Cancer Recurrence and Survival Prediction and Evaluation using Machine Learning

by

Mahtab Farrokh

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

Abstract

As cancer is the leading global cause of death, an ongoing challenge is predicting an individual's cancer progression accurately, to facilitate personalized treatment planning. Individuals diagnosed with cancer may succumb to the illness or face cancer recurrence post-treatment. The first part of this thesis focuses on predicting prostate cancer recurrence using tissue images. Roughly 30% of men with prostate cancer who undergo radical prostatectomy (RP) will suffer biochemical cancer recurrence (BCR). Unfortunately, no current method can effectively predict which patients will experience BCR after RP. We develop and evaluate PathCLR, a novel semi-supervised method that learns a model that can use tissue images along with clinicopathological features to predict prostate cancer recurrence within five years after RP. We built and evaluated models using two prostate cancer datasets: CPCTR and JHU. PathCLR's (10-fold cross-validation) F1 score was 0.61 for CPCTR and 0.85 for JHU, which were statistically superior to the best-learned model that relied solely on clinicopathological features. This finding suggests that there is essential predictive information in tissue images at the time of surgery that goes beyond the knowledge obtained from reported clinicopathological features, helping predict the patient's five-year outcome.

The second part of this dissertation focuses on effective survival prediction and evaluation for cancer patients. In the context of deploying individual survival prediction models, a pivotal question emerges: Are we striving to compare survival durations between patients (*i.e.*, 'Who survives longer between patients A and B?') or are we endeavoring to estimate a specific patient's survival time (*i.e.*, 'How long will patient A survive?'), among other scenarios. We address this fundamental inquiry and conduct a comprehensive evaluation of such predictive models. We consider 9 common solid tumors (breast, lung, prostate, etc.) using data from the Surveillance, Epidemiology, and End Results (SEER) Program. We consider several different possible goals of a survival prediction model and connect each goal to a specific evaluation metric. We propose modified versions of the Mean Absolute Error (MAE) measure tailored to address a query about a patient's expected survival duration. Here, we trained multiple models (including both conventional and advanced machine learning models) on various cancer types and rigorously evaluated those models using the proposed metrics. We demonstrate that a model might be effective for one goal but ineffective for another, and show that we can determine this based on the measure used. Our findings underscore the importance of selecting the evaluation measure that is aligned with the primary objective of a study. This research sets a path for future research that seeks to further refine predictive models for oncological prognostication.

Preface

This thesis is an original work by ‘Mahtab Farrokh’. Chapter 2 reproduces a paper titled “Learning to Predict Prostate Cancer Recurrence from Tissue Images” [1] accepted in the Journal of Pathology Informatics written jointly with Neeraj Kumar, Peter Gann, and Russell Greiner. My contributions were to design, develop, and evaluate the proposed method, and write the paper under the supervision of Russell Greiner and Peter Gann. This research was an international research collaboration with Professor Peter Gann from the Department of Pathology, College of Medicine, University of Illinois at Chicago, United States. Professor Gann provided access to the clinical dataset and shared his expertise in the clinical field.

Chapter 3 reproduces a paper submission titled “Effective Survival Prediction for Cancer Patient” written jointly with Shi-ang Qi, Neeraj Kumar, and Russell Greiner. I contributed by developing and evaluating models to validate the paper’s objective and was responsible for drafting the manuscript under the supervision of Russell Greiner.

"Happiness can be found even in the darkest of times. If one only remembers to turn on the light."

-Dumbledore, J.K. Rowling

To my father and all the people fighting cancer.

Acknowledgements

First of all, I express my deepest appreciation to my supervisor, dear Russ Greiner, the best supervisor that I could have ever asked for. Working with him was truly a dream come true. Thank you for the great human being that you are, and for all that I have learned from you. Thank you for all those times that you reminded me of “Let’s not lose hope” when everything seemed to not be working.

Next, I want to thank the members of the Greiner’s lab, a group of kind and intelligent individuals with whom I had the privilege of working and learning. A special acknowledgment goes to Shi-ang Qi and Neeraj Kumar for generously sharing their knowledge, answering many of my questions, and providing support throughout these past years. I’d like to thank Roberto Vega Romero, for his help during the initial phases of my work and for guiding me in utilizing Google Cloud GPUs. Furthermore, I want to thank Dr. Peter Gann for his valuable clinical insights and for addressing many of my inquiries.

I am also grateful to the Alberta Machine Intelligence Institute (AMII) for providing funding and organizing many social events that brought me closer to many people and gave me a sense of having a community. A special thanks goes to Warren Johnston and Erika Hudon for their sincere and kind support over the past two years.

I extend immense gratitude to my dear friends, as it is through their presence that Edmonton feels like home to me: Faezeh Haghverd, Jordan Coblin, Nikoo Aghaei, Bedir Tapkan, Ehsan Kashfi, Blanca Miller, Vida Goudarzi, Ardavan Mofidi, Hamza Emra, Kiarash Aghakasiri, Amirmohsen Sattarifard, Tahmineh Mohati, Parnian Mehinrad, Milad Zamani, and Aidan Bush.

Also, I consider myself fortunate to have dear friends located beyond Edmonton who regularly check in on me from different parts of the world: Zahra Kalanaki, Monireh Safari, Sina Malakouti, Monireh Seifollahi, Mohsen Sadeghi, Sepideh Mollanorouzi, Melika Mohsenirad, and Mahta Kiaei.

Lastly, I want to express my gratitude and love to my parents, Maryam Ebrahimi and Nader Farrokh, for supporting me and always being there for me, even though it is hard to show their support from a distance. I love you.

Table of Contents

List of Tables

List of Figures

Abbreviations

BCR: Biochemical Cancer Recurrence.

BS: Brier Score.

CI: Confidence Interval.

C-index: Concordance index.

CPCTR: Cooperative Prostate Cancer Tissue Resource.

ECE: Extra Capsular Extension.

ECP: Established Capsular Penetration.

FCP: Focal Capsular Penetration.

GS: Gleason Sum.

IoU: Intersection over Union.

ISD: Individualized Survival Distribution.

JHU: Johns Hopkins University.

KM: Kaplan Meier.

LN: Lymph Node Invasion.

MAE-PO: Mean Absolute Error Pseudo-Observation.

MAE: Mean Absolute Error.

ML: Machine Learning.

MTLR: Multi-Task Logistic Regression.

PathCLR: Pathology Contrastive LeaRning.

PQ: Panoptic Quality.

PSA: Prostate-Specific Antigen.

RP: Radical Prostatectomy.

SEER: Surveillance, Epidemiology, and End Results.

SM: Surgical Margin.

SVI: Seminal Vesicle Invasion.

TL-MAE-PO: Truncated Log Mean Absolute Error Pseudo-Observation.

T-MAE-PO: Truncated Mean Absolute Error Pseudo-Observation.

TMA: Tissue MicroArray.

Chapter 1

Introduction

Cancer is the leading cause of death in the world, with approximately 10 million deaths and 19.3 million new cancer cases in 2020 [2]. Among these newly-detected cancer patients, those of the breast, lung, colorectal, prostate, and stomach are the most frequent cancer types. Numerous cancer-related challenges and applications exist where machine learning (ML) can be effectively applied to develop and train models. These models can aid oncologists in making decisions about personalized treatments and end-of-life care, predicting patient responses to different therapies, assessing risks of cancer recurrence, and optimizing dosages of chemotherapy, among other aspects. Additionally, ML can streamline various time-intensive tasks like cell and tumor annotation in pathology slides, analysis of genomic data, and more.

Among all of these possible applications, this dissertation discusses two of them: First, the focus is on predicting cancer recurrence in patients with prostate cancer who undergo radical prostatectomy (RP) surgery. RP involves the complete removal of the prostate gland and adjacent lymph nodes, serving as a treatment for localized prostate cancer in men. However, there exists a likelihood of experiencing biochemical cancer recurrence (BCR) typically within 5 years post-RP surgery. BCR is identified by an elevation in prostate-specific antigen (PSA) levels detected in blood tests conducted after RP surgery, indicating the occurrence of either local recurrence or the spread of cancer to distant sites. The objective of this task is framed as a binary classification specific to the 5-year time point, with the goal of predicting whether a patient will or will not experience BCR within five years after undergoing RP surgery.

Presently, physicians employ the CAPRA-S scoring method to estimate the likelihood of biochemical cancer recurrence (BCR), relying solely on clinicopathological features. Clinicopathological features encompass both clinical factors (e.g., age, race) and pathology-related characteristics determined by a pathologist during the analysis of tissue images. An example of a pathology-related feature is the Gleason Score, which characterizes the abnormality and aggressiveness of cancer cells based on their appearance under a microscope. In this work, we explore if using information from tumor tissue slides obtained after surgery would help to produce a more accurate prediction than a prediction that is based on only clinicopathological features.

We specifically focus on more challenging cases for BCR prediction, with matching clinicopathological characteristics, where the majority of patients fall into an intermediate-risk category, which is more complex to manage compared to the lower (less aggressive cancer) or higher scores (more aggressive). Therefore, Chapter 2 of this thesis explores how to accurately predict BCR within five years after RP surgery. Our contribution in this work is that we show there is critical information in tissue slides obtained after surgery by introducing a new method, called PathCLR (Pathology Contrastive LeaRning) – an automated pipeline capable of efficiently processing a patient's tissue cores and providing BCR probability predictions in less than two seconds. Additionally, we compare our method with a supervised CNN model, CAPRA-S score, and a previous benchmark for BCR prediction, and we demonstrate that our pipeline predicts recurrence more accurately than these alternative approaches.

Secondly, we engage with another challenging task: rather than considering only predicting a binary result at 5 years, here we instead predict survival probability over all future time points based on a patient's clinical features. It is important to emphasize that this investigation is centered on forecasting survival probability, distinct from the earlier study which concentrated on cancer recurrence prediction. Furthermore, this study expands beyond the exclusive focus on prostate cancer survival prediction. It encompasses nine common solid tumors, specifically those affecting the brain, breast, kidney, liver, lung, stomach, prostate, thyroid, and urinary bladder.

To predict survival probability over all future time points, we learn various individualized survival distribution (ISD) predictor models, that each predict a patient's expected time from detection to death. We elaborate on the significance of ISD models and how they offer more comprehensive information about individual patients.

Moreover, this study has a specific focus on rigorously and effectively evaluating ISD models. We discuss the need for evaluation metrics that are relevant to the research objectives and as a result, we identify which objective leads to which evaluation metric. Additionally, this research places a particular emphasis on the thorough and effective evaluation of ISD models. We discuss the general importance of utilizing evaluation metrics that align with the research objectives, thereby establishing a clear connection between each objective and its corresponding evaluation metric.

Chapter 3 describes this work, and our distinctive contribution lies in proposing a more effective way of evaluating survival prediction models by extending the recent MAE Pseudo-Observation (MAE-PO) measure for evaluating survival models and introducing the truncated and truncated-log variations of MAE-PO (T-MAE-PO and TL-MAE-PO). We train and evaluate multiple common ISD predictor models for each cancer type, and identify the top-performing model for each specific cancer type.

Lastly, Chapter 4 concludes this thesis and discusses potential future applications and research directions.

Chapter 2

Learning to Predict Prostate Cancer Recurrence from Tissue Images

2.1 Introduction

World-wide each year, around 1.6 million men are diagnosed with prostate cancer, and 366,000 will die of the disease [3]. A majority of these patients undergo radical prostatectomy, either as an initial treatment choice or following a period of active surveillance [4]. Within five years of the surgery, about 20% to 40% of these men who undergo radical prostatectomy experience biochemical cancer recurrence (BCR), which is detected by elevated prostate-specific antigen (PSA) levels [5]. The accurate prediction of patients prone to experiencing BCR after surgery is crucial for determining the most appropriate post-surgery course of action. This includes identifying patients who might benefit from additional treatment, advanced imaging to detect metastases, genomic testing, or more frequent monitoring.

More specifically, while it is true that most patients with BCR do not die from the disease, it is still a strong risk factor for subsequent metastasis and mortality. Patients with a “Yes” prediction could benefit from more frequent PSA monitoring. In addition, this “Yes” prediction suggests that the patient could benefit from new imaging techniques (e.g., PSMA-targeted PET) (PSMA = prostate-specific mem-

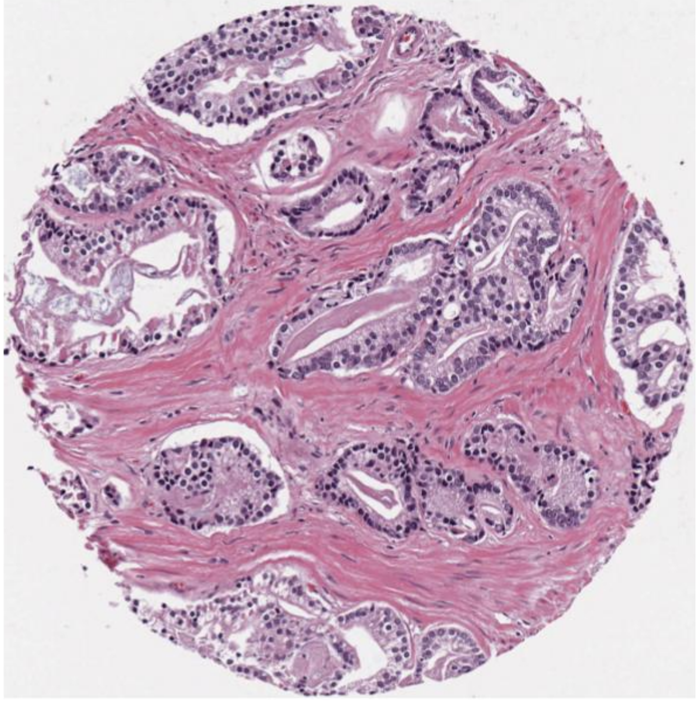

Figure 2.1: Example of a hematoxylin and eosin (H&E) stained tissue microarray (TMA) image.

brane antigen; PET = positron emission tomography) to detect occult metastases or local/regional spread soon after surgery. If this produces visualized lesions, the patient could opt for local or systemic therapy or both.

Due to the frequency of BCR occurrence, several projects have been developed to create multivariable prediction tools aimed at predicting BCR at the time of diagnosis. These tools include models utilizing clinicopathological features and gene profiling [6]. One such model, the CAPRA-S score, employs clinicopathological features such as PSA level, Gleason sum score, and other pathology-reported variables to calculate a risk score associated with progression-free probability at three and five years; it has shown favorable results for BCR risk assessment [7, 8]. Additionally, some approaches use gene profiling biomarkers, such as the Decipher test, on bulk samples of surgical tissue, to forecast recurrence after prostatectomy [6, 9, 10]. Despite the advancements facilitated by these models, widespread clinical adoption has not occurred due to cost barriers or failure to achieve the required level of accuracy.

Therefore, in this study, we introduce a novel deep learning-based method to construct BCR prediction models that leverage both clinicopathological features and information from inexpensive, routinely available images of tumor tissue slides stained with hematoxylin and eosin (H&E), obtained after surgery. Figure 2.1 shows an example of H&E stained tissue image. Recent research has already demonstrated that deep neural networks can identify recurring tumors [11]. These earlier systems rely on common approach *supervised* learning, as each involves directly learning a model that labels each instance with its outcome. By contrast, our method, called PathCLR (Pathology Contrastive LeaRning), takes a *semi-supervised* approach, which involves a self-supervised learning step before the supervised step. Note that emerging evidence suggests that semi-supervised learning often outperforms traditional supervised learning methods, especially when labeled data is limited, as is often the case in medical research [12–14].

This motivates our PathCLR approach, which first trains a self-supervised SimCLR model [15] based on tissue microarray images (TMAs) (see Figure 2.1 as a TMA example) to generate a set of latent representations for a given tissue core of a new patient. Note this self-supervised step is *unsupervised*, as it does not require a BCR label for each instance. Subsequently, PathCLR utilizes this set of latent representations, along with relevant clinical features and reported pathology variables of the patient, to train a neural network (NN) classifier. The goal is to predict whether a specific patient will experience BCR within five years after surgery. This personalized approach sets PathCLR apart from previous methods, as PathCRL (1) provides a prediction about an individual patient (and not about *groups* of patients) and (2) provides a specific individual Yes, No prediction for the patient, (and not a hazard ratios relative to a baseline).

Our research findings demonstrate that incorporating a novel semi-supervised machine learning method shows promise for rapid and inexpensive prediction of BCR directly from routinely stained tumor images, especially when combined with clinicopathological variables. Our proposed method is cost-effective because many centers now routinely capture high-resolution H&E scans for all cancer types, with each image costing approximately $4 and a scan duration of around 1 minute per slide (or about 3 minutes at 40x magnification). Consequently, PathCLR, leveraging these readily available images without the need for additional expert annotations, presents a highly cost effective approach.

We suggest that, indeed, the pathology slides contain valuable information beyond the conventional variables such as Gleason grade, PSA level, surgical margin, etc. Furthermore, we demonstrate that the label-free learning approach on TMA cores, which produces latent representations of their histological patterns, plays a critical role in achieving improved BCR prediction accuracy. Our main contributions are:

1. Our work focuses on challenging datasets where the majority of Gleason grade sum scores are 7 (3 + 4 or 4 + 3). This differs from many previous studies that focus on BCR prediction without matching patients with respect to their clinicopathological characteristics, specifically the Gleason sum score. As we conduct a study on BCR on a carefully selected cohort of patients with matched clinicopathological variables, we focus on more challenging cases for BCR prediction.

2. We introduce PathCLR, an automated pipeline capable of efficiently processing a patient's TMA cores and providing BCR probability predictions in less than two seconds.

3. To demonstrate the effectiveness of our pipeline in recurrence prediction, we compare our semi-supervised PathCLR method with a supervised CNN model, CAPRA-S score, and a previous benchmark [16] for BCR prediction. This comparison shows that our pipeline predicts recurrence more accurately than these alternative approaches.

This chapter is organized as follows: Section 2.2 describes the two datasets that we used and then proposes the PathCLR pipeline. Section 2.3 provides the evaluation metrics that we used in Section 2.4 to present our results. We compare our proposed approach to previous work on BCR prediction in Section 2.5, before concluding in Section 2.6.

2.2 Material and Methods

2.2.1 Tissue microarrays and image pre-processing

We utilized images and data from two prostate cancer cohorts: TMAs and clinicopathological data from the Cooperative Prostate Cancer Tissue Resource (CPCTR) [17], funded by the National Cancer Institute, USA and the PSA Progression dataset from the Prostate Cancer Biorepository Network (PCBN), funded by the Prostate Cancer Research Program of the US Department of Defense. We refer to the later cohort as JHU as all patients in this dataset were treated at Johns Hopkins University.

Both TMAs in our study employed a matched case-control design to define recurrence based on BCR or clinical progression after surgery. Guidelines describe post-RP BCR defined as a serum PSA of more than 0.2 ng/ml for both CPCTR and JHU datasets. The BCR label in both of the datasets is based on a single evaluation unless the reported PSA level was very close to the threshold. However, they differed in their approach to selecting control tumors. In the CPCTR dataset, controls were chosen from patients who survived at least five years without experiencing BCR. The matching was done on a 1:1 basis, considering factors such as age, race, Gleason grade (primary and secondary), and the treating hospital.

On the other hand, the JHU cohort used an incidence density sampling approach. For each case with BCR, one control was selected from the pool of patients who had not experienced BCR up to that specific time point. Case-control matching was done based on age, race, pathologic stage, and Gleason sum. This sampling approach allowed for the possibility that initial controls could later become cases and also that controls could be selected more than once.

Since our primary goal was to compare the risk of developing BCR within five years, we limited the selection of control patients to those who survived the entire five-year period without experiencing BCR. TMA images were pre-processed to remove spots with insufficient tissue, ensuring that those with more than 80% white background were eliminated. The flow chart in Figure 2.2 shows how images from recurrent or non-recurrent patients were selected for final analyses. The CPCTR dataset comprised 189 cases and 185 controls, with a total of 1,281 TMA cores. In the JHU dataset, there were 451 cases and 195 controls comprising 2,983 cores. Both datasets included up to four cores per patient.

The main characteristics of the patients and tumors included in our analysis are summarized in Table 2.1 and Table 2.2, for CPCTR and JHU datasets, respectively. Note that the PSA feature in Tables 2.1 and 2.2 represents the PSA laboratory readings immediately post-prostatectomy. This is distinct from the later rise in PSA levels, which indicates BCR. The TMA image resolution in the CPCTR dataset is 40x, while in the JHU dataset is 20x. We utilized the Python programming language along with OpenCV[1] and the Python Image Library (PIL)[2] for image visualization.

To fine-tune the patch extractor component of PathCLR, we employed the MoNuSAC dataset [18] for learning a model to segment epithelial cells. This dataset is significant as it is extensive, diverse, and hand-annotated, specifically designed for the MoNuSAC2020 challenge, which aimed to identify cancerous epithelial cells and three types of immune cells from H&E stained tissue images. Notably, the top methods submitted to this challenge achieved inter-human agreement levels in their performance.

[1] https://opencv.org/

[2] https://pillow.readthedocs.io/

Table 2.1: Characteristics of CPCTR dataset after data cleaning. Abbreviations: Gleason Sum (GS), Prostate Specific Antigen (PSA), Lymph Node Invasion (LN), Seminal Vesicle Invasion (SVI), Surgical Margin (SM), Extra Capsular Extension (ECE)

CPCTR	**Total, n(%)**	**No-BCR, n(%)**	**BCR, n(%)**
#Patients	374	185	189
GS = 3 + 2	2 (0.5)	1 (0.5)	1 (0.5)
GS = 3 + 3	85 (22.7)	41 (22.1)	44 (23.2)
GS = 3 + 4	196 (52.4)	98 (52.9)	98 (51.8)
GS = 3 + 5	7 (1.8)	2 (1.1)	5 (2.6)
GS = 4 + 3	56 (14.9)	27 (14.5)	29 (15.3)
GS = 4 + 4	18 (4.8)	11 (5.9)	7 (3.7)
GS = 4 + 5	9 (2.4)	5 (2.7)	4 (2.1)
GS = 5 + 3	1 (0.2)	0 (0.0)	1 (0.5)
PSA	10.5 ± 11.5	8.8 ±6.18	12.2 ± 14.8
LN	17 (4.5)	6 (3.2)	11 (5.8)
SVI	17 (4.5)	8 (4.3)	9 (4.7)
SM = tumor free	226 (60.4)	124 (67)	102 (53.9)
SM = tumor focal at margin	114 (30.4)	49 (26.4)	65 (34.3)
SM = tumor widespread at margin	30 (8.6)	9 (4.8)	21 (11.1)
SM = unknown	4 (1)	3 (1.6)	1 (0.5)
ECE = none	235 (62.8)	111 (60)	124 (65.6)
ECE = multifocal	28 (7.4)	13 (7)	15 (7.9)
ECE = focal	98 (26.2)	56 (30.2)	42 (22.2)
ECE = established	12 (3.2)	4 (2.16)	8 (4.2)
ECE = unknown	1 (0.2)	1 (0.5)	0 (0)

2.2.2 The PathCLR model

Taking inspiration from the success of recent studies in applying machine learning techniques to the medical field, we propose a semi-supervised method for BCR prediction. Semi-supervised learning has proven to achieve notable performance using

Table 2.2: Characteristics of JHU dataset after data cleaning. Abbreviations: Gleason Sum (GS), Prostate Specific Antigen (PSA), Lymph Node Invasion (LN), Seminal Vesicle Invasion (SVI), Surgical Margin (SM), Established Capsular Penetration (ECP), Focal Capsular Penetration (FCP)

JHU	**Total, n(%)**	**No-BCR, n(%)**	**BCR, n(%)**
#Patients	646	195	451
GS = 2 + 3	2 (0.3)	2 (1.0)	0 (0)
GS = 3 + 2	5 (0.7)	3 (1.5)	2 (0.4)
GS = 3 + 3	99 (15.3)	49 (25.1)	50 (11.0)
GS = 3 + 4	263 (40.7)	94 (48.2)	169 (37.4)
GS = 3 + 5	13 (2.0)	3 (1.5)	10 (2.2)
GS = 4 + 3	136 (21.0)	29 (14.8)	107 (23.7)
GS = 4 + 4	60 (9.2)	8 (4.1)	52 (11.5)
GS = 4 + 5	56 (8.66)	6 (3.0)	50 (11.0)
GS = 5 + 3	3 (0.4)	0 (0)	3 (0.6)
GS = 5 + 4	9 (1.3)	1 (0.5)	8 (1.7)
PSA	11.72 ± 9.66	10.21 ± 7.75	12.38 ± 10.32
LN	99 (15.3)	11 (5.6)	88 (19.5)
SVI	149 (23.0)	29 (14.8)	120 (26.6)
SM = postive	196 (30.3)	35 (17.9)	161 (35.6)
ECP = positive	413 (63.9)	111 (56.9)	321 (71.1)
FCP = positive	235 (36.3)	102 (52.3)	133 (29.4)

fewer labeled samples for training when compared to previous fully supervised approaches [12]. In our proposed PathCLR algorithm, the initial step involves learning the latent representations in a self-supervised manner, which means no diagnostic label is required during this training phase. Subsequently, PathCLR utilizes these latent representations to describe each training patient and then employs their outcome labels to train a model capable of predicting BCR in a supervised manner.

To learn the feature representation of tissue images, we adopted the state-of-the-art self-supervised SimCLR algorithm, which is a simple framework for contrastive

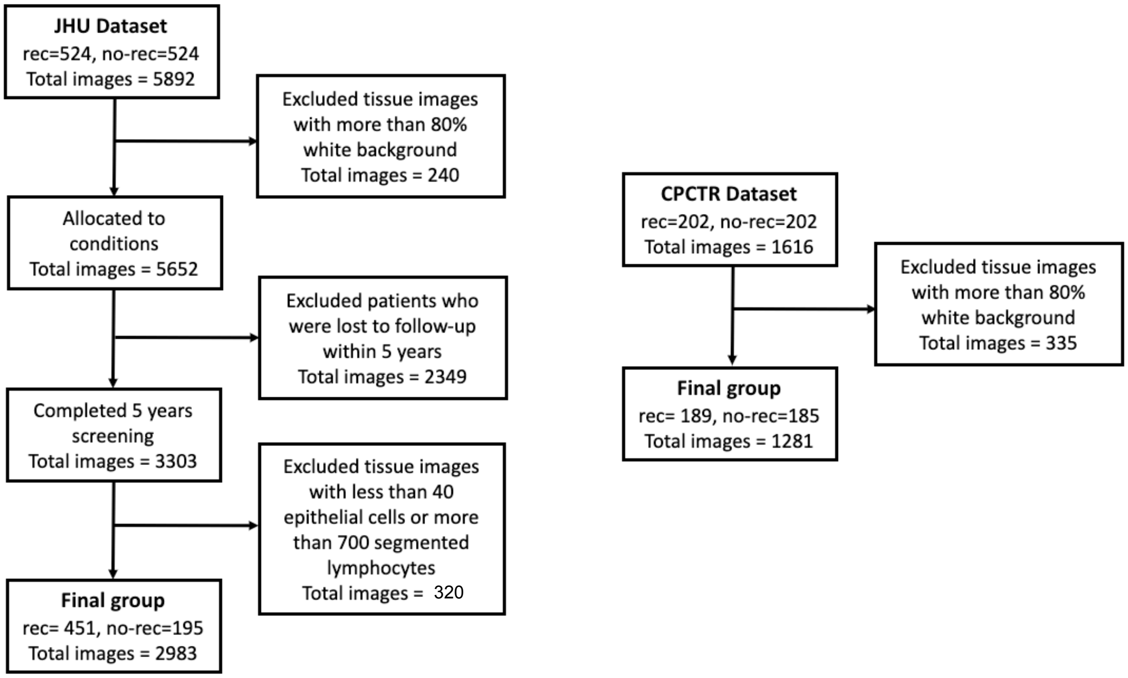

Figure 2.2: Pre-processing flow chart for JHU (left) and CPCTR (right) datasets.

learning of visual representations. In essence, SimCLR learns the latent representation of an image by maximizing agreement between differently augmented views of the same image through a contrastive loss [15]. To use the training examples effectively, we implemented various data augmentations, such as random cropping, color distortion, and blurring. We employed contrastive learning [19] to acquire highly contrasting features from TMA cores. An advantage of the proposed PathCLR model is its adaptability with varying numbers of TMA cores per patient.

To begin our analysis, we recognize the importance of preprocessing tissue images, considering that these images may vary in size and degree of centering. Extracting informative image patches is a crucial initial step. Additionally, we believe that the most relevant regions for BCR prediction are likely to be around epithelial cells and the surrounding stroma, as demonstrated in a prior study [20].

To segment epithelial cells, we employed the HoVer-Net model [21], known for its capabilities in automatic nuclear instance segmentation and classification in histology images. This model simultaneously detects epithelial cells, lymphocytes, macrophages, and neutrophils. The HoVer-Net model is currently a state-of-the-art model for segmenting these nuclei with high accuracy. We fine-tuned the HoVer-Net model on the MoNuSAC2020 dataset, utilizing image augmentation and pretrained weights to achieve even higher segmentation accuracy [18].

To accommodate SimCLR's requirement of fixed-size inputs, we randomly sampled 200 patches, each measuring 128×128 pixels in spatial dimensions, from every TMA image. These patches were centered at the positions of each nucleus detected by the HoVer-Net model. However, the number of detected epithelial cells in each TMA core varied, which could differ from the requirement that each patch include exactly 200 nuclear centers.

To address this, if we identified more than 200 epithelial cells in a TMA core, we employed K-Means clustering to group the (x, y) epithelial cell center positions into 200 clusters. From these clusters, we selected the center of each cluster as one of the 200 regions of epithelial cells of interest. This clustering approach aimed to maximize the coverage of tumor-rich areas and minimize the overlapping of extracted patches.

For TMA cores with fewer than 200 epithelial cells, we took the (x, y) spatial position of the k detected epithelial cell centers and randomly added $200-k$ additional spatial positions in the vicinity of the detected cells. This process allowed us to generate 200 patches for each TMA core. Figure 2.3 visually illustrates the described patch extraction process from each TMA core image.

As depicted in Figure 2.4, the PathCLR pipeline consists of two main components. First, the self-supervised SimCLR model learns the latent representation of all extracted tissue patches. In this step, we used ResNet34 [22] as the base architecture in SimCLR to produce latent representations of input image patches.

Secondly, the final supervised binary classification neural network incorporates all relevant information specific to each patient. This includes both the imaging features expressed through the SimCLR encoding and seven observed clinicopathological features, such as PSA level, surgical margin, Gleason grade, and others. The objective

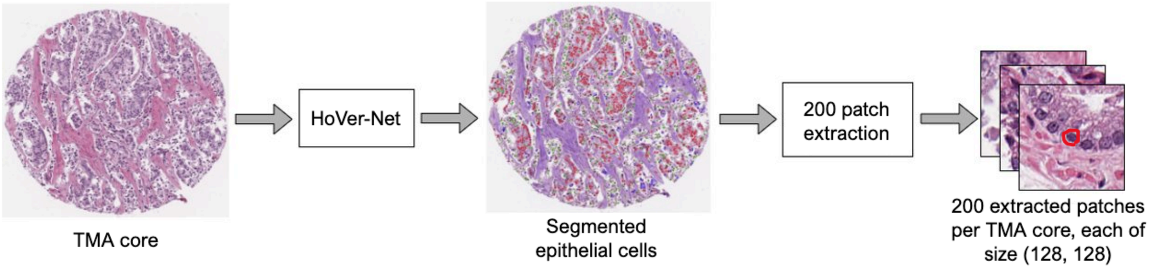
TMA core
HoVer-Net
Segmented epithelial cells
200 patch extraction
200 extracted patches per TMA core, each of size (128, 128)

Figure 2.3: The first part of the PathCLR process extracts 200 patches in size 128 by 128 pixels in RGB format around segmented epithelial cells. The output of HoVer-Net is a segmented tissue core that shows each epithelial cell with a red circle.

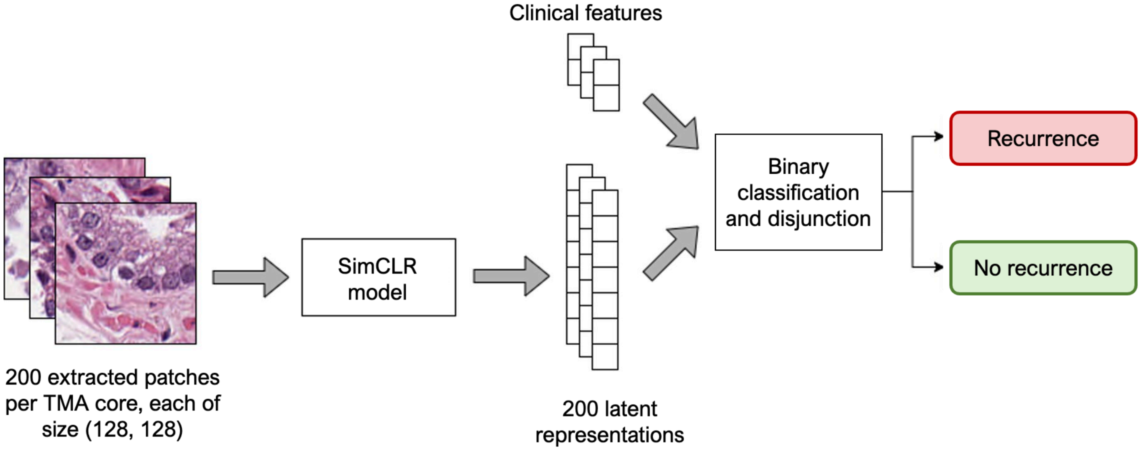
Clinical features
200 extracted patches per TMA core, each of size (128, 128)
SimCLR model
200 latent representations
Binary classification and disjunction
Recurrence
No recurrence

Figure 2.4: The second part of the PathCLR pipeline, which predicts prostate cancer recurrence using 200 extracted patches per TMA core. First, SimCLR learns the latent representation of each tissue patch. Next, a learned binary classifier predicts BCR using the learned latent representations along with 7 clinicopathological features, following a disjunction function to decide patient-level recurrence.

of this step is to train a neural network model that can accurately predict whether a patient will experience BCR within five years after surgery or not.

Note that the final supervised binary classification neural network can capture nonlinear interactions among various input features, encompassing both latent representations of TMA image patches and clinicopathological variables. This capability is important for accurate BCR prediction.

The described binary classifier predicts the recurrence probability within five years for each extracted patch. We then compute the average recurrence probabilities of all patches from a single TMA core. If the average probability from a TMA core is higher than a specified threshold (in this case, 0.6), we predict that the patient will experience BCR. The threshold was empirically determined by internal cross-validation. If a patient has multiple tissue cores available, we calculate the average recurrence prediction probability for each TMA core and return the "disjunction" – indicating a positive prediction if any of the tissue cores is predicted as a recurrence case. We considered various combination rules (e.g., conjunction, majority) but found that the disjunction method performed better (see Section 2.4).

We also compared the results from the binary classification neural network using (1) only clinicopathological features, (2) only the SimCLR output, and (3) both input sources combined. The PathCLR pipeline is implemented using the Python programming language and the Keras library[3]. The HoVer-Net and SimCLR parts of the pipeline are implemented using the PyTorch library[4]. Image pre-processing and transformations were done using the transformers in the torchvision[5] library. Table 2.3 provides more information regarding the main components of the PathCLR pipeline. Note, that the provided configurations for SimCLR in this table are given for the fine-tuning phase because we used pre-trained weights of SimCLR.

Table 2.3: Configuration and overview of components of PathCLR.

Model	**#Parameters**	**#Layers**	**Activation**	**Optimization**	**# Epochs**	**Batch size**
SimCLR	21M	34	ReLU	SGD/Adam	100	256
Binary Classifier	32K	4	ReLU	Adam	200	32

2.3 Evaluation Metric

In the context of the BCR binary classification task, selecting an appropriate evaluation metric is crucial. In our study, we reported two main metrics: accuracy and F1-score:

[3] https://keras.io/
[4] https://pytorch.org/
[5] https://pytorch.org/vision/

$$\text{Accuracy} \quad = \quad \frac{TP \ + \ TN}{TP \ + \ FP \ + \ TN \ + \ TP} \tag{2.1}$$

$$\text{F1} \quad = \quad \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} \quad = \quad \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{2.2}$$

where TP (resp., TN, FP, FN) is the number of true positive (resp., true negative, false positive, false negative) instances. These metrics are commonly used for binary classification tasks and provide valuable insights into the performance of our predictive model.

To assess the performance of PathCLR, we employed 10-fold cross-validation and reported a 95% confidence interval. This involved training-&-testing 10 times, with each iteration training on nine parts (representing 90% of the dataset) and testing on one left-out part (representing 10% of the dataset). For evaluating our epithelial segmentation, we used the panoptic quality (PQ) metric,

$$\text{PQ} \quad = \quad \frac{\sum_{(p,g)\in TP} IoU(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{2.3}$$

which was also used in the MoNuSAC2020 data challenge to evaluate the effectiveness of nuclear segmentation and classification algorithms. Here, p represents the predicted set of pixels in the nucleus, g represents the set of pixels in the ground truth, and $IoU(p, g)$ calculates the Intersection over Union (IoU) by dividing the number of elements in the intersection of the sets by the size of their union.

To determine the significance of our results, we utilized a paired two-sided t-test with a significance level of $p < 0.05$ – *i.e.*, a result is considered significant if the p-value obtained from the t-test is less than 0.05.

2.4 Results

As described in Section 2.2, the initial step of the PathCLR pipeline involves segmenting epithelial cells and utilizing the segmented epithelial cell centers to extract 200 patches from a TMA core. Our fine-tuned HoVer-Net model demonstrated superior performance compared to the winning results from the MoNuSAC2020 challenge, indicating its promising capabilities for epithelial cell detection. While the best reported PQ for the validation dataset in the MoNuSAC2020 challenge website is 0.611, our model achieved a higher PQ of 0.675, showing its improved performance in epithelial cell segmentation.

To investigate the impact of using H&E-stained tissue slides on recurrence prediction, we trained models in three distinct settings:

1. Using only clinicopathological features
2. Using only H&E-stained TMA cores
3. Using both clinicopathological features and H&E-stained TMA cores

In order to conduct a fair evaluation of setting #1, we systematically explored various combinations of clinicopathological features and different learning models. After thorough experimentation, we found that the best outcome was achieved by employing a two-layer feed-forward neural network with cross-entropy loss. Importantly, the optimal result was obtained by using only the precise set of CAPRA-S score's features [7]. The selected clinicopathological features for the first setting are as follows:

clinicopathological_features = { PSA Level, Surgical Margin, Primary Gleason Grade, Secondary Gleason Grade, Extracapsular Extention, Lymph Node, Seminal Vesicle }

In setting #3, we maintained the same listed set of clinicopathological features for a fair comparison with the other settings. In both settings #2 and #3, we employed the SimCLR network using ResNet34 with Xavier weight initialization [23]. Note, the other setting #1 does not use ResNet34.

Table 2.4:
10-fold cross-validation result with 95% CI. Setting #1 uses only clinicopathological features, setting #2 uses only H&E stained TMA cores, and setting #3 uses both clinicopathological features and H&E stained TMA cores.

	Accuracy		**F1-Score**	
	CPCTR	**JHU**	**CPCTR**	**JHU**
Setting #1	53.7 ± 2.77	69.81 ± 0.54	0.544	0.82
Setting #2	56.1 ± 3.47	69.82 ± 0.39	0.6	0.82
Setting #3	**59.5 ± 3.29**	**75.25 ± 3.48**	**0.614**	**0.85**

To handle binary classification in all three settings, we experimented with different numbers of layers and regularization techniques (both L1 and L2 regularization penalties along with early stopping) to prevent overfitting and optimize the model's performance.

Table 2.4 and Figure 2.5 present the results obtained from 10-fold cross-validation on both the CPCTR and JHU datasets. Notably, using solely clinicopathological features yields results that are almost on par with a trivial "just say BC" prediction, where BCR is predicted for all cases regardless of their features. This "just say BCR" prediction would achieve an accuracy of 69.81% (451/646) on the JHU dataset and 50.5% (189/374) accuracy on the CPCTR dataset.

We assume the reason clinicopathological variables alone are insufficient for BCR prediction lies in the fact that the case-control pairs are carefully matched for clinicopathological features in both the CPCTR and JHU datasets. Consequently, there are patients with similar clinicopathological features but different outcomes. As a result, classifiers relying solely on clinicopathological features struggle to distinguish between recurrence and non-recurrence cases, which hampers their predictive performance.

In setting #1, we compared the results of our pipeline with a conventional model that relies solely on the CAPRA-S score [7]. However, upon experimentation, we observed that applying a learned neural network to the clinicopathological features yielded slightly better results. We attribute this improvement to the neural network's ability to discover non-linear patterns within the features, enhancing the predictive accuracy.

When using the CAPRA-S score with 10-fold cross-validation, we obtained an accuracy of 53.22 ± 5.93 on the CPCTR dataset, and 69.81 ± 0.0 on the JHU dataset. These results indicate that even the CAPRA-S score fails to accurately predict BCR in cases where patients have matched clinicopathological features, highlighting the challenges in making accurate predictions solely based on clinicopathological variables.

In setting #3, where we combined the clinicopathological features with latent representations of tissue cores, we observed that the accuracy of this model is at least 5% better than using only clinicopathological features. This improvement demonstrates that there is valuable additional information related to cancer recurrence embedded in the tissue slides. To confirm the significance of this improvement, we conducted paired two-sided t-tests on both the JHU and CPCTR datasets, resulting in p-values of 0.005 and 0.04, respectively. Both p-values are below the significance threshold of 0.05, indicating that the accuracy improvement is statistically significant.

Comparing setting #2 to setting #1, we found that the accuracy and F1-score of setting #2 were higher. However, the paired t-tests revealed that the difference between the two settings was not statistically significant, with p-values of 0.19 in the JHU dataset and 0.48 in the CPCTR dataset. Despite not being statistically significant, the improved accuracy in setting #2 suggests the potential benefits of incorporating the learned latent representations of tissue cores along with clinicopathological features in BCR prediction. We also compared setting #2 versus setting #3: for the JHU dataset, setting #3 is significantly better ($p = 0.007$). However, for the CPCTR dataset, there is no ($p < 0.05$) significant difference ($p = 0.17$). Note, we did not apply any false discovery rate (FDR) correction since the number of involved comparisons is small.

As mentioned in the previous section, we have up to 4 TMA cores per patient,

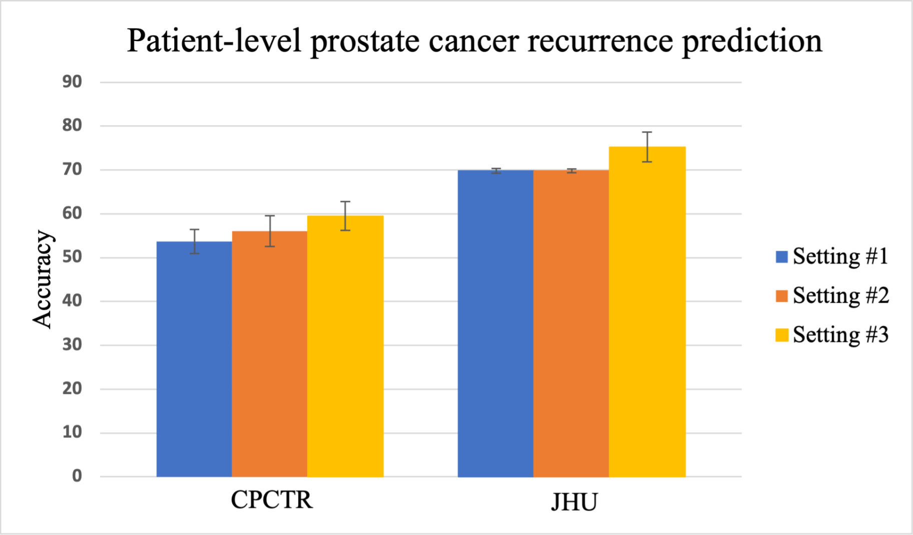

Figure 2.5: Accuracy on JHU and CPCTR datasets 10-fold cross-validation. Error bar shows the 95% CI.

and we predicted recurrence for each TMA core. To achieve this, we used various functions to combine the individual core scores, including disjunction, conjunction, and majority voting. The results of 10-fold cross-validation accuracy for these variants are presented in Table 2.5 for both the CPCTR and JHU datasets, corresponding to settings #2 and #3.

The results indicate that the accuracy of the disjunction method is higher than that of conjunction and majority voting. We hypothesize that some TMA cores may not provide informative data for cancer recurrence prediction. Therefore, if at least one TMA core shows a high probability of BCR, we consider the corresponding patient to have a high probability of experiencing BCR within five years. This approach allows us to leverage the most informative TMA cores for each patient, resulting in improved accuracy in predicting recurrence at the patient level.

In addition to our PathCLR approach, we explored a fully supervised convolutional neural network (CNN) approach using the ResNet34 model. Instead of training on extracted patches of TMA cores, we trained the CNN directly on the whole tissue cores as images. For patients with multiple TMA cores, we still used the disjunction method for both the PathCLR and supervised models.

The results, as shown in Table 2.6, indicate that the accuracy of the supervised approach was approximately 2% lower than the PathCLR approach on both the CPCTR and JHU datasets. This suggests that our semi-supervised PathCLR approach, which leverages both clinicopathological features and learned latent representations from tissue images, outperforms the fully supervised CNN approach in predicting prostate cancer recurrence.

Drawing inspiration from a previous benchmark by Leo et al. [16], which predicted BCR using H&E slides and meticulously crafted features of gland morphology, we performed a further comparison of PathCLR. However, rather than focusing on the lumen, we segmented the epithelial cells and derived an identical set of Histotyping features via the HistomicsTK library[6]. Our focus shifted to epithelial cells primarily because the TMA cores we examined contained limited lumen regions. Moreover, as previously stated, we hypothesize that epithelial cells provide more valuable information for predicting BCR. The results outlined in Table 2.6 show that PathCLR performs better than these Histotyping features. Hence, the superior performance of PathCLR can be attributed to the learning of each image patch representation through contrastive loss. This embedded information proves to be more effective than the morphology features obtained through Histotyping.

In terms of the run-time computational efficiency of the learned system, the processing time for each TMA core input of a new patient is under two seconds. Considering the number of TMA cores (up to four in this study), the prediction of BCR for an individual, on average, is completed in less than 7.3 seconds.

2.5 Discussion

Recently, the combination of computational pathology with machine learning (ML) techniques has shown the potential to enhance diagnostic accuracy and optimize

[6] https://digitalslidearchive.github.io/HistomicsTK/index.html

Table 2.5: 10-fold cross-validation accuracy with 95% CI in different combination modes.

	Setting #2		**Setting #3**	
	CPCTR	**JHU**	**CPCTR**	**JHU**
Majority	55.08 ± 0.26	**69.82 ± 0.39**	55.08 ± 0.55	74.44 ± 2.1
Conjunction	55.62 ± 0.26	69.82 ± 0.24	55.88 ± 0.55	74.93 ± 1.7
Disjunction	**56.06 ± 3.47**	**69.82 ± 0.39**	**59.49 ± 3.29**	**75.25 ± 3.48**

Table 2.6: Comparing PathCLR, a semi-supervised approach, with supervised and Histotyping approaches, using 10-fold cross-validation accuracy and a 95% CI.

	Setting #2		**Setting #3**	
	CPCTR	**JHU**	**CPCTR**	**JHU**
PathCLR	**56.06 ± 3.47**	**69.82 ± 0.39**	**59.49 ± 3.29**	**75.25 ± 3.48**
Supervised	54.32 ± 2.2	69.4 ± 0.32	57.4 ± 4.14	70.59 ± 0.81
Histotyping	53 ± 0.01	68.3 ± 0.02	54 ± 0.02	70.00 ± 1.00

patient care [24]. Nonetheless, there are many challenges in applying ML to pathology tasks, including the limited availability of labeled and annotated tissue samples.

Eksi et al. [25] conducted a study exploring projects that applied ML techniques to only clinicopathological parameters, to develop models for predicting BCR. Their findings revealed that all ML models outperformed the traditional statistical regression method. The use of ML methods demonstrated the potential for more accurate risk classification, improved prognosis estimation, earlier intervention, reduced unnecessary treatments, and lower morbidity and mortality. However, it is important to note that their study focused exclusively on clinicopathological features and did not integrate tissue microarray histopathology images into their analysis.

In previous studies, ML techniques have been employed to automate Gleason grading of H&E-stained tissue images [26–28]. Another study by Yan et al. [29] utilized contrastive learning on tissue images for cancer diagnosis. However, the application of ML to predict BCR has been relatively scarce.

Kumar et al. [20] attempted BCR prediction using two separate convolutional neural networks (CNNs) on a small number of cases from the CPCTR dataset, evaluating their model on only 30 cases. In contrast, our study utilizes 10-fold cross-validation on the entire CPCTR dataset, providing a more comprehensive evaluation. Additionally, Kumar et al.'s approach required manual annotation of tumors to predict BCR for a test patient, and their prediction was based solely on tissue images. In contrast, our approach does not require additional annotations and benefits from the integration of both clinicopathological features and learned latent representations from tissue images.

Manual annotation of histopathology images is a labor-intensive process, often requiring expert pathologists or even a group of pathologists to vote and annotate a tissue image. To overcome these challenges, we proposed a semi-supervised method that first learns latent representations of unlabeled tissue images and subsequently utilizes each patient's outcome as the label to predict BCR.

Yamamoto et al. [30] employed H&E-stained tissue images to predict BCR and generated low dimensional features using an autoencoder [31] in two different resolutions. Then, they used a support vector machine (SVM) [32] and classical regression [33, 34] to predict BCR; without incorporating clinicopathological features. In contrast, the PathCLR pipeline first generates latent representations then a neural network learns the non-linear relations between a combination of the latent representation of a TMA image and clinicopathological features. This allows the PathCLR model to leverage all available information, both from tissue images and clinicopathological data, to predict BCR for a specific patient. Moreover, we explored using lower resolutions, or the combination of different resolutions but observed no enhancement in accuracy.

A recent study by Leo et al. [16] employs deep learning for the segmentation of lumen glands and subsequently extracts handcrafted features. They implement feature selection and pinpoint 6 features for BCR risk stratification. While our approach is distinct in its focus on predicting BCR for individual patients, we took inspiration from their methodology to establish a baseline model. This served as a benchmark against which we contrasted the performance of PathCLR, with the results detailed in Section 2.4.

The most recent work by Pinckaers et al. [11] utilized deep learning to develop a biomarker for BCR prediction using H&E-stained tissue images. They reported the odds ratio and hazard ratio for the identified group of patients using the developed biomarker. Conversely, our methodology in this investigation steers towards generating individualized BCR predictions, supplying a precise prediction for each individual patient, as opposed to the common task of identifying group biomarkers. A fundamental issue with hazard ratio predictions lies in its nature as a ratio - meaning it can be used to compare one patient to another – e.g., predict that patient A will live longer than patient B – but it does not provide a direct prediction about an individual patient. In contrast, PathCLR predicts a Yes/No prediction for the patient. This enables clinicians to devise more targeted treatment and follow-up strategies that are aligned with each patient's unique circumstances.

Lastly, it is worth mentioning that even though we had access to two datasets with prostate cancer slides, we decided to not train on one dataset and test on the other, as there are significant differences between these two datasets: large variations in the range of clinical feature values and slide resolutions, as well as covariate shift issues, including differences in BCR rates, stage distribution, etc. These differences mean that we do not anticipate that a model trained on one would apply to the other.

2.6 Conclusion

We introduced PathCLR, a semi-supervised machine learning approach designed to predict prostate cancer biochemical recurrence (BCR) within five years following radical prostatectomy surgery. PathCLR utilizes H&E-stained TMA cores and clinicopathological variables for BCR prediction. Notably, PathCLR stands out as the first approach that predicts BCR without the need for time-consuming manual nuclear or tumor annotations. Instead, it learns contrastive latent representations of H&E-stained TMA cores, contributing to its predictive capabilities.

Our results underscore that the integration of latent representations derived from TMA core images and clinicopathological attributes in PathCLR yields statistically significant improvements of at least 5% compared to using clinicopathological features alone for BCR prediction. This highlights the critical information present in diagnostic TMA images, which complements the data from clinicopathological variables and aids in predicting a patient's BCR within five years after surgery. This work represents the initial step towards developing an accurate personalized BCR predictor and sets the stage for future research leading to a deployed system for clinical applications.

Chapter 3

Effective Survival Prediction for Cancer Patient

3.1 Introduction

Cancer is the leading cause of death in the world, with approximately 10 million deaths in 2020 [2]. The most commonly diagnosed cancers in humans include the cancers of breast, lung, colorectal, prostate, and stomach. Accurately predicting the prognosis of cancer can help inform personalized treatment strategies, lead to better patient outcomes, reduced side effects, and more efficient allocation of healthcare resources [35].

This research primarily concentrates on prostate cancer survival analysis and compares it with a recent study by Lee et al. [36]. We then extend our analysis to other common solid tumor types [37]: brain, breast, kidney and renal pelvis, liver, lung and bronchus, stomach, thyroid, and urinary bladder.

Our research focuses on predicting individualized survival distribution (ISD) [38] for a given cancer patient. An ISD is a survival curve that provides personalized survival probabilities at all future time points $t > 0$, based on attributes of patient $\boldsymbol{x}_i$ (see Appendix A for survival dataset definition), defined as $S(\, t \mid \boldsymbol{x}_i \,) = P(T > t \mid \mathbf{X} = \boldsymbol{x}_i)$, to represent the probability that $\boldsymbol{x}_i$ will survive until (at least) time t. When considering all $t > 0$, this $S(\, t \mid \boldsymbol{x}_i \,)$ produces a *survival curve* that starts with a probability of 1.0 at time zero and gradually decreases over time. Figure 3.1 presents ISD curves for three patients, and Appendix B offers further details about ISD. We will use the median time – the time point when the survival curve crosses the probability of 0.5 (see the orange dotted line in Figure 3.1) – to estimate when the event will occur for the patient.

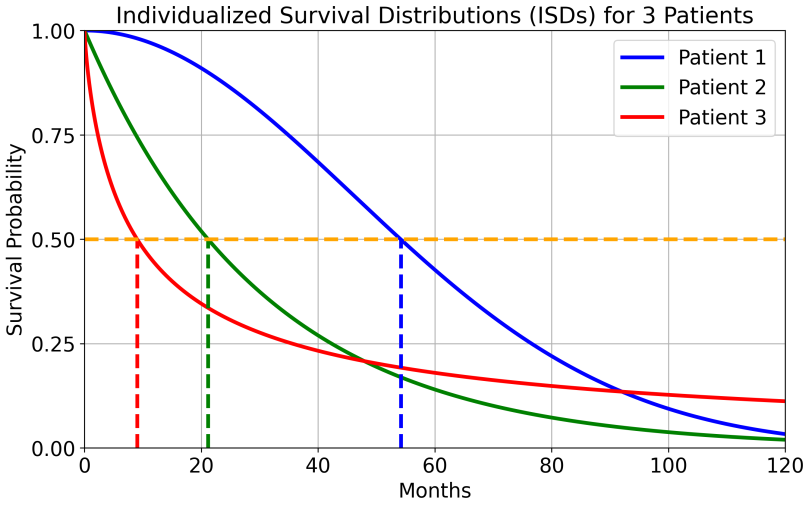

Figure 3.1: ISD curves for three patients. The horizontal orange dotted line shows the probability of 0.5, allowing us to read that the median time (where the survival curve crosses the orange line) for patients 1 to 3 is 54, 21, and 9 months, respectively.

An ISD provides a thorough description of a patient's cancer trajectory for three primary reasons. First, it is tailored to an *individual*, as opposed to generalized group predictions, such as the Kaplan Meier (KM) curve [39]. Second, it provides personalized probabilities of survival over *all future times*, which provides more information and insights to form clinical decisions, compared to models that deal with only a single time point, such as the 5-year and 10-year focus [40], or those that cover only specific time intervals [41]. Lastly, the ISD curve provides the survival *probabilities* rather than risk scores, that alone offer little insight. For example, a model that assigns patient A a risk score of 6, and patient B a risk of 5 is only predicting that A will die before B. This does not specify when A or B will die, or the probability of A surviving beyond one year, etc. While most previous works focused on such risk stratification [42–46], we believe that the ISD curve offers a detailed, meaningful profile, from which one could derive (1) a risk score (by using the negative of the median time), (2) a single-time probability for any time t (by using $S(\, t \mid \boldsymbol{x}_i \,)$), and (3) a regression score that predicts survival timing (by using the median time).

Many studies, including those focusing on ISD models [36, 47, 48], use the C-index [49] as the primary evaluation measure, which only evaluates how well the model is predicting the relative *ordering* of the time to death. We argue that the selected evaluation metrics should align with the specific questions the study aims to answer, a criterion frequently not met by the C-index metric. In this study, we seek to learn ISD models with the goal of accurately predicting the expected survival time for cancer patients. Section 3.2 describes the datasets that we use. Section 3.3 further explains our objective, and proposes evaluation metrics that match our objective, and explains why other commonly used metrics are not relevant for this task. Section 3.4 employs a variety of ML models to compute a patient's ISD curve, and conducts an extensive analysis using various metrics. Section 3.5 further discusses different scenarios, identifies which objective corresponds to which evaluation metric.

3.2 Datasets

Our study utilizes the public Surveillance, Epidemiology, and End Results (SEER) Program dataset[1] [50, 51], which consists of information about cancer patients, and their diagnoses and survival time, reported from registries that cover about 49% of the United States population.

For prostate cancer, we consider two SEER datasets. Our "prostate #1" dataset matches the data (with the same data selection and pre-processing) that was used in a recent work that predicts the ISD curve using the Survival Quilts model [36]. We also used an updated version of the SEER dataset and expanded our scope to encompass different cancer types, including common solid tumors – prostate (denoted as prostate #2), brain, breast, kidney and renal pelvis, liver, lung and bronchus,

[1] https://seer.cancer.gov

Table 3.1: Statistics of patients included in each cancer type, including the number of cases (#Patinets), number of reported deaths (#Death), average patient (Mean age), number of female patients (#Female), number of male patients (#Male), and the number of features (#Feat.).

Cancer	#Patient	#Death (%)	Mean age	#Female (%)	#Male (%)	#Feat.
Prostate #1	171,942	4,157 (2.4%)	65.60	0 (0%)	171,942 (100%)	11
Brain	78,006	47,220 (60.5%)	49.24	34,370 (44.1%)	43,636 (55.9%)	10
Breast	899,341	150,956 (16.7%)	60.80	899,341(100%)	0 (0%)	14
Kidney and renal pelvis	218,373	56,126 (25.7%)	62.77	82,577 (37.8%)	135,796 (62.2%)	15
Liver	93,190	59,568 (63.9%)	63.42	26,204 (28.1%)	66,986 (71.9%)	14
Lung and bronchus	607,701	382,099 (62.8%)	67.87	295,870 (48.7%)	311,831 (51.3%)	15
Prostate #2	745,342	81,619 (10.9%)	66.78	0 (0%)	745,342 (100%)	20
Stomach	108,440	62,418 (57.5%)	66.55	43,076 (39.7%)	65,364 (60.3%)	14
Thyroid	156,023	7,340 (4.7%)	50.30	111,795 (71.6%)	44,228 (28.4%)	13
Urinary bladder	247,505	65,089 (26.2%)	70.00	67,219 (27.2%)	180,286 (72.8%)	14

stomach, thyroid, and urinary bladder. For pre-processing each dataset, we discarded any feature with over 70% missing or unknown values. Additionally, we excluded patients lacking the final time (death or censoring). Finally, we removed patients with matching clinical features and outcomes to avoid duplication (leaving only one instance for each patient). Table 3.1 shows the statistics for each cancer type after this pre-processing. Appendix C presents the features included for each cancer type. Figure 3.2 shows the KM curves [39] of these SEER datasets.

In the realm of survival analysis, individuals who have not experienced the event of interest are considered "censored", perhaps because the study ended before the patient underwent the event, or the patient left the study (perhaps by relocating to a different city), meaning we do not know what happened after the time when they left the study. Consequently, we only have a lower bound of the exact time-to-event duration. For example, if a patient is censored at 10 years, it is clear that they survived at least 10 years, but beyond 10 years, we do not know if they lived for 10 years and a day or 30 years. It is important not to simply remove these censored instances for

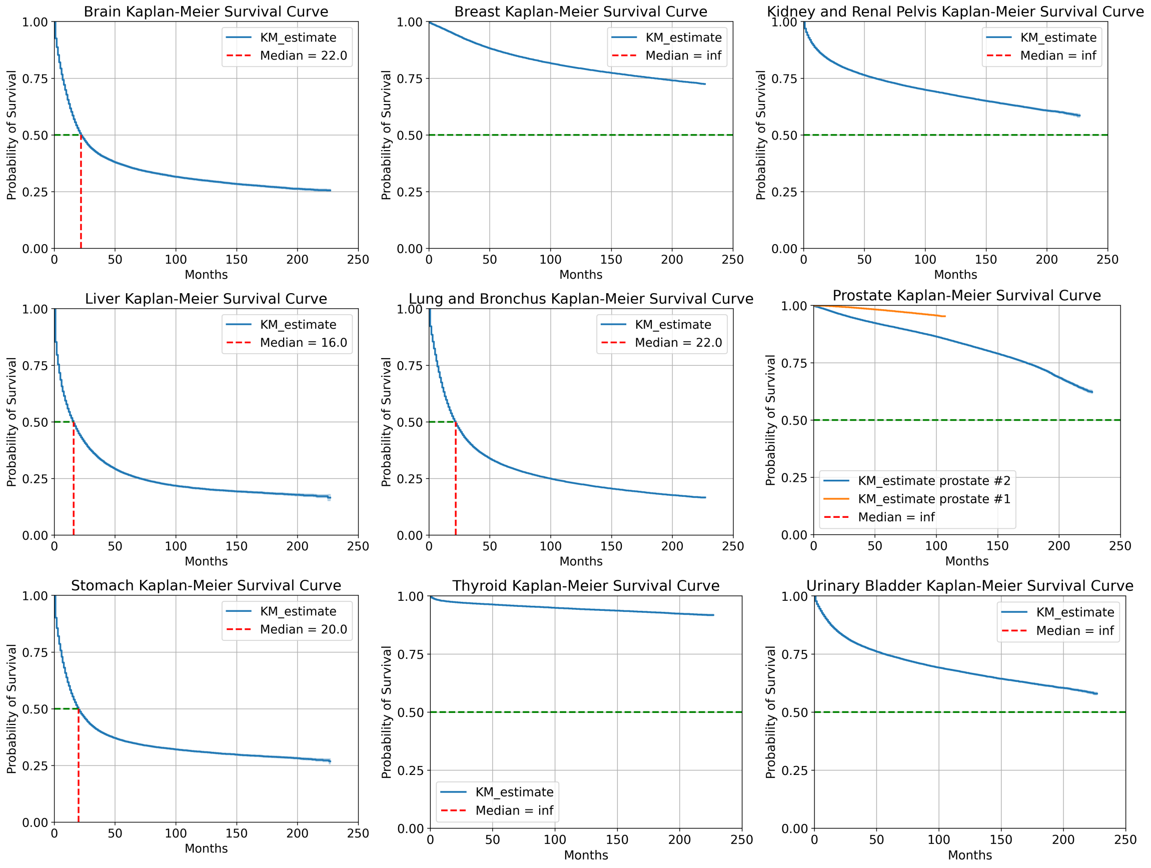

Figure 3.2: Kaplan-Meier survival curve for different types of cancer. The red line indicates the median time for the whole population, and the blue shadow (visible under high magnification) indicates the 95% confidence interval.

two primary reasons: (1) there is valuable information in this observed lower bound of survival, and (2) often a high proportion of the patients are censored. For instance, in the prostate #1 dataset, there is a 97.6% censoring rate, with only 4,157 deaths observed among 171,942 patients. Therefore, a method that effectively incorporates the lower-bound information is essential. The following section will explore ways to address the challenges raised by censorship.

3.3 Methods and Evaluation Metrics

The objective of a supervised learning task is to train a model that can perform a specific task *effectively*. To achieve this objective, it is critical to define what it means for this learned model to be effective, which requires identifying the precise question (or application) the learned model is designed to answer. Once we have a clear understanding of the objective, then we can identify the appropriate metric to evaluate such a model.

It has been reported that cancer patients often seek information regarding their expected lifespan [52], which also proves useful for oncologists in planning personalized treatments, conducting risk-benefit analyses, making end-of-life care decisions, among other aspects [53, 54]. For instance, an oncologist wants to decide on treatment regarding a patient's expected lifespan, in this case, it is useful to precisely predict the time until death for each patient. Here, predicting whether a patient has only a month, four months, or two years to live significantly influences the choice of treatment and end-of-life care strategies, this is only achievable by precise expected lifespan prediction. Hence, we set our objective to use a patient's features to address this question: "*How long should patient A expect to live?*".

To predict the expected lifespan of a given cancer patient, we trained various ISD-based survival prediction models, from traditional statistical tools to cutting-edge deep learning-based methods including: Cox Proportional Hazard (Cox-PH) [55], Accelerate Failure time (AFT) [56], Random Survival Forest (RSF) [57], Multi-Task Logistic Regression (MTLR) [58, 59], Deep MTLR [60], and DeepHit [61]. For all these models we use the median of the learned ISD models as the predicted time to death for the i-th patinet ($\hat{t}_i = S^{-1}(0.5|x_i)$). Appendix F includes a short description of each model and the details of implementation and hyperparameters. Note that we also included the KM curve [39] as a baseline for comparison, even though it is a *population-wide curve* and not an ISD predictor.

To evaluate the mentioned survival prediction models, it is vital to choose an appropriate metric that matches our "How long?" objective. Here, an intuitive metric is the mean absolute error (MAE) – the average absolute difference between the actual time (t_i) and the predicted time ($\hat{t}_i$) – since it exactly measures the error that we aim to minimize [62]. This is challenging to compute since the actual time of the event for censored subjects is unknown. Thus, it becomes essential to utilize alternative forms of MAE suitable for such datasets. Qi et al. [62] explored various MAE forms to measure the difference between the predicted time, and the actual true time, which is unknown for censored patients. They proposed and recommended MAE Pseudo-Observation (MAE-PO), which deals with censored instances by estimating its best guess time, since the actual event time remains unknown due to censorship. They showed that MAE-PO can accurately rank learned models, and often closely match the true MAE. Inspired by their findings, and given our study's goal of precisely predicting survival durations for various cancer types, we employed the MAE-PO metric in our analysis.

We choose to cap the model's prediction and actual time values, at the study's duration, denoted as τ, since predictions beyond this time point lack reliability. Appendix D further motivates this truncated adaptation. The truncated prediction time given by the ISD curve is defined as the minimum of the predicted time and τ (see Equation B.2 in Appendix B for details and the formula) – *i.e.*, truncated median time (Equation B.2). For instance, in Prostate #1, we set τ to 10 years, meaning the truncated prediction time to event is the minimum of 10 years and the median time of the ISD.

In addition, when we cap the prediction time to be less than τ, then we need to ensure that the event time we use is similarly restricted. For all uncensored patients, this value is less than τ as the event occurs within the study period. For censored patients, we first compute MAE-PO estimated best guess, but this might surpass the study's duration τ. To prevent excessively high and inaccurate error values, we also constrain this estimate. Hence, the truncated variation of MAE-PO, called T-MAE-PO, bounds the best guess and predicted survival time to be less than τ. Appendix E includes the definitions and mathematical details for T-MAE-PO.

Additionally, we propose that the survival time should be understood logarithmically. As time progresses, the precision of the prediction becomes less critical. Take, for instance, a situation where the actual event time for a patient is 100 months, and the model predicts 110 months – MAE computes a difference of 10 months. Now consider another patient where the true event time is 4 months and the model's prediction is 14 months. Although the deviation is 10 months in both examples, the second one should attract a stiffer penalty because the gap between 4 months, and 14 months is more critical.

Given this pattern, we advocate for the truncated-log variation as the most fitting metric for our goal. Continuing the previous example, the difference of $log(110)$ and $log(100)$ is equal to $log(110/100) = 0.09$, while the difference of $log(14)$ and $log(4)$ is $log(14/4) = 1.25$. Hence, by applying the logarithm to both predicted and actual times, we give more penalty to a specified difference in short-term predictions compared to long-term prediction errors. We denote this truncated log variation of MAE-PO as TL-MAE-PO (see Appendix E for the formula and details).

While TL-MAE-PO is our chosen evaluation metric that matches our objective, we also aim to compare our results with Lee et al. [36] who evaluated its models on the prostate #1 dataset using C-index and Brier Score. C-index, detailed in Appendix E, is a popular metric to evaluate the performance of survival prediction models in terms of accurately *ranking* patients in applications where relative order matters (see Scenario 2 in Section 3.5). However, the C-index is not relevant to our objective, when the exact timing of the event, rather than ranking, is of primary importance. Lee et al. [36] also used the Brier score, which is the mean squared difference between predicted probabilities and the actual outcomes at a specific time point (here, they use 10 years) [63] (see Appendix E). While the Brier score estimates the mean squared error at a specific time point, but it is not clear why they use 10 years as a target time for evaluation. Moreover, it again fails to assess our objective concerning the expected survival time.

3.4 Results

As we mentioned, the previous work on the prostate #1 cancer dataset learned Survival Quilts models [36], evaluated using C-index and Brier Score. Accordingly, we first evaluated our models on the prostate #1 dataset using the same metrics.

We conducted 10-fold cross-validation and provided the results within a 95% confidence interval (CI) [64]. Figure 3.3 shows that the RSF, Deep-MTLR, Cox-PH, and AFT models outperform Survival Quilts in terms of C-index. Additionally, among all evaluated methods, Deep-MTLR is the top performer with a C-index of 86.05 ± 0.71. In terms of the Brier Score, all of the models had an equally low value of 0.03 ± 0.00 at 10 years. Note that the Brier Score of the degenerate "no one dies in 10 years" model is a low value of 0.05 since only 2.4% of the population died before 10 years.

Note that the Survival Quilts approach combines the Cox-PH, RSF, conditional inference survival forest, and DeepHit models, making it computationally demanding. Yet, each of the individual RSF, Deep-MTLR, Cox-PH, and AFT models demonstrates superior performance and is more time-efficient in training. We suspect that incorporating the DeepHit model in Survival Quilts could be a reason for the lower performance, especially as Figure 3.3 shows that DeepHit significantly underperforms (two-sided t-test, p-value < 0.05) compared to all other ISD models.

Given the issues raised in the prior section, we extend our models' evaluations to include T-MAE-PO and TL-MAE-PO. Table 3.2 shows that AFT performs the best with a TL-MAE-PO of 0.62 ± 0.0015, closely followed by MTLR with a TL-MAE-PO of 0.62 ± 0.0266 (statistically, these results are considered tied). Due to the high computational demands of the Survival Quilts model, it was not feasible to retrain it. Consequently, we are unable to present the outcomes with respect to TL-MAE-PO and include only the originally reported results from the study.

Additionally, we trained our set of ISD predictor models on the other types of

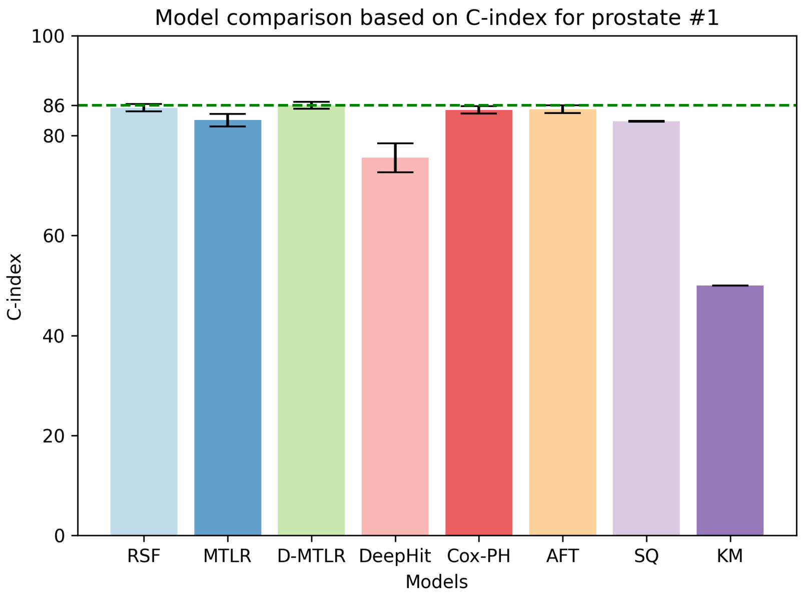

Figure 3.3: Model comparison based on C-index in predicting 10-year prostate cancer-specific mortality (for Prostate # 1 dataset), using a 10-fold cross-validation with 95% CI. SQ is Survival Quilt, and D-MTLR is Deep MTLR.

Table 3.2: Model comparison using the **TL-MAE-PO** metric for all types of cancer, based on monthly survival data. The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate superior performance compared to other models.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
Prostate #1	0.64 ± 0.00	0.62 ± 0.02	0.63 ± 0.02	0.65 ± 0.01	0.63 ± 0.00	**0.62 ± 0.00**	0.67 ± 0.00
Brain	**1.42 ± 0.01**	1.52 ± 0.02	1.44 ± 0.01	1.70 ± 0.00	1.48 ± 0.01	1.49 ± 0.01	2.75 ± 0.00
Breast	**0.93 ± 0.01**	0.97 ± 0.03	0.94 ± 0.02	1.14 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.22 ± 0.00
Kidney	**1.20 ± 0.02**	1.43 ± 0.04	1.21 ± 0.03	1.71 ± 0.03	1.45 ± 0.02	1.46 ± 0.02	2.00 ± 0.02
Liver	**1.93 ± 0.05**	2.07 ± 0.03	1.95 ± 0.04	2.24 ± 0.01	2.07 ± 0.05	2.11 ± 0.05	3.62 ± 0.11
Lung	**1.44 ± 0.00**	1.57 ± 0.02	1.47 ± 0.01	1.66 ± 0.01	1.55 ± 0.01	1.55 ± 0.01	2.81 ± 0.01
Prostate #2	**0.81 ± 0.01**	1.11 ± 0.17	0.89 ± 0.04	1.06 ± 0.02	0.88 ± 0.01	0.89 ± 0.00	1.14 ± 0.01
Stomach	**1.53 ± 0.00**	1.72 ± 0.02	1.61 ± 0.01	1.85 ± 0.02	1.74 ± 0.01	1.78 ± 0.01	3.03 ± 0.02
Thyroid	**1.29 ± 0.01**	1.43 ± 0.06	1.33 ± 0.06	1.66 ± 0.03	1.52 ± 0.01	1.55 ± 0.01	1.81 ± 0.00
Urinary bladder	1.23 ± 0.01	1.44 ± 0.11	**1.22 ± 0.04**	1.68 ± 0.02	1.46 ± 0.01	1.46 ± 0.01	1.82 ± 0.02

cancer. Table 3.2 shows that RSF has the best (10-fold cross validation) TL-MAE-PO loss across many cancer types, followed by Deep-MTLR which is a close second in performance (except for urinary bladder cancer, where it has the best TL-MAE-PO score). Recognizing the common usage of the C-index as a metric in academic studies,

Table 3.3: Model comparison using the **C-index** for all types of cancer, based on monthly survival data. The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate superior performance compared to other models.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
Prostate #1	85.59 ± 0.71	83.11 ± 1.27	**86.05 ± 0.71**	75.57 ± 2.86	85.14 ± 0.74	85.31 ±0.75	50.00 ± 0.00
Brain	**74.87 ± 0.36**	72.46 ± 0.39	74.68 ± 0.39	73.45 ± 0.28	72.63 ± 0.41	72.64 ± 0.43	50.00 ± 0.00
Breast	**82.66 ± 0.08**	77.23 ± 0.53	80.71 ± 0.15	80.76 ± 0.22	76.86 ± 0.09	77.03 ± 0.09	50.00 ± 0.00
Kidney	**86.26 ± 0.06**	79.65 ± 0.24	86.22 ± 0.11	84.59 ± 0.07	79.94 ± 0.14	80.03 ± 0.14	50.00 ± 0.00
Liver	74.50 ± 0.21	70.30 ± 0.26	**74.56 ± 0.18**	73.41 ± 0.22	70.35 ± 0.25	70.30 ± 0.25	50.00 ± 0.00
Lung	**73.02 ± 0.07**	68.88 ± 0.23	72.20 ± 0.11	71.72 ± 0.12	68.96 ± 0.09	69.06 ± 0.09	50.00 ± 0.00
Prostate #2	**88.31 ± 0.10**	75.39 ± 5.93	84.50 ± 0.40	85.94 ± 1.10	83.60 ± 0.08	84.15 ± 0.09	50.00 ± 0.00
Stomach	**77.00 ± 0.14**	72.76 ± 0.33	76.68 ± 0.13	75.25 ± 0.15	71.80 ± 0.21	71.87 ± 0.20	50.00 ± 0.00
Thyroid	93.22 ± 0.22	89.38 ± 0.34	**93.30 ± 0.34**	92.22 ± 0.31	90.27 ± 0.20	90.20 ± 0.20	50.00 ± 0.00
Urinary bladder	**81.73 ± 0.18**	74.41 ± 0.68	81.34 ± 0.27	80.94 ± 0.20	74.64 ± 0.19	74.73 ± 0.18	50.00 ± 0.00

we also compared the models using the C-index in Table 3.3, which shows that both RSF and D-MTLR appear as top performers in terms of C-index, and following that, DeepHit achieves strong results. Appendix G presents the results on these datasets, using several other metrics.

Additionally, in certain cancer types, there is a noticeable variation in model rankings when comparing the T-MAE-PO with the TL-MAE-PO metric. For instance, with kidney and renal pelvis, Deep-MTLR is the best based on T-MAE-PO, whereas RSF is the best with TL-MAE-PO. Hence, if our specific application is more concerned about short-term accuracy over long-term precision, it would be advisable to utilize the log variations of the MAE for evaluation.

Consequently, the ranking of top models changes based on the chosen metric, highlighting the importance of appropriate metric selection. This insight emphasizes the need to first determine the primary objective of a study involving survival prediction models and subsequently select a proper metric for performance evaluation. This becomes particularly important when we are considering multiple ISD models and want to select the optimal one based on a specific evaluation criterion.

3.5 Discussion and Conclusion

This research conducted a comprehensive study on various SEER datasets, to determine which learned model can best predict time to mortality subsequent to cancer diagnosis. While many past studies have focused on survival prediction for a single type of cancer at a particular time point as a classification task [65, 66], our research stands out by addressing a more challenging and general survival prediction task, specifically focusing on ISD models (which also allows us to address other tasks) and effective evaluation.

Additionally, our emphasis on rigorous evaluation sets us apart from earlier studies. There are different scenarios in which we can use survival prediction, and the selection of evaluation metrics must resonate with our primary objectives.

Scenario 1: In a foundational study, one might be interested in binary classification and predicting the probability of survival at a specific time point. For instance, a hospital aims to decide about providing end-of-life care for those in need, and therefore, needs to predict 1-year survival outcome and categorize each patient as either surviving beyond one year versus those who experience the event within the first year. Hence, we need a model that can predict "*What is the probability that patient A will live more than 1 year?*". In this case of binary classification, metrics such as accuracy, F1 score, Area Under the Curve (AUC), 1-Calibration, etc., have been used in previous studies [67–70].

Scenario 2: In a given scenario such as liver transplant prioritization, we need to decide which patient should receive the liver transplant. Here, the objective is to discern which patient will die soonest after diagnosis, as this identifies which patient should receive the transplant. This situation emphasizes the comparison between all pairs of patients, focusing on accurately determining the order in which each patient will experience the event, and answering "*Among the current set of patients, who will experience the event first?*". In such a context, the exact prediction of the time to the event is not the priority; instead, a correct ordering of the patients in terms of survival risk or mortality time suffices. The proper evaluation metric in this scenario is the C-index metric, specifically designed for measuring the correctness of ranking [71–73].

Scenario 3: In a different context, one might aim to decide about allocating hospital beds. Here, it is useful to predict the duration of a patient's hospital stay or time until death. In this context, we want to answer a question concerning "*How long?*". This situation (including our research objective) primarily concerns the time until the event, making the MAE-PO a proper evaluation metric [74, 75]. In some cases of this scenario, predictions are targeted up to a specific time, such as 10 years, without considering or trusting further forecasts. Here, T-MAE-PO is advisable for evaluation. Finally, for predicting " *How long?*" up to a set time with a focus on the accurate prediction of earlier events, we recommend TL-MAE-PO.

Of course, if our model attains a perfect MAE score (essentially zero), we could then address the queries posed in scenarios 1 and 2 using precise time-to-event predictions. Nonetheless, achieving an MAE of close to zero in such predictions is notably a more challenging task than, for example, addressing a 1-year classification task, where the focus is on correct prediction at a specific time rather than all future time points. Hence, if the objective aligns with scenarios 1 and 2, the basic evaluation measures highlighted for each such case could be adequate, eliminating unnecessary complexity.

In conclusion, it is important to first determine the study objective prior to selecting an appropriate evaluation metric for survival prediction problems. In specific circumstances of the third scenario, we recommend utilizing the proposed T-MAE-PO or TL-MAE-PO instead of other metrics, for analyzing survival data. Lastly, our research shows that RSF, followed by Deep MTLR, ranks as the top learning model across various types of cancer based on both TL-MAE-PO and C-index.

Chapter 4

Conclusions

This thesis encompasses two pivotal studies in the realm of cancer-related research and machine learning. The first study proposes PathCLR, a novel semi-supervised contrastive learning approach designed for predicting BCR within five years following RP surgery. PathCLR effectively leverages H&E-stained TMA cores and clinicopathological variables, without a need for labor-intensive manual annotations. We showed that PathCLR performs better than previous works – suggesting that there is critical information in diagnostic TMA images, which complements the data from clinicopathological variables and aids in predicting a patient's BCR within five years after surgery. Future research can build upon this foundation to develop a more precise and efficient model, perhaps by improving the representation learner part of this pipeline and using other more recent models such as transformers [76–78]. Further, our current approach is not robust to the source-site variation. This should be further investigated to train a robust model that is transferable to other datasets from different hospitals. Additionally, this work can be extended to predict the probability of recurrence over all future time points, rather than limiting the prediction to a single time point (5 years).

The second study addresses several of these limitations and conducts a comprehensive analysis using the SEER dataset across various cancer types and focuses on a more general task where the aim is to predict the expected lifespan of a specific patient using ISD models. It emphasizes the crucial role of selecting the right evaluation metrics that align with the study's objectives and suggests T-MAE-PO and TL-MAE-PO metrics to determine whether a model achieves the objective, which here is to predict the expected lifespan. Future directions for this study could involve broadening the dataset variety, incorporating data from sources other than the SEER database, or considering other cancer types. Additionally, there is potential in merging genetic and molecular information, pathology slides (utilizing the latent representations learned from PathCLR), or various other data forms beyond clinical information, to determine whether these enhancements could improve expected survival time prediction.

In summary, it is our hope that this thesis makes a valuable contribution to the fields of oncology and machine learning by introducing novel methods and perspectives for predicting cancer recurrence and expected survival time, thereby laying a robust groundwork for future research and clinical applications.

Bibliography

[1] M. Farrokh, N. Kumar, P. H. Gann, and R. Greiner, "Learning to predict prostate cancer recurrence from tissue images," *Journal of Pathology Informatics*, p. 100 344, 2023.

[2] H. Sung *et al.*, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[3] C. H. Pernar, E. M. Ebot, K. M. Wilson, and L. A. Mucci, "The epidemiology of prostate cancer," *Cold Spring Harbor perspectives in medicine*, vol. 8, no. 12, a030361, 2018.

[4] H. G. Welch and P. C. Albertsen, "Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986–2005," *Journal of the National Cancer Institute*, vol. 101, no. 19, pp. 1325–1329, 2009.

[5] R. Tourinho-Barbosa *et al.*, "Biochemical recurrence after radical prostatectomy: What does it mean?" *International braz j urol*, vol. 44, pp. 14–21, 2018.

[6] P. J. Boström *et al.*, "Genomic predictors of outcome in prostate cancer," *European urology*, vol. 68, no. 6, pp. 1033–1044, 2015.

[7] M. R. Cooperberg, J. F. Hilton, and P. R. Carroll, "The capra-s score: A straightforward tool for improved prediction of outcomes after radical prostatectomy," *Cancer*, vol. 117, no. 22, pp. 5039–5046, 2011.

[8] S. Punnen *et al.*, "Multi-institutional validation of the capra-s score to predict disease recurrence and mortality after radical prostatectomy," *European urology*, vol. 65, no. 6, pp. 1171–1177, 2014.

[9] A. E. Ross, A. D'amico, and S. Freedland, "Which, when and why? rational use of tissue-based molecular testing in localized prostate cancer," *Prostate Cancer and Prostatic Diseases*, vol. 19, no. 1, pp. 1–6, 2016.

[10] A. E. Ross *et al.*, "Tissue-based genomics augments post-prostatectomy risk stratification in a natural history cohort of intermediate-and high-risk men," *European urology*, vol. 69, no. 1, pp. 157–165, 2016.

[11] H. Pinckaers *et al.*, "Predicting biochemical recurrence of prostate cancer with artificial intelligence," *Communications Medicine*, vol. 2, no. 1, pp. 1–9, 2022.

[12] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," *arXiv preprint arXiv:2006.05278*, 2020.

[13] O. Ciga, T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," *Machine Learning with Applications*, vol. 7, p. 100 198, 2022.

[14] S. Azizi *et al.*, "Robust and efficient medical imaging with self-supervision," *arXiv preprint arXiv:2205.09723*, 2022.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[16] P. Leo *et al.*, "Computer extracted gland features from h & e predicts prostate cancer recurrence comparably to a genomic companion diagnostic test: A large multi-site study," *NPJ precision oncology*, vol. 5, no. 1, p. 35, 2021.

[17] A. A. Patel *et al.*, "The development of common data elements for a multi-institute prostate cancer tissue bank: The cooperative prostate cancer tissue resource (cpctr) experience," *BMC cancer*, vol. 5, no. 1, pp. 1–14, 2005.

[18] R. Verma, N. Kumar, A. Patil, N. C. Kurian, S. Rane, and A. Sethi, "Multi-organ nuclei segmentation and classification challenge 2020," *IEEE transactions on medical imaging*, vol. 39, no. 1380-1391, p. 8, 2020.

[19] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020.

[20] N. Kumar *et al.*, "Convolutional neural networks for prostate cancer recurrence prediction," in *Medical Imaging 2017: Digital Pathology*, SPIE, vol. 10140, 2017, pp. 106–117.

[21] S. Graham *et al.*, "Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Medical Image Analysis*, vol. 58, p. 101 563, 2019.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[24] M. Cui and D. Y. Zhang, "Artificial intelligence and computational pathology," *Laboratory Investigation*, vol. 101, no. 4, pp. 412–422, 2021.

[25] M. Ekşi *et al.*, "Machine learning algorithms can more efficiently predict biochemical recurrence after robot-assisted radical prostatectomy," *The Prostate*, vol. 81, no. 12, pp. 913–920, 2021.

[26] K. Nagpal *et al.*, "Development and validation of a deep learning algorithm for gleason grading of prostate cancer from biopsy specimens," *JAMA oncology*, vol. 6, no. 9, pp. 1372–1380, 2020.

[27] A. Hossain *et al.*, "Automated approach for estimation of grade groups for prostate cancer based on histological image feature analysis," *The Prostate*, vol. 80, no. 3, pp. 291–302, 2020.

[28] W. Bulten *et al.*, "Automated deep-learning system for gleason grading of prostate cancer using biopsies: A diagnostic study," *The Lancet Oncology*, vol. 21, no. 2, pp. 233–241, 2020.

[29] J. Yan, H. Chen, X. Li, and J. Yao, "Deep contrastive learning based tissue clustering for annotation-free histopathology image analysis," *Computerized Medical Imaging and Graphics*, vol. 97, p. 102 053, 2022.

[30] Y. Yamamoto *et al.*, "Automated acquisition of explainable knowledge from unannotated histopathology images," *Nature communications*, vol. 10, no. 1, pp. 1–9, 2019.

[31] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[32] V. Vapnik, *The nature of statistical learning theory.* Springer science & business media, 1999.

[33] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[34] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[35] F. C. Hamdy *et al.*, "10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer," *New England Journal of Medicine*, vol. 375, no. 15, pp. 1415–1424, 2016.

[36] C. Lee, A. Light, A. Alaa, D. Thurtle, M. van der Schaar, and V. J. Gnanapragasam, "Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the surveillance, epidemiology, and end results (seer) database," *The Lancet Digital Health*, vol. 3, no. 3, e158–e165, 2021.

[37] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *Ca Cancer J Clin*, vol. 73, no. 1, pp. 17–48, 2023.

[38] H. Haider, B. Hoehn, S. Davis, and R. Greiner, "Effective ways to build and evaluate individual survival distributions," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 3289–3351, 2020.

[39] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[40] M. Mourad *et al.*, "Machine learning and feature selection applied to seer data to reliably assess thyroid cancer prognosis," *Scientific reports*, vol. 10, no. 1, p. 5176, 2020.

[41] H. Shakir, B. Aijaz, T. M. R. Khan, and M. Hussain, "A deep learning-based cancer survival time classifier for small datasets," *Computers in Biology and Medicine*, vol. 160, p. 106 896, 2023.

[42] R. Zelic *et al.*, "Predicting prostate cancer death with different pretreatment risk stratification tools: A head-to-head comparison in a nationwide cohort study," *European urology*, vol. 77, no. 2, pp. 180–188, 2020.

[43] S. P. Basourakos *et al.*, "Tissue-based biomarkers for the risk stratification of men with clinically localized prostate cancer," *Frontiers in Oncology*, vol. 11, p. 676 716, 2021.

[44] M. G. Sanda *et al.*, "Clinically localized prostate cancer: Aua/astro/suo guideline. part i: Risk stratification, shared decision making, and care options," *The Journal of urology*, vol. 199, no. 3, pp. 683–690, 2018.

[45] M. R. Cooperberg, J. M. Broering, and P. R. Carroll, "Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis," *JNCI: Journal of the National Cancer Institute*, vol. 101, no. 12, pp. 878–887, 2009.

[46] R. N. Hansen *et al.*, "Long-term survival trends in patients with unresectable stage iii non-small cell lung cancer receiving chemotherapy and radiation therapy: A seer cancer registry analysis," *BMC cancer*, vol. 20, no. 1, pp. 1–6, 2020.

[47] D. R. Thurtle, D. C. Greenberg, L. S. Lee, H. H. Huang, P. D. Pharoah, and V. J. Gnanapragasam, "Individual prognosis at diagnosis in nonmetastatic prostate cancer: Development and external validation of the predict prostate multivariable model," *PLoS medicine*, vol. 16, no. 3, e1002758, 2019.

[48] C. Nagpal, S. Yadlowsky, N. Rostamzadeh, and K. Heller, "Deep cox mixtures for survival regression," in *Machine Learning for Healthcare Conference*, PMLR, 2021, pp. 674–708.

[49] F. E. Harrell Jr, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in medicine*, vol. 15, no. 4, pp. 361–387, 1996.

[50] N Howlader, A. Noone, M Krapcho, J Garshell, D Miller, S. Altekruse, *et al.*, "Surveillance, epidemiology, and end results (seer) program (www. seer. cancer. gov) seer* stat database: Incidence-seer 18 regs research data+ hurricane katrina impacted louisiana cases, nov 2015 sub (2000-2013)¡ katrina/rita population adjustment¿-linke. 2015," *Rita population adjustment¿ e Linke*, 2015.

[51] N. C. Institute, "Surveillance, epidemiology, and end results (seer) program," *Cancer Statistics, SEER Data & Software, Registry Operations*, 2018.

[52] C Alifrangis *et al.*, "The experiences of cancer patients," *QJM: An International Journal of Medicine*, vol. 104, no. 12, pp. 1075–1081, 2011.

[53] B. H. Osse, M. J. Vernooij-Dassen, E. Schadé, B. de Vree, M. E. van den Muijsenbergh, and R. P. Grol, “Problems to discuss with cancer patients in palliative care: A comprehensive approach,” *Patient education and counseling*, vol. 47, no. 3, pp. 195–204, 2002.

[54] B. Saraiya, S. Bodnar-Deren, E. Leventhal, and H. Leventhal, “End-of-life planning and its relevance for patients’ and oncologists’ decisions in choosing cancer therapy,” *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 113, no. S12, pp. 3540–3547, 2008.

[55] D. R. Cox, “Regression models and life-tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[56] W. Stute, “Consistent estimation under random censorship when covariables are present,” *Journal of Multivariate Analysis*, vol. 45, no. 1, pp. 89–103, 1993.

[57] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, “Random survival forests,” *The Annals of Applied Statistics*, pp. 841–860, 2008.

[58] C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos, “Learning patient-specific cancer survival distributions as a sequence of dependent regressors,” *Advances in Neural Information Processing Systems*, vol. 24, pp. 1845–1853, 2011.

[59] P. Jin, “Using survival prediction techniques to learn consumer-specific reservation price distributions,” M.S. thesis, University of Alberta, 2015.

[60] S. Fotso, “Deep neural networks for survival analysis based on a multi-task framework,” *arXiv preprint arXiv:1801.05512*, 2018.

[61] C. Lee, J. Yoon, and M. 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*, vol. 67, no. 1, pp. 122–133, 2019.

[62] S.-A. Qi *et al.*, “An effective meaningful way to evaluate survival models,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 2023, pp. 28 244–28 276.

[63] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, “Assessment and comparison of prognostic classification schemes for survival data,” *Statistics in medicine*, vol. 18, no. 17-18, pp. 2529–2545, 1999.

[64] M. Stone, “Cross-validatory choice and assessment of statistical predictions,” *Journal of the royal statistical society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[65] A. Safiyari and R. Javidan, “Predicting lung cancer survivability using ensemble learning methods,” in *2017 intelligent systems conference (IntelliSys)*, IEEE, 2017, pp. 684–688.

[66] M. S. I. Polash, S. Hossen, R. K. R. Sarker, M. A. Bhuiyan, and A. Taher, "Functionality testing of machine learning algorithms to anticipate life expectancy of stomach cancer patients," in *2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, IEEE, 2022, pp. 1–6.

[67] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[68] S. Hegselmann, L. Gruelich, J. Varghese, and M. Dugas, "Reproducible survival prediction with seer cancer data," in *Machine Learning for Healthcare Conference*, PMLR, 2018, pp. 49–66.

[69] L. Han *et al.*, "Nomogram of conditional survival probability of long-term survival for metastatic colorectal cancer: A real-world data retrospective cohort study from seer database," *International Journal of Surgery*, vol. 92, p. 106 013, 2021.

[70] W. Wang, J. Liu, and L. Liu, "Development and validation of a prognostic model for predicting overall survival in patients with bladder cancer: A seer-based study," *Frontiers in Oncology*, vol. 11, p. 692 728, 2021.

[71] A. E. Braat *et al.*, "The eurotransplant donor risk index in liver transplantation: Et-dri," *American Journal of Transplantation*, vol. 12, no. 10, pp. 2789–2796, 2012.

[72] J. L. A. van Vugt *et al.*, "A model including sarcopenia surpasses the meld score in predicting waiting list mortality in cirrhotic liver transplant candidates: A competing risk analysis in a national cohort," *Journal of hepatology*, vol. 68, no. 4, pp. 707–714, 2018.

[73] P. Burra *et al.*, "Limitations of current liver donor allocation systems and the impact of newer indications for liver transplantation," *Journal of Hepatology*, vol. 75, S178–S190, 2021.

[74] M. Tello *et al.*, "Machine learning based forecast for the prediction of inpatient bed demand," *BMC medical informatics and decision making*, vol. 22, no. 1, p. 55, 2022.

[75] E. Kutafina, I. Bechtold, K. Kabino, and S. M. Jonas, "Recursive neural networks in hospital bed occupancy forecasting," *BMC medical informatics and decision making*, vol. 19, pp. 1–10, 2019.

[76] K. He *et al.*, "Transformers in medical image analysis," *Intelligent Medicine*, vol. 3, no. 1, pp. 59–78, 2023.

[77] F. Shamshad *et al.*, "Transformers in medical imaging: A survey," *Medical Image Analysis*, p. 102 802, 2023.

[78] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[79] Q. Gong and L. Fang, "Asymptotic properties of mean survival estimate based on the kaplan–meier curve with an extrapolated tail," *Pharmaceutical statistics*, vol. 11, no. 2, pp. 135–140, 2012.

[80] N. R. Latimer, "Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data: Inconsistencies, limitations, and a practical guide," *Medical Decision Making*, vol. 33, no. 6, pp. 743–754, 2013.

[81] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. Voelker, B. Nussenbaum, and E. W. Wang, "A practical guide to understanding kaplan-meier curves," *Otolaryngology—Head and Neck Surgery*, vol. 143, no. 3, pp. 331–336, 2010.

[82] H. Kvamme and Ø. Borgan, "The brier score under administrative censoring: Problems and solutions," *arXiv preprint arXiv:1912.08581*, 2019.

Appendix A: Survival Dataset

In a survival analysis dataset consisting of N patients, the data for each individual is represented as (x_i, t_i, δ_i), where $x_i \in \mathbb{R}^d$ is the feature set for the i-th patient and $t_i \in \mathbb{R}_+$ is the survival time which is either the censored time or event time. $\delta_i \in \{0, 1\}$ indicates if the patient was censored or not, where $\delta_i = 0$ indicates that the i-th patient was censored and t_i is the censoring time, and $\delta_i = 1$ means that the patient experienced the event (death), and t_i is the time to event. Hence, a survival dataset is represented as $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^N$.

Appendix B: Individualized Survival Distribution (ISD)

A patient's ISD curve shows their likelihood of survival as time progresses. This is the probability of survival until time t given the patient's features of x_i and time t, and it is represented as $S(\,t \mid \boldsymbol{x}_i\,) = P(T > t \mid \mathbf{X} = \boldsymbol{x}_i)$. The ISD curve begins with a survival probability of 1 at time zero and gradually declines thereafter. Each ISD is specific to an instance using specific clinical data from that individual patient (x_i), distinguishing them from curves like the KM curve [39], which are derived from an entire population's data.

If one needs a single value, many use time-to-event prediction given by the model's output (ISD curve), either mean (denoted by $\mathbb{E}_t[\,S(\,t \mid \boldsymbol{x}_i\,)\,]$) or median survival time (denoted by $\text{median}(S(\,t \mid \boldsymbol{x}_i\,))$. The truncated adaptations of the mean (expected) and median survival time with respect to time τ are defined as follows:

$$\begin{aligned} \hat{t}_{i,\text{T-mean},\tau} &= \min\{\, \mathbb{E}_t[\,S(\,t \mid \boldsymbol{x}_i\,)\,]\,, \quad \tau\} \\ &= \min\{\, \int_0^\infty S(\,t \mid \boldsymbol{x}_i\,)\ dt,\ \tau\} \end{aligned} \tag{B.1}$$

and

$$\begin{aligned} \hat{t}_{i,\text{T-median},\tau} &= \min\{\, \text{median}(S(\,t \mid \boldsymbol{x}_i\,)),\ \tau\} \\ &= \min\{\, S^{-1}(\,P = 0.5 \mid \boldsymbol{x}_i\,),\ \tau\}, \end{aligned} \tag{B.2}$$

where τ represents the time point that we truncate, $\boldsymbol{x}_i$ denotes the attributes of patient i, $S(\,t \mid \boldsymbol{x}_i\,)$ is the predicted ISD curve for this patient, and S^{-1} is the inverse function of the survival function S. τ can be set to any time point depending on the application, here in this study, we set it to be the final time point (length of the study). In this study, we use the truncated median time (Equation B.2) as the prediction time.

Note that the ISD curve often does not cross the probability of 0.5. In such cases, the common approach for calculating the standard median time is to linearly extrapolate the curve until it reaches the 0.5 probability – we draw a line from the initial time point with a probability of 1 to the final time point, then continue this line until it intersects with the probability of either 0 or 0.5. However, in the case of truncated median time (Equation B.2), extrapolation is not required, as we bound the prediction by τ. For example, in Figure B.1 left, the median of the ISD curve is 22 months, which is less than $\tau = 200$ months, which is the end of the study – here, we set the time to event prediction to the median time (22 months). For Figure B.1 right, the ISD curve ends before reaching the probability of 0.5, and as a result, we know that the median time is after the end of the study. Since we take the minimum of the median time and the end of the study time (200), we set the truncated prediction time to 200 months. Note this means that we do not need to extrapolate the ISD curve.

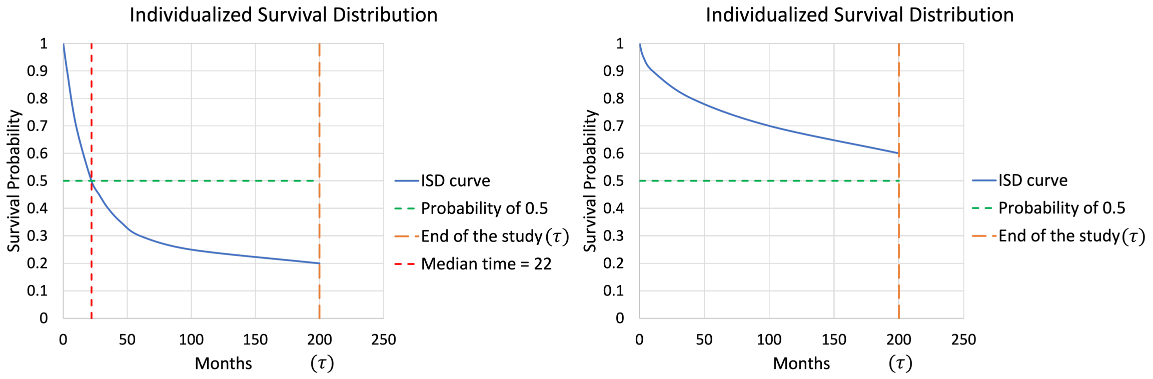

Figure B.1: ISD curve for two patients. The end of the study time (τ) for both curves is 200 months. The truncated time to event prediction using the median time of the left side ISD curve is 22 months and for the right one is 200 months.

Appendix C: SEER Features

Table C.1 lists the features we used for each type of cancer. Note that we chose not to do the feature selection step as: (1) the number of included features was less than 20, and (2) we viewed this as a distraction from the primary objective of our research.

Table C.1: List of features included in each type of cancer dataset. In this table, prostate refers to prostate # 2.

Feature	Brain	Breast	Kidney	Liver	Lung	Prostate	Stomach	Thyroid	Urinary
Age	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sex	✓	-	✓	✓	✓	-	✓	✓	✓
Behavior recode for analysis	✓	-	-	-	✓	-	-	-	-
Combined Summary Stage	✓	✓	✓	✓	✓	✓	✓	✓	✓
Grade	✓	✓	✓	✓	✓	✓	✓	-	✓
RX Summ–Scope Reg LN Sur	✓	✓	✓	✓	✓	✓	✓	✓	✓
RX Summ–Surg Oth Reg/Dis	✓	✓	✓	✓	✓	✓	✓	✓	✓
RX Summ–Surg Prim Site	✓	✓	✓	✓	✓	✓	✓	✓	✓
Summary stage 2000	✓	✓	✓	✓	✓	✓	✓	✓	✓
SEER historic stage A	✓	✓	✓	✓	✓	✓	✓	✓	✓
Derived AJCC T, 6th ed (2004-2015)	-	✓	✓	✓	✓	✓	✓	✓	✓
Derived AJCC N, 6th ed (2004-2015)	-	✓	✓	✓	✓	✓	✓	✓	✓
Derived AJCC M, 6th ed (2004-2015)	-	✓	✓	✓	✓	✓	✓	✓	✓
Breast - Adjusted AJCC 6th Stage	-	✓	-	-	-	-	-	-	-
Derived AJCC Stage Group, 6th ed	-	✓	✓	✓	✓	✓	✓	✓	✓
Derived AJCC Stage Group, 7th ed	-	✓	✓	✓	✓	✓	✓	✓	✓
Invasion Beyond Capsule Recode	-	-	✓	-	-	-	-	-	-
Gleason Patterns Clinical Recode	-	-	-	-	-	✓	-	-	-
Gleason Patterns Pathological Recode	-	-	-	-	-	✓	-	-	-
Gleason Score Clinical Recode	-	-	-	-	-	✓	-	-	-
Gleason Score Pathological Recode	-	-	-	-	-	✓	-	-	-
PSA Lab Value Recode	-	-	-	-	-	✓	-	-	-
Number of Cores Positive Recode	-	-	-	-	-	✓	-	-	-
Number of Cores Examined Recode	-	-	-	-	-	✓	-	-	-

Appendix D: Motivation for the Truncated Variation of MAE

In this study, we proposed the truncated variation of MAE-PO; this section motivates this variation one step further. In Section 3.3, we discussed truncating the predicted time-to-event, which is the median time of the ISD curve, and the same issue is raised in the context of KM curves. When dealing with the KM curve of datasets with high censorship, this curve often fails to descend to zero and might not even cross the 0.5 survival probability threshold. Consequently, the median time, typically employed as a time-to-event prediction, is unknown. Among our included datasets, as illustrated in Figure 3.2, for cancers of breast, kidney and renal pelvis, prostate, thyroid, and urinary bladder, the blue KM curve does not intersect the green line (representing 0.5 probability) by the study's conclusion.

Some prior studies have attempted to address this matter, proposing: (1) dropping the curve vertically to zero post-study conclusion, (2) employing linear extrapolation (which we illustrated in Figure D.1), and (3) applying a specific function or distribution to extend the curve [79, 80]. However, Rich et al. [81] noted that any form of KM curve extrapolation lacks justification, and any prediction after the study conclusion is unreliable. Take the prostate # 1 dataset as an instance, where the survival curve does not reach the probability of 0.5. If we use linear extrapolation – from the starting point of the curve (0,1) to the final time point, then continue the line to reach the probability of 0.5 or 0 – to continue the curve and find the median time, as demonstrated in Figure D.1, then we can see that linear extrapolation exceeds 2100 months (175 years) of survival, and the median time is 1121 months (93 years). Given that the age average for the prostate #1 cancer dataset is 65 years, then a prediction of $65 + 93 = 158$ years is a wrong and unrealistic prediction.

Therefore, we follow the same suggestion as Rich et al [81], meaning that we drop the ISD cure vertically to zero post-study conclusion, bound the predictions of trained models and the best guess estimate for actual time to event by the length of the study (τ), as any prediction beyond the conclusion of the study is unreliable and lacks justification.

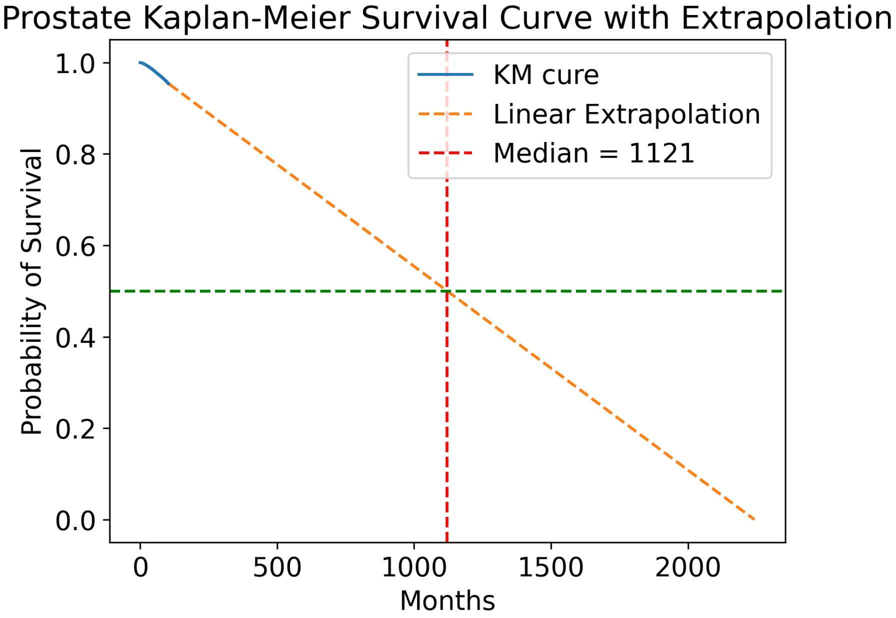

Figure D.1: KM curve linear extrapolation for prostate #1 dataset.

Appendix E: Evaluation Metrics in Details

In this section, we explain the formula of evaluation metrics and describe them in detail.

1. **C-index:**

   The C-index of a model, on a labeled survival dataset, is given by

   $$\text{C-index}(S(.|.), \mathcal{D}) \quad = \quad \frac{\text{Number of concordant pairs}}{\text{Number of comparable pairs}}\,, \tag{E.1}$$

   where a pair of instances is considered concordant if the predicted and the actual outcome follow the same ranking. Among all possible combinations of two subjects from a sample size of N, a comparable pair means we know which one of the subjects experienced the event first. For example, as shown in Figure E.1, patients A and B can be considered a comparable pair because it is clear that the event occurred first with patient A. In contrast, patients B and C do not form a comparable pair since patient B is censored prior to patient C's event, leaving ambiguity about whether patient B experienced the event before or after patient C. Hence, for patients B and C we do not know who experienced the event first, and remains uncertain and incomparable. Additionally, any two patients who are not censored are comparable, making patients A and C a comparable pair. Therefore in Figure E.1, we have 2 comparable pairs: {A, B}, and {A, C}.

After computing the number of comparable pairs, given the model's prediction versus the ground truth, we compute the number of concordant pairs. So following our example, if we predict the following time to events: $A = 5$, $B = 13$, $C = 8$, then we have correctly ranked both of our comparable pairs, since time to event prediction for B is greater than A, and C is also greater than A. Thus, C-index is equal to 1.

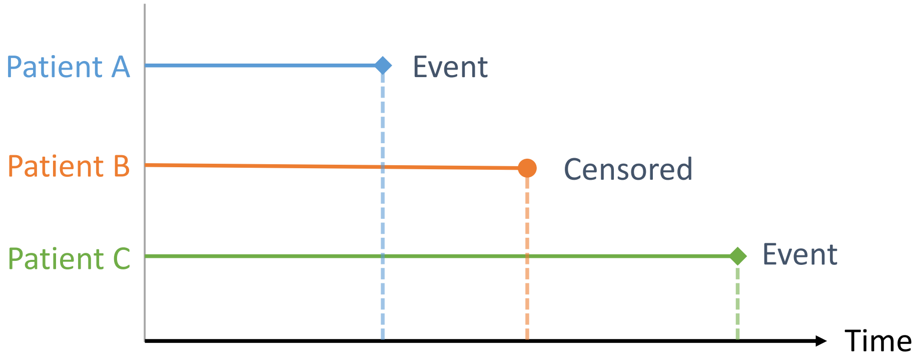

Figure E.1: Time to event/censorship for three patients.

2. **Brier Score:**

Brier Score (BS) is the squared difference between the predicted probability of survival at a specific time t and the true event value (0 or 1) [63]. It ranges between zero to one, and a value of zero means perfect prediction. For censored patients with unknown event values, BS uses the inverse probability censoring weight (IPCW) [82], which uniformly transfers each censored patient's weight to uncensored patients after that time.

BS is defined as:

$$BS(t, \mathcal{D}) \quad = \quad \frac{1}{N}\sum_{i=1}^{N} \frac{(0 - S_{\mathrm{m}}(\, t \mid \boldsymbol{x}_i \,))^2 \cdot 1_{t_i \le t, \delta_i = 1}}{G_i(t_i)} + \frac{(1 - S_{\mathrm{m}}(\, t \mid \boldsymbol{x}_i \,))^2 \cdot 1_{t_i > t}}{G_i(t)} \;,$$

where $G_i(t)$ is the probability of not being censored until time t, which is commonly estimated by running the KM algorithm, but with the censor-bit (event flag) flipped.

3. MAE:

Mean absolute error (MAE) measures the average absolute difference between the predicted time ($\hat{t}_i$) and the actual (truth) time (t_i):

$$\text{MAE}(\{\hat{t}_i\},\ \{t_i\}) \quad = \quad \frac{1}{N}\sum_{i=1}^{N} |\hat{t}_i - t_i| \ . \tag{E.2}$$

For the prediction time ($\hat{t}_i$), we use the median time of the ISD model ($\hat{t}_i = S^{-1}(\ P\ =\ 0.5\ |\ \boldsymbol{x}_i\ ))$. However, to compute this MAE, the actual time (t_i) is unknown for censored patients. Hence, we need to use another variation of MAE that can estimate the truth time for censored patients. In this study, we use the MAE-PO that employs pseudo-observation to estimate the actual time of survival for censored patients [62].

4. MAE-PO:

Qi et al. [62] proposed the MAE-PO that employs pseudo-observation to estimate the actual time of survival for censored patients and uses $\hat{\theta}$ as a predictor, which can be based on the mean value of the KM estimator, $\hat{\theta} = \mathbb{E}_t[S_{\text{KM}(\mathcal{D})}(\ t\ )]$, where $S_{\text{KM}(\mathcal{D})}(\ t\ )$ is the group-level survival probability, estimated using KM model on the dataset $\mathcal{D}$. The idea here is that we measure the contribution of patient i to the unbiased predictor $\hat{\theta}$. The best guess for MAE-PO can be defined as:

$$e_{\text{T-pseudo-obs}}(t_i, \mathcal{D}) \quad = \quad N \times \hat{\theta}\ -\ (N-1) \times \hat{\theta}^{-i}\ , \tag{E.3}$$

where $\hat{\theta}^{-i}$ is $\mathbb{E}_t[S_{\text{KM}(\mathcal{D}^{-i})}(\ t\ )]$, the predictor applied to the $N-1$ data instances, after removing the patient i. This best guess can be unreliable for patients who get censored earlier in the study since we do not have much information about them. Therefore, as suggested by Haider et al. [38], we assign less confidence weight to the best guess of early censored patients. This confidence weight is calculated as:

$$\omega_i \quad = \quad 1 - S_{\text{KM}(\mathcal{D})}(t_i) \ . \tag{E.4}$$

Note ω_i is zero in the beginning (at time zero), and increases after that. Lastly, MAE-PO is defined as:

$$\mathbb{E}_{i \sim \mathcal{D}}[\mathcal{R}_{\text{MAE-PO}}(\hat{t}_i, t_i, \delta_i)] = \tag{E.5}$$
$$\frac{1}{\sum_{i=1}^{N} \omega_i} \sum_{i=1}^{N} \omega_i \left| [(1-\delta_i) \cdot e_{\text{T-pseudo-obs}}(t_i, \mathcal{D}) + \delta_i \cdot t_i] - \hat{t}_i \right| ,$$

where symbol $\mathcal{R}$ means a scoring rule, which is used to compute the MAE-PO error. Note that here the prediction time ($\hat{t}_i$) is the median of the ISD model.

5. **Truncated MAE-PO:**

As discussed in Section 3.3, we choose to bound the prediction time and best guess by the end of the study and use the truncated variation of MAE-PO. Hence, the best guess for truncated MAE-PO can be defined as:

$$e_{\text{T-pseudo-obs},\tau}(t_i, \mathcal{D}) \quad = \quad \min\{e_{\text{pseudo-obs}}(t_i, \mathcal{D}) \ \ , \tau\}, \tag{E.6}$$

where $e_{\text{pseudo-obs}}(t_i, \mathcal{D})$ is defined using Euqation E.3. Further, we use the same weighting as described in Equation E.4. Therefore, the truncated MAE-PO is defined as follows:

$$\mathbb{E}_{i\sim\mathcal{D}}[\mathcal{R}_{\text{T-MAE-PO},\tau}(\hat{t}_i, t_i, \delta_i)] = \frac{1}{\sum_{i=1}^{N}\omega_i}\sum_{i=1}^{N}\omega_i\left|[(1-\delta_i)\cdot e_{\text{T-pseudo-obs},\tau}(t_i,\mathcal{D}) + \delta_i\cdot t_i] - \hat{t}_i\right|, \tag{E.7}$$

where the prediction time ($\hat{t}_i$) is the **truncated median time** ($\hat{t}_i = \hat{t}_{i,\text{T-median},\tau}$) of the ISD defined in Equation B.2, and we use the **truncated best guess** ($e_{\text{T-pseudo-obs},\tau}(t_i, \mathcal{D})$) defined in Equation E.6.

6. **Truncated-Log MAE-PO:**

The truncated-log (TL) adaptation of MAE-PO is:

$$\mathbb{E}_{i\sim\mathcal{D}}[\mathcal{R}_{\text{TL-MAE-PO},\tau}(\hat{t}_i,\, t_i,\, \delta_i)] = \frac{1}{\sum_{i=1}^{N}\omega_i}\sum_{i=1}^{N}\omega_i\left|[(1-\delta_i)\cdot\log(e_{\text{T-pseudo-obs},\tau}(t_i,\mathcal{D})) + \delta_i\cdot\log(t_i)] - \log(\hat{t}_i)\right|, \tag{E.8}$$

where again the prediction time ($\hat{t}_i$) is the **truncated median time** ($\hat{t}_i = \hat{t}_{i,\text{T-median},\tau}$) of the ISD defined in Equation B.2, and we use the **truncated best guess** ($e_{\text{T-pseudo-obs},\tau}(t_i, \mathcal{D})$) defined in Equation E.6. For Equation E.8, if the predicted time to event, the ground truth, or the best guess is zero, we initially add a small value (ϵ) to prevent the logarithm function from yielding minus infinity. Moreover, we choose to use log base e.

To further understand how we can interoperate error measured by TL-MAE-PO, recall that the TL-MAE-PO for AFT on the Prostate #1 dataset is 0.62 ± 0.001. Here, given that $\exp(0.62) = 1.86$, this is claiming that we expect each prediction to be within a multiplicative factor of 1.86 of the correct value. So, for instance, if we predict patient A will live 9.02 months, we are saying that we anticipate that patient A will live between $(9.02/1.86,\ 9.02 \times 1.86) = (4.82,\ 31.19)$

months. If another patient was predicted to live 9.02 days, then we would anticipate that person would live between (4.82, 31.19) days. This is the nature of multiplicative bounds.

Appendix F: Model Implementation Details

In this section, we included details of the model implementation that was used in this paper.

- **Kaplan Meier (KM)** is a popular estimator that uses the information of a group of patients. The KM curve provides a stepwise estimate of the probability of event occurrence. We used `KaplanMeierFitter` class from `lifelines` library, and we used the median time of the training population as the time to event prediction for the test set.

- **Random Survival Forest (RSF):** is an extension of the Random Forest algorithm for time-to-event data, offering a non-parametric approach to model survival outcomes. RSF is an ensemble of survival trees, each learned on a bootstrapped version of the training dataset. We used `RandomSurvivalForest` class from `sksurv.ensemble` library for implementation, with 150 trees, min samples split of 25, and min samples leaf of 20.

- **Multi-Task Logistic Regression (MTLR)** is a machine learning approach designed for survival analysis and gives individualized curve prediction. The model is implemented using `MTLR` class from `torchmtlr` package, with a learning rate of 0.001, batch size of 512, and 500 epochs.

- **Deep MTLR** is another method that predicts individualized survival distribution and uses the MTLR models as its base and a deep learning model as its core. We implemented D-MTLR using `DeepMTLR` class from `torchmtlr` package, with the same configuration as MTLR. The architecture of the model is provided in the code base, and it includes layers of NN nodes, dropout of 0.4, and Exponential Linear Units (ELUs).

- **DeepHit** learns the individualized survival distribution using deep learning. The model is implemented using `DeepHitSingle` class from `pycox.models` package. We used Adam optimizer with early stopping.

- **Cox Proportional Hazard (Cox-PH)** is a semi-parametric method used in survival analysis to assess the impact of several risk factors on survival time. It provides hazard ratios, indicating the relative risk of event occurrence given a change in predictor variables. It is composed of a baseline hazard function at the population level (non-parametric) and a parametric partial hazard function. We implemented Cox-PH using `CoxPHSurvivalAnalysis` class from `sksurv.linear_model` library.

- **Accelerate Failure time (AFT)** is a parametric survival analysis technique that directly models the time to event and provides individualized prediction. We employed AFT with Weibull parametric assumption. For implementation, we employed the `WeibullAFTFitter` class from `lifelines` library. We used median time for the time-to-event prediction, and based on our experiments, it works better than using the average time.

Appendix G: Detailed evaluation of selected cancer types

Tables G.1, G.2, G.3, G.4, G.5, G.6, G.7, G.8, G.9, and G.10 show the evaluation of various models by each of the discussed metrics for the selected cancer types. For all the tables, the reported C-index and BS are computed at the median time, except for table G.6, in which we computed the C-index and BS at the 10-year time point since we wanted to compare our results with the reported results of Survival Quilts model [36]. In terms of TL-MAE-PO and T-MAE-PO, our results show that RSF followed by Deep-MTLR are the top-performing methods in all the datasets except for the prostate # 1 dataset where AFT is the best.

Table G.1: Model comparison using all the discussed metrics based on monthly survival data for **brain** cancer. The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate the best performance over the included set of models.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**74.91 ± 0.35**	72.46 ± 0.39	74.68 ± 0.39	73.45 ± 0.28	72.63 ± 0.41	72.64 ± 0.43	50.00 ± 0.00
BS(median)	**0.16 ± 0.00**	0.18 ± 0.00	**0.16 ± 0.00**	0.20 ± 0.00	0.17 ± 0.00	0.17 ± 0.00	0.22 ± 0.00
T-MAE-PO	**46.12 ± 0.53**	51.97 ± 0.82	47.99 ± 0.80	58.49 ± 0.26	48.99 ± 0.51	49.30 ± 0.57	156.56 ± 0.77
TL-MAE-PO	**1.42 ± 0.01**	1.52 ± 0.02	1.44 ± 0.01	1.70 ± 0.00	1.48 ± 0.01	1.49 ± 0.01	2.75 ± 0.00

Table G.2: Model comparison using all the discussed metrics based on monthly survival data for **breast** cancer. The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate the best performance over the included set of models.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**82.66 ± 0.08**	77.23 ± 0.53	80.71 ± 0.15	80.76 ± 0.22	76.86 ± 0.09	77.03 ± 0.09	50.00 ± 0.00
BS(median)	**0.06 ± 0.00**	0.07 ± 0.00	**0.06 ± 0.00**	0.07 ± 0.00	0.07 ± 0.00	0.07 ± 0.00	0.08 ± 0.00
T-MAE-PO	**65.66 ± 0.13**	69.76 ± 3.17	65.74 ± 1.87	85.09 ± 0.63	72.46 ± 0.16	72.71 ± 0.15	99.15 ± 0.01
TL-MAE-PO	**0.93 ± 0.01**	0.97 ± 0.03	0.94 ± 0.02	1.14 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.22 ± 0.00

Table G.3: Model comparison using all the discussed metrics based on monthly survival data for **kidney and renal pelvis** cancer. The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate the best performance over the included set of models.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**86.26 ± 0.06**	79.65 ± 0.24	86.22 ± 0.11	84.59 ± 0.07	79.94 ± 0.14	80.03 ± 0.14	50.00 ± 0.00
BS(median)	**0.07 ± 0.00**	0.09 ± 0.00	**0.07 ± 0.00**	0.10 ± 0.00	0.09 ± 0.00	0.09 ± 0.00	0.12 ± 0.00
T-MAE-PO	51.94 ± 1.26	66.59 ± 2.13	**51.59 ± 1.53**	78.70 ± 2.36	68.56 ± 1.18	69.55 ± 1.2	117.74 ± 1.47
TL-MAE-PO	**1.20 ± 0.02**	1.43 ± 0.04	1.21 ± 0.03	1.71 ± 0.03	1.45 ± 0.02	1.46 ± 0.02	2.00 ± 0.02

Table G.4: Model comparison using all the discussed metrics based on monthly survival data for **liver** cancer. The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate the best performance over the included set of models.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	74.50 ± 0.21	70.30 ± 0.26	**74.56 ± 0.18**	73.41 ± 0.22	70.35 ± 0.25	70.30 ± 0.25	50.00 ± 0.00
BS(median)	**0.17 ± 0.00**	0.19 ± 0.00	**0.17 ± 0.00**	0.22 ± 0.00	0.19 ± 0.00	0.19 ± 0.00	0.23 ± 0.00
T-MAE-PO	**39.98 ± 0.95**	42.99 ± 3.36	40.86 ± 3.14	45.99 ± 4.50	44.10 ± 1.49	46.37 ± 1.34	174.60 ± 5.54
TL-MAE-PO	**1.93 ± 0.05**	2.07 ± 0.03	1.95 ± 0.04	2.24 ± 0.01	2.07 ± 0.05	2.11 ± 0.05	3.62 ± 0.11

Table G.5: Model comparison using all the discussed metrics based on monthly survival data for **lung and bronchus** cancer. The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate the best performance over the included set of models.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**73.02 ± 0.07**	68.88 ± 0.23	72.20 ± 0.11	71.72 ± 0.12	68.96 ± 0.09	69.06 ± 0.09	50.00 ± 0.00
BS(median)	**0.18 ± 0.00**	0.20 ± 0.00	**0.18 ± 0.00**	0.21 ± 0.00	0.20 ± 0.00	0.20 ± 0.00	0.23 ± 0.00
T-MAE-PO	**42.81 ± 0.50**	47.93 ± 1.55	43.62 ± 1.44	51.08 ± 1.71	47.59 ± 0.78	47.43 ± 0.99	164.66 ± 1.49
TL-MAE-PO	**1.44 ± 0.00**	1.57 ± 0.02	1.47 ± 0.01	1.66 ± 0.01	1.55 ± 0.01	1.55 ± 0.01	2.81 ± 0.01

Table G.6: Model comparison using all the discussed metrics based on monthly survival data for **prostate #1** cancer. The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate the best performance over the included set of models.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	Survival Quilts	KM (baseline)
C-index	85.59 ± 0.71	83.11 ± 1.27	**86.05 ± 0.71**	75.57 ± 2.86	85.14 ± 0.74	85.31 ± 0.75	82.90 ± 0.09	50.00 ± 0.00
BS (10-years)	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.04 ± 0.00
T-MAE-PO	46.52 ± 0.30	44.65 ± 2.54	45.74 ± 2.20	47.39 ± 0.89	45.05 ± 0.25	**44.69 ± 0.32**	-	49.37 ± 0.22
TL-MAE-PO	0.64 ± 0.00	0.62 ± 0.02	0.63 ± 0.02	0.65 ± 0.01	0.63 ± 0.00	**0.62 ± 0.00**	-	0.67 ± 0.00

Table G.7: Model comparison using all the discussed metrics based on monthly survival data for **prostate #2** cancer. The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate the best performance over the included set of models.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**88.31 ± 0.10**	75.39 ± 5.93	84.5 ± 0.40	85.94 ± 1.10	83.60 ± 0.08	84.15 ± 0.09	50.00 ± 0.00
BS(median)	**0.04 ± 0.00**	0.10 ± 0.06	0.05 ± 0.00	0.05 ± 0.00	0.05 ± 0.00	0.05 ± 0.00	0.06 ± 0.00
T-MAE-PO	**56.01 ± 1.17**	81.89 ± 6.31	65.08 ± 3.52	80.36 ± 3.52	62.76 ± 1.26	64.15 ± 1.30	94.68 ± 1.49
TL-MAE-PO	**0.81 ± 0.01**	1.11 ± 0.17	0.89 ± 0.04	1.06 ± 0.02	0.88 ± 0.01	0.89 ± 0.00	1.14 ± 0.01

Table G.8: Model comparison using all the discussed metrics based on monthly survival data for **stomach** cancer. The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate the best performance over the included set of models.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**77.00 ± 0.14**	72.76 ± 0.33	76.68 ± 0.13	75.25 ± 0.15	71.80 ± 0.21	71.87 ± 0.20	50.00 ± 0.00
BS(median)	**0.16 ± 0.00**	0.18 ± 0.00	**0.16 ± 0.00**	0.20 ± 0.00	0.18 ± 0.00	0.18 ± 0.00	0.22 ± 0.00
T-MAE-PO	**40.21 ± 0.52**	50.41 ± 0.93	49.26 ± 2.22	54.76 ± 2.40	51.07 ± 0.69	52.13 ± 1.31	159.69 ± 2.36
TL-MAE-PO	**1.53 ± 0.00**	1.72 ± 0.02	1.61 ± 0.01	1.85 ± 0.02	1.74 ± 0.01	1.78 ± 0.01	3.03 ± 0.02

Table G.9: Model comparison using all the discussed metrics based on monthly survival data for **thyroid** cancer. The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate the best performance over the included set of models.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	93.22 ± 0.22	89.38 ± 0.34	**93.30 ± 0.34**	92.22 ± 0.31	90.27 ± 0.20	90.20 ± 0.20	50.00 ± 0.00
BS(median)	**0.01 ± 0.00**	0.02 ± 0.00	**0.01 ± 0.00**	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
T-MAE-PO	**56.42 ± 0.34**	65.71 ± 3.70	57.35 ± 3.31	78.16 ± 3.16	72.01 ± 0.72	74.94 ± 0.60	98.00 ± 0.13
TL-MAE-PO	**1.29 ± 0.01**	1.43 ± 0.06	1.33 ± 0.06	1.66 ± 0.03	1.52 ± 0.01	1.55 ± 0.01	1.81 ± 0.00

Table G.10: Model comparison using all the discussed metrics based on monthly survival data for **urinary bladder** cancer. The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate the best performance over the included set of models.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**81.73 ± 0.18**	74.41 ± 0.68	81.34 ± 0.27	80.94 ± 0.20	74.64 ± 0.19	74.73 ± 0.18	50.00 ± 0.00
BS(median)	**0.08 ± 0.00**	0.12 ± 0.01	0.08 ± 0.00	0.10 ± 0.00	0.10 ± 0.00	0.10 ± 0.00	0.12 ± 0.00
T-MAE-PO	64.55 ± 1.69	80.42 ± 10.29	**62.40 ± 2.95**	97.11 ± 2.92	79.88 ± 1.03	80.21 ± 1.05	117.51 ± 2.83
TL-MAE-PO	1.23 ± 0.01	1.44 ± 0.11	**1.22 ± 0.04**	1.68 ± 0.02	1.46 ± 0.01	1.46 ± 0.01	1.82 ± 0.02