score $\widehat{\mathrm{BS}}_{\mathrm{IPCW}}(t)$ using the inverse probability of censoring weights (IPCW) from a Kaplan–Meier estimator of the censoring distribution, $\hat{G}(.)$, as proposed in [19, 22].

$$\begin{aligned}\mathrm{BS}(t) &= \mathbb{E}_{\mathcal{D}}\big[\big(\mathbf{1}\{T_i > t\} - \widehat{\mathbb{P}}(T > t|X = \boldsymbol{x})\big)^2\big] \\ \widehat{\mathrm{BS}}_{\mathrm{IPCW}}(t) &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{f(\boldsymbol{x}_i,t)^2\mathbf{1}\{T \le t, \delta_i = 1\}}{\hat{G}_i(T_i)} + \frac{\big(1 - f(\boldsymbol{x}_i,t)\big)^2\mathbf{1}\{T > t\}}{\hat{G}_i(t)}\right].\end{aligned}$$

Area under ROC Curve (AUC): ROC curves are extensively used in classification tasks where the true positive rate (TPR) is plotted against the false positive rate (1-true negative rate, or TNR) to measure the model's discriminative power over all output thresholds.

To enable the use of ROC curves to assess the performance of survival models subject to censoring, we employed the technique proposed by [27, 55], which treats the TPR as time-dependent on a specified horizon, t, and adjusts survival probabilities, $f(\boldsymbol{x}, t)$, using the IPCW from a Kaplan–Meier estimator of the censoring distribution, $\hat{G}(t)$. Estimating the TNR requires observing outcomes for each individual; only uncensored instances are used. Refer to [29] for details on computing ROC curves in the presence of censoring.

$$\widehat{\mathrm{TPR}}(c,t) = \frac{\sum_{i=1}^{n}\frac{\delta_i}{\hat{G}(T_i)}\cdot\mathbf{1}\{f(\boldsymbol{x}_i,t) > c, T_i \le t\}}{\sum_{i=1}^{n}\frac{\delta_i}{\hat{G}(T_i)}\cdot\mathbf{1}\{T_i < t\}\cdot}; \quad \widehat{\mathrm{TNR}}(c,t) = \frac{\sum_{i=1}^{n}\mathbf{1}\{f(\boldsymbol{x}_i,t) \le c, T_i > t\}}{\sum_{i=1}^{n}\mathbf{1}\{T_i > t\}}$$

Time-Dependent Concordance Index (C^{td}): The concordance index compares risks across all pairs of individuals within a fixed time horizon, t, to estimate the ability to appropriately rank instances relative to each other in terms of their risks, $f(\boldsymbol{x}, t)$.

$$C^{\mathrm{td}}(t) = \mathbb{P}\big(f(\boldsymbol{x}_i,t) \le f(\boldsymbol{x}_j,t)|\delta_i = 1, T_i < T_j, T_i \le t\big)$$

We employed the censoring adjusted estimator for C^{td}, which exploits IPCW estimates from a Kaplan–Meier estimate of the censoring distribution. Further details can be found in [56] and [20].

Expected $\ell_{\mathbf{1}}$ Calibration Error (ECE): The ECE measures the average absolute difference between the observed and expected (according to the risk score) event

rates, conditional on the estimated risk score. At time t, let the predicted risk score be $R(t) = \widehat{\mathbb{P}}(T > t|X)$. Then, the ECE approximates

$$\mathrm{ECE}(t) = \mathbb{E}\big[\big|\mathbb{P}(T > t|R(t)) - R(t)\big|\big]$$

by partitioning the risk scores R into q quantiles $\{[r_j, r_{j+1})\}_{j=1}^{q}$ and computing the Kaplan–Meier estimate of the event rate. $\mathrm{KM}_j(t) \approx \mathbb{P}(T > t|R \in [r_j, r_{j+1}))$, and the average risk score, $\overline{R}_j = \frac{q}{n}\sum_{i:R_i\in[r_j,r_{j+1})} R_i$, in each bin. Altogether, the estimated ECE is

$$\widehat{\mathrm{ECE}}(t) = \frac{1}{q}\sum_{j=1}^{q} |\mathrm{KM}_j(t) - \overline{R}_j(t)|.$$

In practice, we fixed the number of quantiles to 20 for our experiments.

A.5 Implementation Details

All experiments were run on NVIDIA RTX A6000, with Python 3.9.12 and Pytorch 1.11.0. Statistical analysis was performed with Python 3.9.12.

We performed a hyperparameter search by searching for the hyperparameters that yielded the lowest validation Brier score.

For DCPH, DSM, and TBC models, layer sizes (the number of neurons in each hidden layer of survival models) are searched from [( ) (no hidden layers), (64) (one hidden layer with 64 neurons), (64, 64) (two hidden layers with 64 neurons each), (128) (one hidden layer with 128 neurons), and (128, 128) (two hidden layers with 128 neurons each)]. For DSM models, in addition to layer sizes, the number of underlying distributions k is searched from [2, 3, 4, 6], temperatures from [1, 100, 500, 1000], and elbo from [True, False]. The configurations with the best validation performance are presented in Table A.1.

Adam optimizer with a learning rate of 3×10^{-4} was used for all models. The distribution for all DSM models was the Weibull distribution. The training batch size was 64, and the validation batch size was 32. Models were trained for 10 epochs, with a patience of 3.

		DSM		DCPH
Layer Size	**k**	**Temperature**	**Elbo**	**Layer Size**
(128)	4	1	True	(64, 64)

TBC, 2-Year	**TBC, 5-Year**	**TBC, 10-Year**
	Layer Size	
(64)	(64, 64)	(128)

Table A.1: Hyperparameter configurations that yielded the best validation performance

Appendix B

Machine Learning Identifies Patients Who Derive Survival Benefit from Coronary Revascularization

B.1 Implementation Details

Models were trained using Python 3.9.12 and the auton-survival package available through GitHub. Deep learning models were trained using a CPU. Model hyperparameters are the following.

1. Horizons: [1, 3, 5]
2. Number of underlying treatment effect phenotypes: [2]
3. Number of training epochs: 100
4. Learning rate: 0.01
5. Batch size: 256
6. Size of the validation split: 0.15
7. Patience: 3
8. Optimizer: Adam

Hyperparameter selection is performed for the following:

1. Number of underlying base survival phenotypes: 1 or 2
2. Number of neurons in each hidden layer: [50, 50] or [50]

The best set of hyperparameters is the one that yielded the lowest integrated Brier score. For mortality, (1, [50, 50]) was chosen, and for MACCE, (2, [50]) was chosen.

Bibliography

[1] The predictive approaches to treatment effect heterogeneity (path) statement. *Annals of Internal Medicine*, 172(1):35–45, 2020. doi: 10.7326/M18-3667. URL https://doi.org/10.7326/M18-3667. PMID: 31711134. 3.4

[2] Gerald L. Andriole, E. David Crawford, Robert L. Grubb, Saundra S. Buys, David Chia, Timothy R. Church, Mona N. Fouad, Claudine Isaacs, Paul A. Kvale, Douglas J. Reding, Joel L. Weissfeld, Lance A. Yokochi, Barbara O'Brien, Lawrence R. Ragard, Jonathan D. Clapp, Joshua M. Rathmell, Thomas L. Riley, Ann W. Hsing, Grant Izmirlian, Paul F. Pinsky, Barnett S. Kramer, Anthony B. Miller, John K. Gohagan, Philip C. Prorok, and PLCO Project Team. Prostate cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: Mortality results after 13 years of follow-up. *Journal of the National Cancer Institute*, 104(2):125–132, January 2012. ISSN 1460-2105. doi: 10.1093/jnci/djr500. 2.2, 2.4

[3] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests, 2018. URL https://arxiv.org/abs/1610.01271. 3.4

[4] Faisal G. Bakaeen, Marc Ruel, John H. Calhoon, Leonard N. Girardi, Robert Guyton, Dawn Hui, Rosemary F. Kelly, Thomas E. MacGillivray, S. Christopher Malaisrie, Marc R. Moon, Joseph F. Sabik, 3rd, Peter K. Smith, Lars G. Svensson, and Wilson Y. Szeto. STS/AATS-Endorsed Rebuttal to 2023 ACC/AHA Chronic Coronary Disease Guideline: A Missed Opportunity to Present Accurate and Comprehensive Revascularization Recommendations. 116(4):675–678. ISSN 0003-4975. doi: 10.1016/j.athoracsur.2023.02.007. URL https://doi.org/10.1016/j.athoracsur.2023.02.007. 3.1

[5] Mihalj Bakator and Dragica Radosav. Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction*, 2(3), 2018. ISSN 2414-4088. doi: 10.3390/mti2030047. URL https://www.mdpi.com/2414-4088/2/3/47. 2.1

[6] Ivo M. Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. *Scientific Reports*, 9(1):6381, April 2019. ISSN 2045-2322.

doi: 10.1038/s41598-019-42294-8. 2.1

[7] Rudolf Beran. Nonparametric regression with randomly censored survival data. 01 1981. 1.1, 2.1

[8] GLENN W. BRIER. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3, 1950. doi: https://doi.org/10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml. 2.2.2

[9] Paidamoyo Chapfuwa, Serge Assaad, Shuxi Zeng, Michael J Pencina, Lawrence Carin, and Ricardo Henao. Enabling counterfactual survival analysis with balanced representations. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 133–145, 2021. 3.4

[10] Jianhong Cheng, John Sollee, Celina Hsieh, Hailin Yue, Nicholas Vandal, Justin Shanahan, Ji Whae Choi, Thi My Linh Tran, Kasey Halsey, Franklin Iheanacho, James Warren, Abdullah Ahmed, Carsten Eickhoff, Michael Feldman, Eduardo Mortani Barbosa, Ihab Kamel, Cheng Ting Lin, Thomas Yi, Terrance Healey, Paul Zhang, Jing Wu, Michael Atalay, Harrison X. Bai, Zhicheng Jiao, and Jianxin Wang. COVID-19 mortality prediction in the intensive care unit with deep learning based on longitudinal chest X-rays and clinical data. *European Radiology*, 32(7):4446–4456, July 2022. ISSN 1432-1084. doi: 10.1007/s00330-022-08588-8. 2.1, 2.4

[11] Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P. Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, pages 231–249. PMLR, December 2022. 2.2.3

[12] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1972.tb00899.x. 1.1, 2.1, A.2

[13] Erin Craig, Chenyang Zhong, and Robert Tibshirani. Survival stacking: casting survival analysis as a classification problem. *arXiv preprint arXiv:2107.13480*, 2021. 2.1

[14] Yifan Cui, Michael R Kosorok, Erik Sverdrup, Stefan Wager, and Ruoqing Zhu. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):179–211, 02 2023. ISSN 1369-7412. doi: 10.1093/jrsssb/qkac001. URL https://doi.org/10.1093/jrsssb/qkac001. 3.4

[15] William F. Fearon, Frederik M. Zimmermann, Bernard De Bruyne, Zsolt Piroth, Albert H.M. van Straten, Laszlo Szekely, Giedrius Davidavičius, Gintaras Kalinauskas, Samer Mansour, Rajesh Kharbanda, Nikolaos Östlund Papadogeorgos, Adel Aminian, Keith G. Oldroyd, Nawwar Al-Attar, Nikola Jagic, Jan-Henk E. Dambrink, Petr Kala, Oskar Angerås, Philip MacCarthy, Olaf Wendler, Filip Casselman, Nils Witt, Kreton Mavromatis, Steven E.S. Miner, Jaydeep Sarma, Thomas Engstrøm, Evald H. Christiansen, Pim A.L. Tonino, Michael J. Reardon, Di Lu, Victoria Y. Ding, Yuhei Kobayashi, Mark A. Hlatky, Kenneth W. Mahaffey, Manisha Desai, Y. Joseph Woo, Alan C. Yeung, and Nico H.J. Pijls. Fractional flow reserve–guided pci as compared with coronary bypass surgery. *New England Journal of Medicine*, 386(2):128–137, 2022. doi: 10.1056/NEJMoa2112299. URL https://www.nejm.org/doi/full/10.1056/NEJMoa2112299. 3.1

[16] Stephane Fotso. Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*, 2018. 1.1, 2.1

[17] Mario Gaudino, Katia Audisio, Whady A. Hueb, Gregg W. Stone, Michael E. Farkouh, Antonino Di Franco, Mohamed Rahouma, Patrick W. Serruys, Deepak L. Bhatt, Giuseppe Biondi Zoccai, Salim Yusuf, Leonard N. Girardi, Stephen E. Fremes, Marc Ruel, and Bjorn Redfors. Coronary artery bypass grafting versus medical therapy in patients with stable coronary artery disease: An individual patient data pooled meta-analysis of randomized trials. *The Journal of Thoracic and Cardiovascular Surgery*, 167(3):1022–1032.e14, 2024. ISSN 0022-5223. doi: https://doi.org/10.1016/j.jtcvs.2022.06.003. URL https://www.sciencedirect.com/science/article/pii/S0022522322006419. 3.1

[18] Michael F Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019. 1.1, 2.1

[19] Thomas A Gerds and Martin Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006. A.4

[20] Thomas A Gerds, Michael W Kattan, Martin Schumacher, and Changhong Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in medicine*, 32(13):2173–2184, 2013. A.4

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, November 2016. ISBN 978-0-262-03561-3. 2.1

[22] Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999. A.4

[23] Lisa Gren, Karen Broski, Jeffery Childs, Jill Cordes, Deborah Engelhard, Betsy

Gahagan, Eduard Gamito, Vivien Gardner, Mindy Geisser, Darlene Higgins, Victoria Jenkins, Lois Lamerato, Karen Lappe, Heidi Lowery, Colleen McGuire, Mollie Miedzinski, Sheryl Ogden, Sally Tenorio, Gavin Watt, Bonita Wohlers, and Pamela Marcus. Recruitment methods employed in the prostate, lung, colorectal, and ovarian cancer screening trial. *Clinical Trials*, 6(1):52–59, 2009. doi: 10.1177/1740774508100974. URL https://doi.org/10.1177/1740774508100974. PMID: 19254935. 2.2

[24] The BARI 2D Study Group. A randomized trial of therapies for type 2 diabetes and coronary artery disease. *New England Journal of Medicine*, 360(24):2503–2515, 2009. doi: 10.1056/NEJMoa0805796. URL https://www.nejm.org/doi/full/10.1056/NEJMoa0805796. 3.2.1, 3.5

[25] Marsha A. Hasson, Richard M. Fagerstrom, Dalia C. Kahane, Judith H. Walsh, Max H. Myers, Clifford Caughman, Blaine Wenzel, Juline C. Haralson, Lynn M. Flickinger, and Louisa M. Turner. Design and evolution of the data management systems in the prostate, lung, colorectal and ovarian (plco) cancer screening trial. *Controlled Clinical Trials*, 21(6, Supplement 1):329S–348S, 2000. ISSN 0197-2456. doi: https://doi.org/10.1016/S0197-2456(00)00100-8. URL https://www.sciencedirect.com/science/article/pii/S0197245600001008. 2.2

[26] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks, January 2018. 2.2.3

[27] Hung Hung and Chin-tsang Chiang. Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian journal of statistics*, 37(4):664–679, 2010. 2.2.2, A.4

[28] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. 2008. 1.1, 2.1

[29] Adina Najwa Kamarudin, Trevor Cox, and Ruwanthi Kolamunnage-Dona. Time-dependent roc curve analysis in medical research: current methods and applications. *BMC medical research methodology*, 17(1):1–19, 2017. A.4

[30] 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, February 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0482-1. 2.4, A.2

[31] David Kent, Ewout Steyerberg, and David Klaveren. Personalized evidence based medicine: Predictive approaches to heterogeneous treatment effects. *BMJ*, 363:k4245, 12 2018. doi: 10.1136/bmj.k4245. 3.4

[32] Márton Kolossváry, Vineet K. Raghu, John T. Nagurney, Udo Hoffmann, and Michael T. Lu. Deep Learning Analysis of Chest Radiographs to Triage Patients

with Acute Chest Pain Syndrome. *Radiology*, 306(2):e221926, February 2023. ISSN 0033-8419. doi: 10.1148/radiol.221926. 2.1, 2.4

[33] 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. 1.1, 2.1

[34] Jennifer S. Lawton, Jacqueline E. Tamis-Holland, Sripal Bangalore, Eric R. Bates, Theresa M. Beckie, James M. Bischoff, John A. Bittl, Mauricio G. Cohen, J. Michael DiMaio, Creighton W. Don, Stephen E. Fremes, Mario F. Gaudino, Zachary D. Goldberger, Michael C. Grant, Jang B. Jaswal, Paul A. Kurlansky, Roxana Mehran, Thomas S. Metkus, Lorraine C. Nnacheta, Sunil V. Rao, Frank W. Sellke, Garima Sharma, Celina M. Yong, and Brittany A. Zwischenberger. 2021 acc/aha/scai guideline for coronary artery revascularization: A report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. *Circulation*, 145(3):e18–e114, 2022. doi: 10.1161/CIR.0000000000001038. URL https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000001038. 3.4

[35] Mingzhu Liu, Chirag Nagpal, and Artur Dubrawski. Deep survival models can improve long-term mortality risk estimates from chest radiographs. *Forecasting*, 6(2):404–417, 2024. ISSN 2571-9394. doi: 10.3390/forecast6020022. URL https://www.mdpi.com/2571-9394/6/2/22. 1

[36] Michael T. Lu, Alexander Ivanov, Thomas Mayrhofer, Ahmed Hosny, Hugo J. W. L. Aerts, and Udo Hoffmann. Deep Learning to Assess Long-term Mortality From Chest Radiographs. *JAMA Network Open*, 2(7):e197416, July 2019. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2019.7416. 2.1, 2.2.3, 2.4

[37] David J. Maron, Judith S. Hochman, Harmony R. Reynolds, Sripal Bangalore, Sean M. O'Brien, William E. Boden, Bernard R. Chaitman, Roxy Senior, Jose López-Sendón, Karen P. Alexander, Renato D. Lopes, Leslee J. Shaw, Jeffrey S. Berger, Jonathan D. Newman, Mandeep S. Sidhu, Shaun G. Goodman, Witold Ruzyllo, Gilbert Gosselin, Aldo P. Maggioni, Harvey D. White, Balram Bhargava, James K. Min, G.B. John Mancini, Daniel S. Berman, Michael H. Picard, Raymond Y. Kwong, Ziad A. Ali, Daniel B. Mark, John A. Spertus, Mangalath N. Krishnan, Ahmed Elghamaz, Nagaraja Moorthy, Whady A. Hueb, Marcin Demkow, Kreton Mavromatis, Olga Bockeria, Jesus Peteiro, Todd D. Miller, Hanna Szwed, Rolf Doerr, Matyas Keltai, Joseph B. Selvanayagam, P. Gabriel Steg, Claes Held, Shun Kohsaka, Stavroula Mavromichalis, Ruth Kirby, Neal O. Jeffries, Frank E. Harrell, Frank W. Rockhold, Samuel Broderick, T. Bruce Ferguson, David O. Williams, Robert A. Harrington, Gregg W. Stone, and Yves Rosenberg. Initial invasive or conservative strategy for stable coronary disease. *New England Journal of Medicine*, 382(15):1395–1407, 2020. doi:

10.1056/NEJMoa1915922. URL https://www.nejm.org/doi/full/10.1056/NEJMoa1915922. 3.1

[38] Anoop Mayampurath, L. Nelson Sanchez-Pinto, Kyle A. Carey, Laura-Ruth Venable, and Matthew Churpek. Combining patient visual timelines with deep learning to predict mortality. *PLOS ONE*, 14(7):e0220640, July 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0220640. 2.1, 2.4

[39] Tomer Meir, Uri Shalit, and Malka Gorfine. Heterogeneous treatment effect in time-to-event outcomes: Harnessing censored data with recursively imputed trees, 2025. URL https://arxiv.org/abs/2502.01575. 3.4

[40] Chirag Nagpal, Xinyu Rachel Li, and Artur Dubrawski. Deep Survival Machines: Fully Parametric Survival Regression and Representation Learning for Censored Data with Competing Risks, June 2021. (document), 1.1, 2.1, A.3

[41] Chirag Nagpal, Mononito Goswami, Keith Dufendach, and Artur Dubrawski. Counterfactual phenotyping with censored time-to-events. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3634–3644. ACM, August 2022. doi: 10.1145/3534678.3539110. URL http://dx.doi.org/10.1145/3534678.3539110. 3.1, 3.4

[42] Chirag Nagpal, Mononito Goswami, Keith Dufendach, and Artur Dubrawski. Counterfactual Phenotyping with Censored Time-to-Events. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3634–3644, August 2022. doi: 10.1145/3534678.3539110. 2.4

[43] Chirag Nagpal, Willa Potosnak, and Artur Dubrawski. Auton-survival: An Open-Source Package for Regression, Counterfactual Estimation, Evaluation and Phenotyping with Censored Time-to-Event Data, August 2022. (document), 2.2.3, A.1, A.2

[44] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005. 2.3

[45] Kai Ninomiya, Shigetaka Kageyama, Hiroki Shiomi, Nozomi Kotoku, Shinichiro Masuda, Pruthvi C. Revaiah, Scot Garg, Neil O'Leary, David van Klaveren, Takeshi Kimura, Yoshinobu Onuma, and Patrick W. Serruys. Can machine learning aid the selection of percutaneous vs surgical revascularization? 82 (22):2113–2124. ISSN 0735-1097. doi: 10.1016/j.jacc.2023.09.818. URL https://www.sciencedirect.com/science/article/pii/S0735109723076349. 3.1, 3.4

[46] Kai Ninomiya, Shigetaka Kageyama, Scot Garg, Shinichiro Masuda, Nozomi Kotoku, Pruthvi Revaiah, Neil O'leary, Yoshinobu Onuma, and Patrick Serruys. Can machine learning unravel unsuspected, clinically important factors predictive

of long-term mortality in complex coronary artery disease? a call for "big data". *European Heart Journal - Digital Health*, 4, 02 2023. doi: 10.1093/ehjdh/ztad014. 3.4

[47] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015. doi: 10.1609/aaai.v29i1.9602. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602. 2.2.2

[48] Sandip S. Panesar, Rhett N. D'Souza, Fang-Cheng Yeh, and Juan C. Fernandez-Miranda. Machine Learning Versus Logistic Regression Methods for 2-Year Mortality Prognostication in a Small, Heterogeneous Glioma Database. *World Neurosurgery: X*, 2:100012, April 2019. ISSN 2590-1397. doi: 10.1016/j.wnsx.2019.100012. 2.1, 2.4

[49] Philip C. Prorok, Gerald L. Andriole, Robert S. Bresalier, Saundra S. Buys, David Chia, E. David Crawford, Ronald Fogel, Edward P. Gelmann, Fred Gilbert, Marsha A. Hasson, Richard B. Hayes, Christine Cole Johnson, Jack S. Mandel, Albert Oberman, Barbara O'Brien, Martin M. Oken, Sameer Rafla, Douglas Reding, Wilmer Rutt, Joel L. Weissfeld, Lance Yokochi, and John K. Gohagan. Design of the prostate, lung, colorectal and ovarian (plco) cancer screening trial. *Controlled Clinical Trials*, 21(6, Supplement 1):273S–309S, 2000. ISSN 0197-2456. doi: https://doi.org/10.1016/S0197-2456(00)00098-2. URL https://www.sciencedirect.com/science/article/pii/S0197245600000982. 2.2

[50] Vineet K. Raghu, Philicia Moonsamy, Thoralf M. Sundt, Chin Siang Ong, Sanjana Singh, Alexander Cheng, Min Hou, Linda Denning, Thomas G. Gleason, Aaron D. Aguirre, and Michael T. Lu. Deep Learning to Predict Mortality After Cardiothoracic Surgery Using Preoperative Chest Radiographs. *The Annals of Thoracic Surgery*, 115(1):257–264, January 2023. ISSN 00034975. doi: 10.1016/j.athoracsur.2022.04.056. 2.1, 2.4

[51] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, December 2017. 2.1

[52] Pranav Rajpurkar, Jeremy Irvin, Robyn L. Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P. Langlotz, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Francis G. Blankenberg, Jayne Seekins, Timothy J. Amrhein, David A. Mong, Safwan S. Halabi, Evan J. Zucker, Andrew Y. Ng, and Matthew P. Lungren. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Medicine*, 15(11):e1002686, November 2018. ISSN 1549-1676. doi: 10.1371/journal.pmed.1002686. 2.1

[53] Pranav Rajpurkar, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L. Ball, Marc Mendelson, Gary Maartens, Daniël J. van Hoving, Rulan Griesel, Andrew Y. Ng, Tom H. Boyles, and Matthew P. Lungren. CheXaid: Deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *NPJ digital medicine*, 3:115, 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-00322-2. 2.1

[54] Joseph F. Sabik, III, Faisal G. Bakaeen, Marc Ruel, Marc R. Moon, S. Christopher Malaisrie, John H. Calhoon, Leonard N. Girardi, and Robert Guyton. The American Association for Thoracic Surgery and The Society of Thoracic Surgeons Reasoning for Not Endorsing the 2021 ACC/AHA/SCAI Coronary Revascularization Guidelines. 113(4):1065–1068. ISSN 0003-4975. doi: 10.1016/j.athoracsur.2021.12.003. URL https://doi.org/10.1016/j.athoracsur.2021.12.003. 3.1

[55] Hajime Uno, Tianxi Cai, Lu Tian, and Lee-Jen Wei. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537, 2007. 2.2.2, A.4

[56] Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, May 2011. ISSN 1097-0258. doi: 10.1002/sim.4154. 2.2.2, A.4

[57] Salim S. Virani, L. Kristin Newby, Suzanne V. Arnold, Vera Bittner, LaPrincess C. Brewer, Susan Halli Demeter, Dave L. Dixon, William F. Fearon, Beverly Hess, Heather M. Johnson, Dhruv S. Kazi, Dhaval Kolte, Dharam J. Kumbhani, Jim LoFaso, Dhruv Mahtta, Daniel B. Mark, Margo Minissian, Ann Marie Navar, Amit R. Patel, Mariann R. Piano, Fatima Rodriguez, Amy W. Talbot, Viviany R. Taqueti, Randal J. Thomas, Sean van Diepen, Barbara Wiggins, and Marlene S. Williams. 2023 aha/acc/accp/aspc/nla/pcna guideline for the management of patients with chronic coronary disease: A report of the american heart association/american college of cardiology joint committee on clinical practice guidelines. *Circulation*, 148(9):e9–e119, 2023. doi: 10.1161/CIR.0000000000001168. URL https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000001168. 3.1

[58] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels, February 2018. 2.1

[59] S Yusuf, D Zucker, E Passamani, P Peduzzi, T Takaro, L.D Fisher, J.W Kennedy, K Davis, T Killip, R Norris, C Morris, V Mathur, E Varnauskas, and T.C Chalmers. Effect of coronary artery bypass graft surgery on survival: overview of 10-year results from randomised trials by the coronary artery bypass graft surgery trialists collaboration. *The Lancet*, 344(8922):563–570, 1994. ISSN

0140-6736. doi: https://doi.org/10.1016/S0140-6736(94)91963-1. URL https://www.sciencedirect.com/science/article/pii/S0140673694919631. Originally published as Volume 2, Issue 8922. 3.1

[60] Jianpeng Zhang, Yutong Xie, Yi Li, Chunhua Shen, and Yong Xia. COVID-19 Screening on Chest X-ray Images Using Deep Learning based Anomaly Detection. *ArXiv*, March 2020. 2.1

[61] Zhongheng Zhang, Jaakko Reinikainen, Kazeem Adedayo Adeleke, Marcel E. Pieterse, and Catharina G. M. Groothuis-Oudshoorn. Time-varying covariates and coefficients in Cox regression models. 6(7):121. ISSN 2305-5839. doi: 10.21037/atm.2018.02.12. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6015946/. 2.1