Counterfactual Inference Framework for Joint Estimation of Medical Costs, Treatment Effect and Time-to-Event

A Project Report

submitted by

ZUBER KHAN

MT21181

in partial fulfilment of the requirements

for the award of the degree of

MASTER OF TECHNOLOGY

ELECTRONICS AND COMMUNICATION ENGINEERING

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

December, 2022

THESIS CERTIFICATE

This is to certify that the thesis titled **Counterfactual Inference Framework for Joint Estimation of Medical Costs, Treatment Effect and Time-to-Event**, submitted by **Zuber Khan**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Master of Technology**, is a bonafide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Ranjitha Prasad
Thesis Supervisor
Assistant Professor
Department of Electronics and Communication Engineering
IIIT Delhi, 110020

Place: New Delhi

December, 2022

ACKNOWLEDGEMENTS

I want to express my deep gratitude to Dr. Ranjitha Prasad for her constant guidance, support, and extraordinary supervision throughout this research work. Further, I am indebted to Ms. Muskan Gupta for her exemplary help during the crucial stage of this research. Finally, I thank my family for giving me all the care, love, and blessings during this journey.

ABSTRACT

KEYWORDS: Medical Cost ; Counterfactual Inference ; Survival Analysis ; Deep Neural Network

Optimal treatment selection is extremely crucial in emergency situations such as for a patient admitted in ICU. However, the chosen medical treatment may not necessarily be a financially favorable choice. In some of the developing countries like India, there is a lack of a good public health insurance system and not everyone can afford private health insurance. Therefore, balancing the trade-off between medical costs and best treatment choice is important not just from patient point of view but also for settings where free healthcare services are provided. In order to make the best decision, it is essential for clinicians to know about the treatment which will lead to shorter duration of inpatient hospital stay as well as the associated medical costs with each potential treatment choice. Various authors have tried to predict duration of stay or medical costs of inpatient ICU stays or the treatment effect with time-to-event as the outcome variable. However, to the best of our knowledge, there is no research work that proposes joint estimation of medical cost and duration of stay in a hospital taking into consideration individual treatment effect. Our research work addresses this issue, and provides two novel frameworks MedCI and MedSCI, that not only predict time-to-stay and associated medical costs for a given treatment and counterfactual treatment choice but also return individual treatment effect for both the outcomes. Our work is a mixture of Survival Analysis, Causal Inference and Deep Learning. The results are obtained on a semi-synthetic and synthetic dataset for MedCI while MedSCI is evaluated on a synthetic dataset.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Problem Definition

1.1.1 Hospital Time-to-Stay and Medical Cost Estimation

The cost of treatment for healthcare management is proportional to resource consumption (in terms of rupees) for a given medical intervention. Effective cost computation is essential in the economic evaluation of medical interventions. It is to be noted that medical cost scales with the number of days of hospital stay (especially ICU stay) during the treatment period. Comparison of the average medical cost and time-to-stay in a hospital associated with alternative therapies can lead to the optimal choice of treatment and substantial cost reduction, leading to better resource management for the institutions providing free health care to the public.

1.1.2 Optimal Treatment Selection

To know the optimal treatment out of potential treatment options, we need to understand the individual and average treatment effect on time to stay and incurred medical cost of admitted patients during that stay. However, the task of predicting time-to-stay and incurred medical expenses in the counterfactual world is not straightforward. There are state-of-the-art methods for predicting time to event in survival analysis literature. Further, estimating treatment effect for binary or continuous outcomes has also received much attention. However, only a few approaches enable causal effect estimation with time to event outcome since one needs to take into consideration two aspects :

1. The treatment assignment should be independent of potential outcomes.
2. There may be patients who drop out of study before the event (discharge from hospital/death) actually happens.

The problem gets even more complex, when we have a multiple outcome scenario i.e. time-to-stay prediction and medical expense estimation.

The subsequent sections introduce the basic concepts and terminologies of survival analysis and causal inference along with assumptions involved in relevance to this work.

1.2 Time-to-Event Analysis

Time-to-event analysis, or survival analysis, refers to statistical methodologies for studying the length of time during a well-defined event interval. In our context, this duration is from some well-defined point of patient entry (e.g., start of treatment) to some well-defined event (e.g. death or discharge). Because not all patients experience the event (e.g. death or discharge) at the conclusion of the observation period, the actual survival times for some patients remain unknown. This tendency, known as censoring, must be accounted for in the study to draw appropriate conclusions. In medical research, the occurrence of a well-defined event, such as discharge from the hospital, is frequently the main outcome and effectively binary in nature, i.e., whether the event has occurred or not. Therefore, in survival analysis, the outcome of interest is not only whether an event occurred but also when that event occurred. We need unique statistical tools for this since traditional regression methods cannot deal with censoring and are unsuitable for including event and temporal elements simultaneously as output in the model.

It is noteworthy that in the study of time to event or survival data, there are four major factors to consider:-

- Target event
- Time origin
- Time scale
- Description of how participants will exit the study

Time origin is the initiation of the survival analysis study. Some examples include but are not limited to the birth of a child, hospital admission time of a patient, disease occurrence, etc. The time scale could be calendar years, age, or the number of days since the beginning of treatment.

Few participants in the study will not experience the target event and thus will be censored. There are three types of censoring in survival analysis data :-

1. If the event happens after the conclusion of the research, the data is **right-censored**.
2. When an event is detected but the actual occurrence time is unclear, the data is **left-censored**.
3. **Interval-censored** data arises when an event is seen but individuals arrive and go, making the actual event time uncertain.

All these types of censoring are illustrated in Figure 1.1 [1].

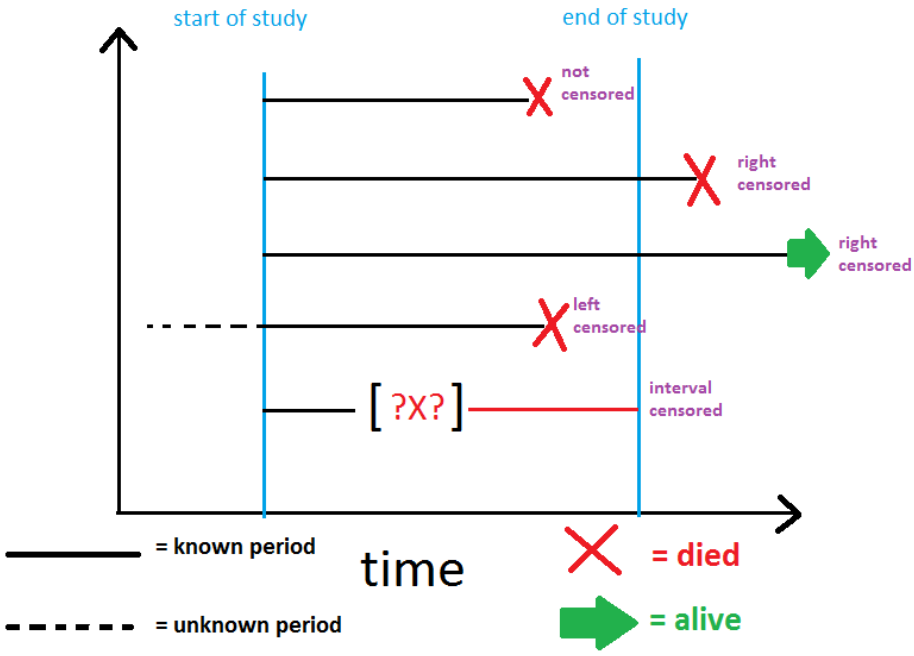

Figure 1.1: Illustration of censoring in survival analysis

Apart from the previously indicated categorization, there is another distinction based on whether censoring is random or informative. Random censoring implies that the reason for the premature exit of participants is unrelated to the treatment assigned or study process. On the other hand, informative censoring occurs when individuals withdraw from the research for reasons related to the study itself. For instance, in research comparing the survival rate of patients in alternative cancer therapies, a high recurrence rate may appear in the ineffective treatment option leading to patients avoiding any follow-up. Contrary to that, patients who received highly effective treatment will have no longer need to stay or visit the hospital and hence will drop out of the study. Therefore, in such a scenario, disease-free survival rates would be predicated on patients

who remained unfettered in the trial and would be inflated for the unsuccessful therapy and understated for the beneficial treatment. Censoring is assumed to be random or non-informative in most analytical methodologies. Also, majority of survival analysis methods are built for right-censored observations, although there are ways for interval and left-censored data as well. Now, there are certain important aspects that one need to know during time-to-event analysis such as :-

- What percentage of people will survive the event after a particular period of time?
- What percentage of people will experience the event after a particular period of time?
- What is the risk of the event occurring at a specific moment in time among individuals who have survived up to that point?

To address these questions, there exist certain probabilistic functions used in survival analysis such as survival function, hazard function, instantaneous hazard rate, cumulative incidence function and cumulative hazard function.

1.2.1 Survival Function

The survival function is the statistical formulation that enables one to inquire whether the participant under observation in the study will survive beyond the specific time point. The participant can be an industrial tool, a patient in a hospital, a manufacturing unit or anything under observation for a well-defined period of time. Mathematically, survival function is the probability that the subject of interest survives beyond a time t i.e

$$S(t) = P(T > t) \tag{1.1}$$

where T refers to time elapsed from beginning or baseline time to the time-to-event of subject under consideration. As time t varies from 0 to ∞ , the survival function has following properties :-

- It is non increasing in nature
- At the origin, i.e. t = 0, S(t) = 1. This is quite obvious since probability of survival at the beginning of study is always unity.
- At time t = ∞, S(t) = S(∞) = 0. As time horizon approaches to infinity, the survival curve approaches zero.

The survival curve is depicted in Figure 1.2 [2].

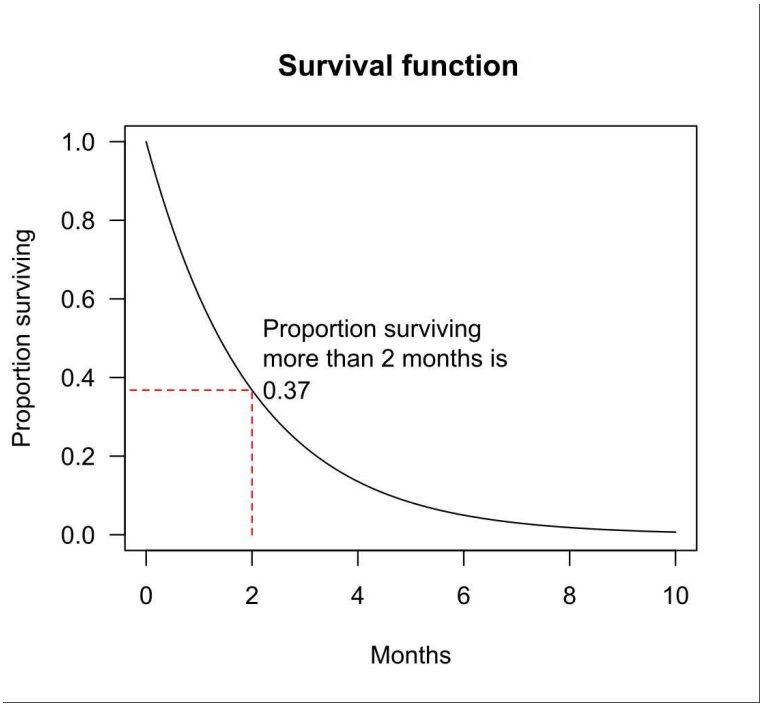

Figure 1.2: Survival Curve

1.2.2 Cumulative Incidence Function

It refers to the probability that the subject of interest will survive until or less than a specific time point t. Mathematically,

$$F(t) = P(T \leq t) \tag{1.2}$$

Clearly,

$$F(t) = 1 - S(t) \tag{1.3}$$

1.2.3 Hazard Function

The hazard function is the instantaneous rate of encountering an event at time t, conditioned that the subject survived up to that point. Mathematically, it is expressed as:-

$$\begin{aligned} h(t) &= \lim_{dt \to 0} \frac{P(t < T \leq t + dt | T > t)}{dt} \\ &= \lim_{dt \to 0} \frac{P(t < T \leq t + dt, T > t)}{P(T > t)dt} \\ &= \lim_{dt \to 0} \frac{P(t < T \leq t + dt)}{P(T > t)dt} \end{aligned} \tag{1.4}$$

Now, recall that the probability density function of the survival distribution is defined as:

$$\begin{aligned} f(t) &= \frac{d}{dt} F(t) \\ &= \lim_{dt \to 0} \frac{F(t + dt) - F(t)}{dt} \\ &= \lim_{dt \to 0} \frac{P(t < T \leq t + dt)}{dt} \end{aligned} \tag{1.5}$$

Also,

$$S(t) = P(T > t) \tag{1.6}$$

Substituting 1.5 and 1.6 in 1.4 , the formula can be simplified as :

$$h(t) = \frac{f(t)}{S(t)} \tag{1.7}$$

The hazard rate can also be expressed as a function of the survival function, as proven below:

Since,

$$h(t) = \frac{f(t)}{S(t)} \tag{1.8}$$

and,

$$f(t) = \frac{d}{dt} F(t) \tag{1.9}$$

This implies,

$$\begin{aligned} h(t) &= \frac{1}{S(t)} \cdot \frac{\mathrm{d}F(t)}{\mathrm{d}t} \\ &= \frac{1}{S(t)} \cdot \frac{\mathrm{d}(1 - S(t))}{\mathrm{d}t} \\ &= \frac{1}{S(t)} \cdot -\frac{\mathrm{d}(S(t))}{\mathrm{d}t} \\ h(t) &= \frac{-d}{dt} \log S(t) \end{aligned} \tag{1.10}$$

1.2.4 Cumulative Hazard Function

The cumulative hazard reflects the hazard that has accrued up to time t. Mathematically, it is the integral of hazard function from origin to time t :-

$$H(t) = \int_0^t h(t)\, dt \tag{1.11}$$

It also implies that the cumulative hazard function is equal to area under the curve of hazard function

1.2.5 Time-to-Event Data Representation in Censoring

In the presence of censoring, survival data can be represented like this :-

- Let T_i be the time-to-event of i^{th} subject under consideration.
- Let C_i be the censoring time of i^{th} subject under consideration.
- δ_i is the censoring indicator i.e.

$$\delta_i = \left\{ \begin{array}{ll} 1, & \text{if the subject was censored } (T_i > C_i) \\ 0, & \text{if the event was observed } (T_i \leq C_i) \end{array} \right\} \tag{1.12}$$

Then, the observed time-to-event Y_i becomes,

$$Y_i = min(T_i, C_i) \tag{1.13}$$

1.2.6 Survival Analysis Techniques

Various survival analysis techniques used to estimate the survival function can be broadly classified in three categories :-

- Non Parametric Estimation
- Parametric Estimation
- Semi Parametric Estimation

Non Parametric Estimation

Non-parametric estimation of the survival curve does not make any assumptions about the underlying distribution or parameters of the function used to estimate it. The estimating function can take any form as per the data. A significant advantage of this approach is that the parameters of the estimator are pretty flexible. However, the model complexity increases as the number of data points grow. Therefore, it is slower as it needs a large amount of data for good estimation. Also, the estimated survival function is not smooth but piece-wise constant when the amount of data is less, resulting in a jump of time-to-event probabilities across the time horizon, which is an unrealistic property. Some prominent examples of non-parametric estimators in survival analysis are :-

- **Kaplan-Meier Estimator**
  - The Kaplan-Meier estimator [3] divides the estimation of S(t) into steps/intervals based on observed event timings resulting into a final estimated survival curve.
  - The observations are used to generate the Kaplan-Meier curve until the time-to-event occurs or the subject is censored.
  - Given that individuals are at risk at the start of the period, the chance of surviving until the conclusion of the interval is determined for each interval. Mathematically, it is represented as :-

$$P_i = \frac{n_i - d_i}{n_i} \tag{1.14}$$

    where n_i represents total subjects at risk in i^{th} interval while d_i represents participants who observed time-to-event in that interval.
  - Then, the estimated survival function value at time t is calculated as:-

$$S(t) = S(t-1) \cdot P_i \tag{1.15}$$

– Finally, the survival function is plotted for the entire duration of study. Once the survival function is known, the hazard function can be easily calculated to plot the hazard curve.

The Kaplan-Meier Curve is shown in Figure 1.3 [4].

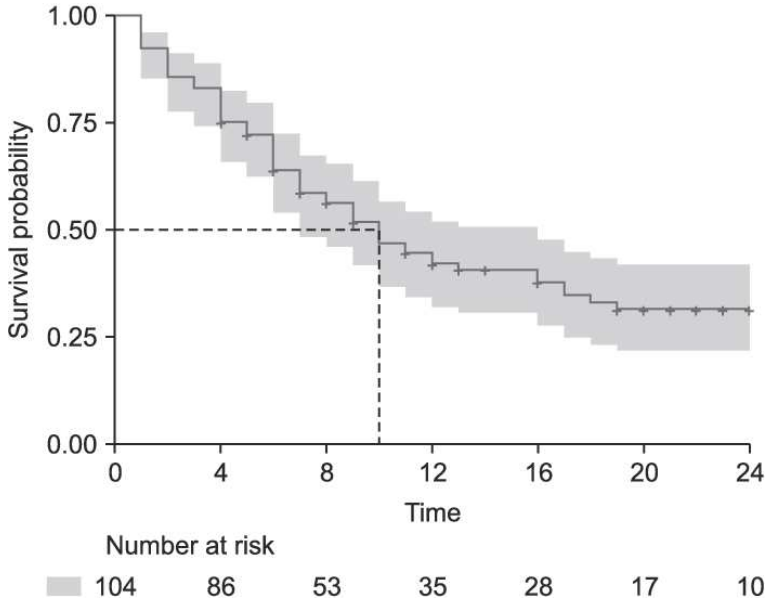

Figure 1.3: Kaplan-Meier Curve with 95% confidence interval (shadow)

- **Nelson-Aalen Estimator**
  - The Nelson-Aalen Estimator [5] models the cumulative hazard function, H(t), using a counting process technique. The mathematical definition is as follows :-

$$H(t) = \sum_{t_i \leq t} \frac{d_i}{n_i} \tag{1.16}$$

where d_i represents the number of events observed at t_i and n_i represents the total individuals at risk at t_i

The Nelson-Aalen Curve for disablity pension due to low back pain is depicted in Figure 1.4 [6].

Parametric Estimation

The parametric estimation assumes that the survival or hazard function follows a particular distribution which means that the estimator's functional form is known beforehand. Unlike the non-parametric approach, the number of parameters of the estimator is fixed here. It overcomes the limitation of step-wise survival function in non-parametric approach for small amount of data because if the chosen estimator is continuous, the

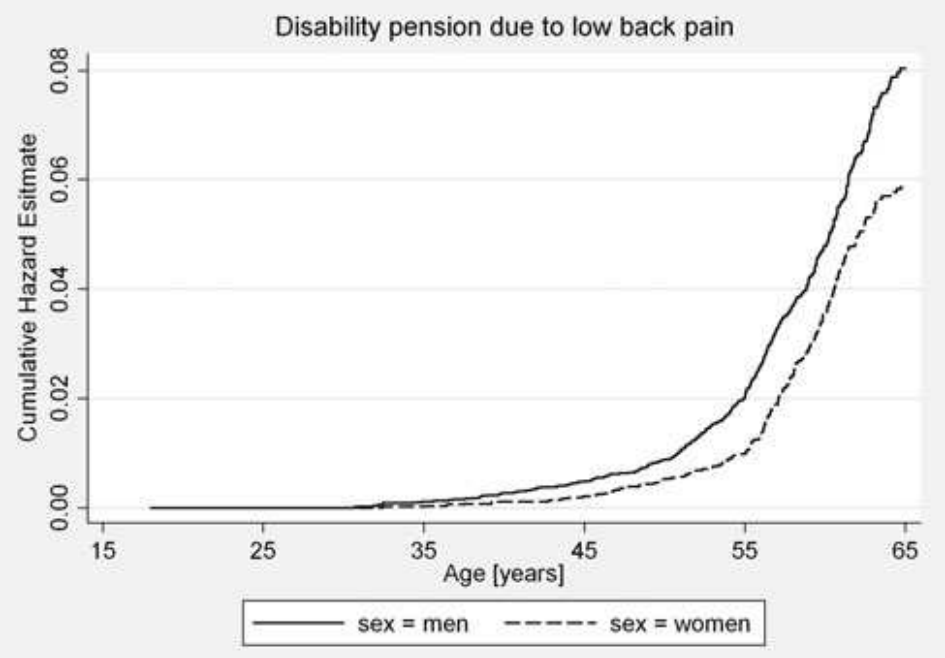

Figure 1.4: Nelson-Aalen Curve

resultant survival curve would be smooth irrespective of the amount of data. A significant challenge in this approach is that the chosen estimator functional form should mirror the actual data distribution. It can be checked using hypothesis testing or visualization techniques. If this condition holds, the parametric approach does not need much data for generalization. Some prominent examples of parametric estimators in survival analysis are :-

- **Accelerated Failure Time Models**
  - These algorithms [7] are a group of parametric survival estimators that establish a linear relationship between the natural log of the time-to-event and covariates. Mathematically, it is represented as:-

$$\log T = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \sigma\epsilon \tag{1.17}$$

    Here, β_0 is intercept term, σ scale parameter, $\beta_1 ... \beta_n$ represents the covariates coefficients while ϵ is the residual term which follows a specific distribution.
  - Based on the residual term ϵ distribution, there is corresponding time-to-event T distribution. For instance, T follows the Weibull distribution if ϵ has an extreme value distribution. Again, if ϵ has a normal distribution, T has a log-normal distribution, and so on.
  - Some examples of distributions used in accelerated failure time algorithms for modelling time-to-event T are log-logistic, weibull, log-normal, exponential, gamma and inverse gaussian.

Semi-Parametric Estimation

The combination of parametric and non-parametric approach is called semi-parametric. The underlying mathematical formulation of the estimator has both parametric as well as non-parametric component. An outstanding example for this category in survival analysis is **Cox Proportional-Hazards model**. The model's [8] goal is to assess the impact of several variables on survival at the same time. In other words, it enables to investigate how certain circumstances impact the rate of occurrence of a specific event (e.g., infection, death) at a given point in time. Recall that the hazard rate reflects the instantaneous rate of occurrence of an event.

- The cox-proportional hazards model formulates the hazard function as follows :-

$$h(t) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) \tag{1.18}$$

  where β_1,β_2....β_p are coefficients of covariates and $h_0(t)$ refers to baseline hazard which is function of time and corresponds to the value of hazard function when all the covariates(x_1,x_2.....x_p) are zero.

- The hazard function is clearly the product of two quantities, one being baseline hazard which is a non-parametric component, and the other is exponentiated linear function of a collection of p explanatory variables, a parametric quantity.

- Since the baseline hazard function is computed non-parametrically, the survival times are not expected to follow a specific statistical distribution, and the form of the baseline hazard is arbitrary.

- In this model, regardless of explanatory variables values (x_1,x_2,....x_p),it is assumed that all participants have the same baseline hazard. Therefore, the ratio of two hazard function is independent of baseline hazard. Hence, to make conclusions about the relative hazard or the hazard ratio, the baseline hazard function does not need to be calculated. It also implies that the hazard ratio is not susceptible to baseline hazard misspecification, making the cox hazard estimator more robust than parametric methods.

- Since the parametric quantity in hazard function is independent of time, it multiplies the baseline hazard by same amount for a participant irrespective of time during the study.

- It also means that given two participants in the study, their hazard ratio stays proportional, which is also known as **proportional hazards assumption**. This can be mathematically proven as follows :-
  Let patient 1 has hazard function $h_1(t) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$
  Let patient 2 has hazard function $h_2(t) = h_0(t) \exp(\beta_1 x_1^{'} + \beta_2 x_2^{'} + ... + \beta_p x_p^{'})$

$$\frac{h_1(t)}{h_2(t)} = \frac{\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{\exp(\beta_1 x_1^{'} + \beta_2 x_2^{'} + ... + \beta_p x_p^{'})}$$

$$h_1(t) = \exp(\beta(x - x^{'}))h_2(t)$$

For instance, if an individual's risk of death is twice as high at an initial time point with respect to another individual, then the risk of death stays twice as high at all subsequent time points.

1.3 Causal Inference

Causal inference as the name suggests, refers to the study of the cause behind a certain effect or outcome. It involves analyzing the intervention that led to the outcome and reasoning whether the applied intervention has actually caused the effect. It also explores the world of counterfactuals and throws light on what will happen in a parallel world with an alternate intervention. The topic is of immense importance in various domains. A few examples are :-

Healthcare: Suppose scientists have developed two vaccines for a virus. It is crucial to figure out which vaccine is more effective against the disease.

Policy Making: Policy makers want to design certain policies to curb the menace of pollution in Delhi. It is essential to know what human activities are majorly causing significant increments in pollution levels. It is equally important to understand which policy will lead to maximum decrease in pollution.

Business: Consider the topic of predicting client churn: the results are frequently meaningless if the cause cannot be identified. One purpose for forecasting churn is to determine what form of action would be most effective in retaining a loyal client.

Psychology: In order to realize the effect of social media on human behaviour, it is essential to know the main social media component causing a specific trait.

1.3.1 Statistical Notation and Definitions in Causal Inference

T : random variable for treatment/intervention

Y : random variable for outcome of interest

X : covariates

Mostly, T is binary in nature { 0,1} but it can also take more than two values or continuous values.

Y(t) : potential outcome which refers to the outcome upon treatment t

A potential outcome differs from the observed outcome in the sense that not all poten-

tial outcomes are actually observed. Instead, all potential outcomes can be potentially observed. The one observed is determined by the value of treatment T.

Let T_i, X_i, Y_i be the treatment, covariates and outcome of the i_{th} individual in the population, then the **individual treatment effect (ITE)** is defined as:

$$\tau_i = Y_i(1) - Y_i(0) \tag{1.19}$$

1.3.2 Fundamental Problem of Causal Inference

Note that in a real world scenario, only one outcome is observed out of all possible potential outcomes. For instance, while comparing the effect of alternative vaccines, a patient can only be given one vaccine to protect against the disease. So for any patient, only one potential outcome is observed which depends on the treatment assigned. This is the fundamental problem because if both $Y_i(1)$ and $Y_i(0)$ are not observed, the causal effect $Y_i(1) - Y_i(0)$ cannot be observed. The potential outcomes that are not observed are called **counterfactuals**.

1.3.3 Solution to the Fundamental Problem

Average Treatment Effect

The average treatment effect (ATE) is obtained by averaging over the individual treatment effect (ITE):

$$\tau = \mathbf{E}[Y_i(1) - Y_i(0)] = \mathbf{E}[Y(1) - Y(0)] \tag{1.20}$$

By linearity of expectation:

$$\tau = \mathbf{E}[Y(1) - Y(0)] = \mathbf{E}[Y(1)] - \mathbf{E}[Y(0)] \tag{1.21}$$

Now, the **associational difference** of the two outcomes is defined as:

$$\mathbf{E}[Y|T = 1] - \mathbf{E}[Y|T = 0] \tag{1.22}$$

Note that the average treatment effect (1.21) is a causal quantity while the associational difference (1.22) is a statistical quantity. The two quantities cannot be equal in general since, association or correlation does not necessarily imply causation. For instance,

- The number of people drowning in swimming pools have high correlation with the number of films Nicolas Cage appears in every year. However, it is foolish to conclude that Nicolas Cage is responsible for all those deaths.
- Similarly, a study found that people who go to bed with their shoes on suffer from a severe headache the next morning. Does it mean that wearing shoes in bed causes headache? However, it turns out that people sleep with their boots on when they are drunk. Also, as they have consumed high alcohol amounts the previous night, it might have led to the headache the next morning. Therefore, drinking at night is the common cause of sleeping in bed with shoes and morning headaches. Such a common cause of attribute and outcome is called **confounding factor**. The example is illustrated in Figure 1.5 [9].

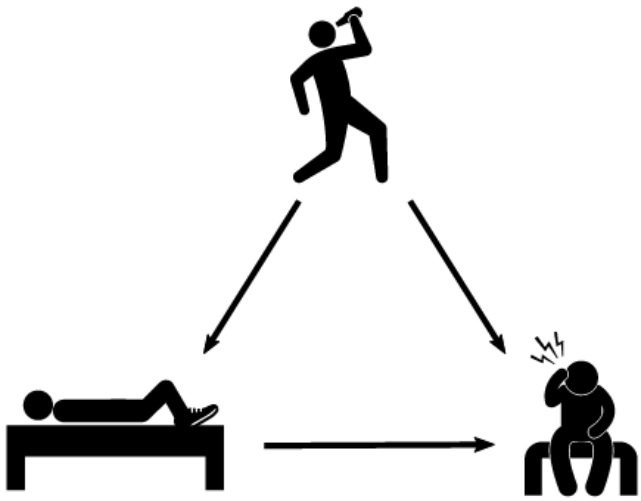

Figure 1.5: Causal structure showing drinking alcohol as the confounding factor of morning headache and sleeping with shoes

However, we can equate 1.21 with 1.22 under certain conditions each of which is explained in following sections.

Ignorability and Exchangeability

The **condition of ignorability** states that the potential outcomes are independent of the treatment assignment. In simpler words, it means that the treatment are assigned completely at random such that the potential outcomes are no longer dependent on the

treatment assignment. Mathematically,

$$(Y(1), Y(0)) \perp\!\!\!\perp T \tag{1.23}$$

This simplifies the average treatment effect (1.20) to as follows:-

$$\mathbf{E}[Y(1)] - \mathbf{E}[Y(0)] = \mathbf{E}[Y(1)|T = 1] - \mathbf{E}[Y(0)|T = 0] \tag{1.24}$$

The condition of ignorability aids in the removal of confounding variables which cause the inability to equate associational difference (statistical quantity) with average treatment effect (causal quantity) as explained earlier.

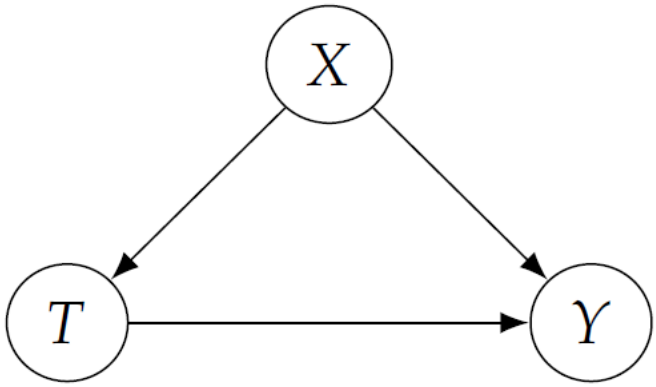

Figure 1.6: Causal structure of X confounding the effect of T on Y

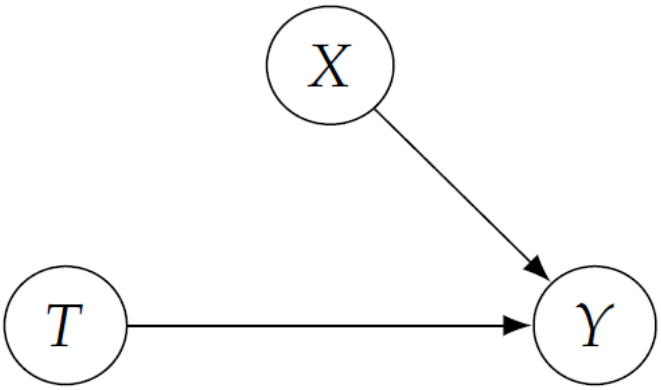

Figure 1.7: Causal structure when ignorability holds. There exist no arrow from X to T, which implies there is no confounding.

The **condition of exchangeability** states that the treatment groups are exchange-

able. In order to understand it, assume there are two treatment groups A and B. The condition of exchangeability implies that if the groups are swapped, the participants in the new group of treatment A will experience the same outcomes as the old treatment A group, and the participants in new treatment B group will experience the same outcomes as the old treatment B group. It means that the two groups which are undergoing treatment are comparable and almost similar to each other in all aspects except the choice of treatment.

Mathematically, the condition of exchangeability implies:

$$\mathbf{E}[Y(1)|T=1] = \mathbf{E}[Y(1)|T=0]$$
$$\mathbf{E}[Y(0)|T=1] = \mathbf{E}[Y(0)|T=0]$$

It implies,

$$\mathbf{E}[Y(1)|T=t] = \mathbf{E}[Y(1)] \tag{1.25}$$
$$\mathbf{E}[Y(0)|T=t] = \mathbf{E}[Y(0)] \tag{1.26}$$

Clearly 1.24, 1.25 and 1.26 implies that the condition of exchangeability and ignorability are mathematically same.

Conditional Exchangeability and Unconfoundedness

It is unreasonable to presume that treatment groups are interchangeable in observational data. In other words, practically there is no reason to anticipate the groups to be identical in all relevant characteristics except treatment. However, if we control for key factors by conditioning, the subgroups may be interchangeable. Let X' be a subset of covariates X conditioned for unconfoundedness, then the conditional exchangeability/ignorability becomes:

$$(Y(1), Y(0)) \perp\!\!\!\perp T|X' \tag{1.27}$$

It establishes that although the treatment and prospective outcomes may be unconditionally related (owing to confounding), they are not associated within levels of X. Conditional exchangeability helps to examine the average treatment effect within levels

of conditioned covariates as follows:

$$\mathbf{E}[Y(1) - Y(0)|X] = \mathbf{E}[Y(1)|X] - \mathbf{E}[Y(0)|X] \tag{1.28}$$

$$= \mathbf{E}[Y(1)|T = 1, X] - \mathbf{E}[Y(0)|T = 0, X] \tag{1.29}$$

$$= \mathbf{E}[Y|T = 1, X] - \mathbf{E}[Y|T = 0, X] \tag{1.30}$$

Linearity of expectation leads to equation 1.28 while equation 1.29 is obtained using conditional exchangeability.

Positivity

In case of binary treatment, the condition of positivity is defined as follows:
For all sets of conditioned covariates, each subgroup has mix population of treated and control participants i.e. all interventions of interest (both treatment option) be seen in every patient subgroup.
Mathematically, it can be written as:

$$0 < P(T = 1|X) < 1 \tag{1.31}$$

No Interference

The notion of no interference indicates that the outcome of every individual is determined only by the therapy provided to that individual and is unaffected by the treatment of someone else. Mathematically,

$$Y_i(t_1, t_2, ..., t_{i-1}, t_i, t_{i+1}, ..., t_n) = Y_i(t_i) \tag{1.32}$$

Consistency

The assumption that the outcome observed Y is the potential outcome under the observed treatment T is known as consistency. Mathematically,

$$T = t \implies Y = Y(t) \tag{1.33}$$

which is equivalent to

$$Y = Y(T) \tag{1.34}$$

1.3.4 Average Treatment Effect Estimation

Theorem 1.1 (Adjustment Formula) If the assumptions of unconfoundedness, positivity, consistency and no interference holds, the average treatment effect is:-

$$\mathbf{E}[Y(1) - Y(0)] = \mathbf{E}_X[\mathbf{E}[Y|T = 1, X] - \mathbf{E}[Y|T = 0, X]] \tag{1.35}$$

Proof of Theorem 1.1.

$$\begin{aligned}
\mathbf{E}[Y(1) - Y(0)] &= \mathbf{E}[Y(1)] - \mathbf{E}[Y(0)] && \text{(linearity of expectation)} \\
&= \mathbf{E}_X[\mathbf{E}[Y(1)|X] - \mathbf{E}[Y(0)|X]] && \text{(law of iterated expectations)} \\
&= \mathbf{E}_X[\mathbf{E}[Y(1)|T = 1, X] - \mathbf{E}[Y(0)|T = 0, X]] && \text{(unconfoundedness and positivity)} \\
&= \mathbf{E}_X[\mathbf{E}[Y|T = 1, X] - \mathbf{E}[Y|T = 0, X]] && \text{(consistency)}
\end{aligned}$$

1.4 Our Contribution

We propose two novel approaches MedCI and MedSCI for the joint estimation of medical cost and time-to-event taking into consideration the treatment effect.

- In MedCI, we extend the CFRNet [10] framework for multi-outcome scenario with one of the outcome being time-to-event. The algorithm is evaluated on a synthetic dataset.
- MedCI is also tested for any category of outcome using a semi-synthetic dataset.
- In MedSCI, we extend the SurvCI [11] framework for multi-outcome scenario with additional hypothesis for medical cost estimation.
- MedSCI is evaluated on a synthetic dataset.

CHAPTER 2

RELATED WORK

2.1 Literature Review

Many authors have developed models for estimating inpatient medical costs using machine learning and deep learning. However, most of these algorithms do not consider the prediction of time to stay, and none talk about treatment effect estimation. Further, extensive algorithms for time-to-event modeling have been developed using survival analysis, and many models for causal inference also exist. However, only a few approaches discuss causal inference with a time-to-event outcome and none of them can be applied in a multiple-outcome scenario (e.g., time-to-event and medical cost as two outcome variables) since they are originally constructed for univariate outcome. Also, a couple of approaches talk about causal inference for multiple outcomes(continuous or binary) but none of them pertains for time-to-event modeling since time as an outcome has certain properties and one needs to take into consideration censoring. To the best of our knowledge, no research work takes into account the evaluation of treatment effect on predicted time-to-stay in hospital along with associated predicted medical costs. We examine related work on medical cost estimation, survival analysis, counterfactual inference, and ultimately introduce the literature that is most relevant to our topic.

2.1.1 Inpatient Medical Cost Prediction

Several researchers leverage machine learning techniques for prediction of medical costs from patients data. Chen et al. [12] used Multiple Linear Regression, Artificial Neural Network(ANN) and Classification and Regression Tree (CART) for the prediction of total hospital expense of patients undergoing breast cancer surgery in Shanghai, China. The dataset consisted of 3699 patients undergoing treatment from 2017 to 2018 which is splitted into train and test in 70:30 ratio. Out of three models, ANN performed best. Muhlestein et al. [13] predicted the total charges involved in transsphenoidal

surgery using 32 machine learning algorithms with the help of 71 predictors and an ensemble model of top preforming algorithms is created. The dataset is taken from NIS database between 2002-2011 with cohort of patients elder than 18 years old . The dataset is splitted with 20% test size and 5 fold cross validation is used. Further, two categories of ensemble models are created: one that takes length of stay as a feature while the other which does not to indicate generalizability since length of stay has a strong influence on the total charges incurred. The total expenses are best predicted by the ensemble model composed of three Gradient Boosted Tree classifiers (root mean square logarithmic error = 0.446). Tong et al. [14] proposed a Bayesian Network fused Locally Weighted Lasso Regression method for prediction of medical cost. In this paper, the dataset consisted of 240 cases of patients in the Department of Enterology in a hospital in Shenyang in March 2016. Data preprocessing included feature selection (removing highly correlated features) that reduced the number of predictors from 26 to 16, followed by missing value imputation using distance based weighted method with the help of clustering in K-Nearest Neighbour. Then the instances (medical records of patients) are fed into a Bayesian Network which categorised them into various treatment schemes based on disease and patient characteristics. Finally, Locally Weighted Lasso Regression method is used to forecast the associated medical costs, achieving highest R^2 value of 0.81 when compared with traditional Linear Regression, Lasso Regression, Neural Network and without fusion Locally Weighted Lasso Regression. Lee et al. [15] developed Artificial Neural Network (ANN) and Classification and Regression Tree (CART) based models to predict the hospital stay charges of colorectal cancer patients in Kyung Hee University Hospital from January 1999 to December 2002. The dataset comprised of 492 patient's 1,022 admission records and 154 predictor variables. Models are developed separately to calculate charges paid by insurance company and charges paid by the patients. With respect to mean absolute error, ANN works better for prediction of medical costs borne by insurance company while CART works better for prediction of medical costs paid by the patient. Taha et al. [16] utilized tree based techniques for prediction of cost associated with patients undergoing colon surgery. Dataset is collected from the Hospital of Wetzikon of 347 patients admitted for colon surgery between January 1, 2013 to December 31, 2019 and splitted into 80:20 ratio for training and testing. Out of Generalized Boosted Regression, Random Forest, and Decision Trees, the minimum mean absolute percentage error (MAPE) is around 21.4% for Ran-

dom Forest. Haddad et al. [17] used machine learning based regression techniques to estimate the medical costs associated with liver transplant patients during their stay as well as their length of stay. The dataset is taken from the Nationwide Inpatient Sample (NIS) database, and it included observations from 2011 (used for training) and 2012 (used for validation) for total 2274 patients. Linear Regression, Principal Components Analysis, Support Vector Machines, Decision Tree, and Nearest Neighbors are the machine learning techniques employed in this study for numeric variables computation such as cost, charges, and duration of stay. The Support Vector Machine with Linear Kernel performed best in forecasting overall cost, with root mean square error (RMSE) value of 0.561, while Principal Component Analysis performed best in predicting length of stay (RMSE=0.743).

2.1.2 Length of Stay Prediction

Elrazek et al. [18] used various machine learning techniques to predict the length of stay of patients admitted in Intensive Care Unit (ICU). The data set is compiled from the medical records of the Suez Canal University Specialized Hospital (SCUSH) ICU from September 2015 to September 2017. The data set included 233 records with 31 descriptive attributes. The output variable (Length of Stay) ranges between 1-142 days and is divided into 3 classes (upto two days, two days - one week, more than one week.) Out of Neural Network, Decision Tree, Tree Bagger, Random Forest, Support Vector Machine, K-Nearest Neighbor, Naive Bayes and Fuzzy Rule models, the Fuzzy logic model performed best with accuracy of 92% evaluated via Bootstrap Aggregation experiments. Rocheteau et al. [19] developed deep learning based temporal pointwise convolution networks for the prediction of length of stay in the intensive care unit. The eICU Database (Pollard et al. [20]) is used, which is a multi-center dataset from 208 hospitals in the United States. The final cohort included 118,534 distinct patients, with 70%, 15%, and 15% of them used for training, validation, and testing, respectively. The model outperformed all the existing approaches with a mean squared error of 21.7. Sheikhalishahi et al. [21] used BiLSTM for the same task on eICU critical care dataset with R^2 value of 0.643.

Clearly, the above mentioned authors treat the medical cost or length of stay prediction as an ordinary regression problem without taking into account any sort of treatment

effect. We now briefly review some popular survival analysis, causal inference and survival analysis-causal inference based techniques in subsequent paragraphs.

2.1.3 Survival Analysis

Various survival analysis techniques can be categorised into two classes depending on whether time is discrete or continuous. Some of the standard survival analysis based models include Cox-Proportional Hazard [8] model and Accelerated Failure Time [[7] models which estimate survival/hazard density of continuous time distribution. A state of the art algorithm, DeepSurv [22] is a deep learning based survival time estimator that utilizes proportional hazard assumption. It is a continuous time survival analysis method. However, the assumption of proportional hazard does not always hold true. Another recent deep learning based development in survival analysis is DeepHit [23], a discrete time survival analysis model which is independent of any underlying assumptions. Deep Survival Machines (DSM) [24] is another cutting-edge model that is not bound by the proportional hazard assumption [8]. It is a continuous-time parametric mixture-model method for time-to-event prediction. DSM is an advanced approach for survival analysis, outperforming benchmarks like DeepHit and DeepSurv.

2.1.4 Causal Inference

Treatment Agnostic Regression Network (TARNet) [10] is a baseline for treatment effect estimation where input covariates are mapped to representation space which is then fed as an input to treatment and control arm for respective prediction. Dragonnet [25] is a propensity score-based adaptive neural network technique for evaluating treatment impact. It is an extension of the TARNet architecture with additional prospensity score estimation head. Counter-factual Regression network (CFRNet) [10] is state-of-the-art deep learning based method for estimation of individual treatment effect in balanced representations. It is also an extension of TARNet [10] with an additional loss term that minimizes the distributional difference between the treatment and control population. Our work is an extension of CFRNet for multi-outcome scenario with one of the outcome as time-to-event.

2.1.5 Causal Inference-Survival Analysis

There exist very few papers that handle counterfactual inference task in a time-to-event setting. Counterfactual survival analysis (CSA) [26] is a continuous time survival analysis technique that learns balanced representations for counterfactual inference in a time-to-event setting using non-parametric adversarial model. SurvITE [27] performs the same task but for discrete intervals of time with the help of a hazard prediction model. Finally, SurvCI [11] is a recent advancement that models the continuous time survival function using primitive mixture of distributions such as log-normal or Weibull. It uses restricted mean survival time to estimate average treatment effect. The algorithm performed at par and addressed the limitations of existing techniques. Our work is an extension of SurvCI framework with additional hypothesis for estimation of medical costs.

2.2 Novelty in Our Work

To the best of our knowledge, there is no single framework that reliably estimates treatment effect for time to stay modelling along with total medical expense computation in that duration. This research work addresses this issue and establishes two counterfactual inference frameworks for joint estimation of time to event and medical cost in balanced representations. The problem is of importance since the number of ICU beds are critical (as highlighted during the pandemic), and it is important for even practitioners to know the personalized duration of stay and which treatment may lead to shorter stay, and lesser cost.

CHAPTER 3

BASELINES

3.1 Counterfactual Regression Network (CFRNet)

Shalit et. al [10] proposed CFRNet framework for estimation of individual treatment effect in balanced representations. We discuss the approach briefly since our work is an extension of this framework and all the assumptions involved hold true for our work as well.

The notations used and assumptions involved are defined as follows:

- Space of covariates $\mathcal{X}$, is a subset of d-dimensional real space $\mathbb{R}^d$ i.e. $\mathcal{X} \subset \mathbb{R}^d$
- The outcome space $\mathcal{Y} \subset \mathbb{R}$
- The treatment a is binary in nature {0,1}
- It is assumed that strong ignorability (1.27) and positivity (1.31) holds.
- The covariates $\mathcal{X}$ are mapped to a representation space $\mathcal{R}$ using the function $\phi : \mathcal{X} \to \mathcal{R}$. It is assumed that ϕ is a twice differentiable, one-to-one function.
- The hypothesis is defined as $h : \mathcal{R} \times \{0, 1\} \to \mathcal{Y}$ while the loss function is $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$

Various definitions involved in the paper are as follows:

- The treatment effect for an instance x is obtained using:-

$$\tau(x) = \mathbf{E}[Y_1 - Y_0|x] \tag{3.1}$$

- The hypothesis is proposed as $f : \mathcal{X} \times \{0, 1\} \to \mathcal{Y}$ such that $f(x, t) = h(\phi(x), a)$.
- The estimated treatment effect of hypothesis f for an instance x is :

$$\hat{\tau}_f(x) = f(x, 1) - f(x, 0) \tag{3.2}$$

- The expectation of square of difference between estimated and actual treatment effect , also called as expected Precision in Estimation of Heterogeneous Effect (PEHE) (Hill et. al [28]) loss is calculated as:

$$\epsilon_{PEHE} = \int_{\mathcal{X}} (\hat{\tau}_f(x) - \tau(x))^2 p(x) dx \tag{3.3}$$

- In order to compute the distance between treatment and control distributions, a probability distribution metric called as Integral Probability Metric (IPM) is used which is defined as :-

$$IPM_G(p,q) = \sup_{g \in G} \left| \int_{\mathcal{S}} g(s)(p(s) - q(s))ds \right| \tag{3.4}$$

where p and q are two probability density functions defined over $\mathcal{S} \subset \mathbb{R}^d$ and G is a family of functions such that $g : \mathcal{S} \to \mathbb{R}$.

The primary objective of CFRNet is to identify a representation $\phi : \mathcal{X} \to \mathcal{R}$ and hypothesis $h : \mathcal{X} \times \{0,1\} \to \mathcal{Y}$ that minimises the PEHE loss ϵ_{PEHE}. In order to achieve this, Shalit et. al [10] utilized a deep learning architecture to simultaneously model $\phi(x)$ and $h(\phi(x), a)$. The covariates $\mathcal{X}$ are transformed to representation space, $\phi(x)$ which then act as an input to the hypothesis layer segmented into two branches, h_1 and h_0 based on whether treatment assigned is 1 or 0. Also, the difference between the treatment and control distribution is minimized using an IPM term. The architecture is represented in Fig. 3.1.

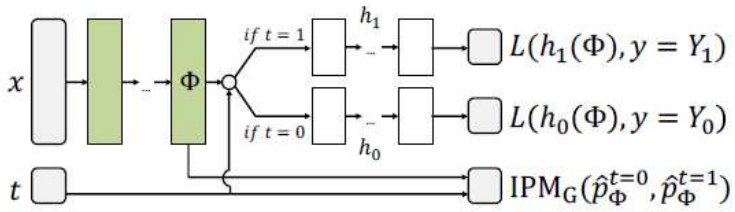

Figure 3.1: Neural Network architecture of CFRNet. L is a loss function, IPM_G is an integral probability metric

The optimal solution is obtained by minimizing loss function as depicted in equation 3.5 using stochastic gradient descent where the error is backpropagated via both hypothesis and representation networks.

$$\mathcal{L} = \frac{\beta}{n} \sum_{i=1}^{n} s_i \cdot L(h(\phi(x_i), a_i), y_i) + \lambda \cdot \mathcal{R}(h) + \gamma \cdot IPM_G(\{\phi(x_i)\}_{i:a_i=0}, \{\phi(x_i)\}_{i:a_i=1}) \tag{3.5}$$

where $s_i = \frac{a_i}{2v} + \frac{1-a_i}{2(1-v)}$, $v = \frac{1}{n}\sum_{i=1}^{n} a_i$ and $\mathcal{R}(.)$ is a model complexity term. Also v refers to the proportion of treated instances ($a_i = 1$), $\{\beta, \lambda, \gamma\}$ are the hyperparameters and $L(.,.)$ is squared error loss. The IPM term utilized Maximum Mean Discrepancy (Sriperumbudur et al.[29]) distance metric.

3.2 SurvCI

Gupta et al. [11] proposed the SurvCI framework, a continuous time survival analysis-causal inference algorithm used to model survival time and to infer treatment effect with that 'time' being the outcome variable. SurvCI utilizes the parametric mixture based generative modelling technique of Deep Survival Machines [24] where the predicted survival distribution is a result of weighted average of K parametric survival distributions. The strategy is described as follows:-

- The covariates, x_i of i^{th} instance are drawn from the distribution.
- The parameters of the model (w, ζ, ξ) are sampled from a normal distribution.
- The latent variable z_i is drawn using the rule: $z_i \sim \text{Discrete}(\sigma_K(\psi_\theta(x_i)^T w_i))$ such that ψ_θ is a multi-layer perceptron with θ as parameters and σ_K is softmax over K mixture distributions.
- $\tilde{\beta}_k$ and $\tilde{\eta}_k$, the mixture component parameters are drawn from two gaussian distributions $\mathcal{N}(\beta_0, \frac{1}{\lambda})$ and $\mathcal{N}(\eta_0, \frac{1}{\lambda})$ respectively for $1 \leq k \leq K$. Note that β_0, η_0, λ are prior parameters.
- Finally, the time to event for the i^{th} instance t_i, is obtained by $t_i \sim R(\beta_k, \eta_k)$, where $\beta_k = \tilde{\beta}_k + \delta(\psi_\theta(x_i)^T \zeta)$ and $\eta_k = \tilde{\eta}_k + \delta(\psi_\theta(x_i)^T \xi)$. Here, $R(.)$ could be Weibull or Log-Normal distribution, and the SELU or Tanh activation functions for the Weibull or Log-Normal distributions are represented by $\delta(.)$.

This results in a model with $\{\theta, \xi, \zeta, \beta, \eta, w\}$ as the learnable parameters. Once the model is trained, the per individual survival density function is estimated and the survival time corresponding to specific treatment (a) is obtained using the concept of Restricted Mean Survival Time (RMST) as follows:-

$$\tilde{y}^a_{T,i} = \int_0^{T*} \tilde{S}^a(y^a_T | x_i) dy^a_T \tag{3.6}$$

In SurvCI, authors have also related survival distribution with individual treatment effect using the same concept of RMST. The proposition is that the individual treatment effect for an i_{th} individual (x_i) is given by:-

$$\tau(x_i) = \int_0^\infty S^1(y|x_i) dy - \int_0^\infty S^0(y|x_i) dy \tag{3.7}$$

The bayesian network and architecture of SurvCI are shown in Figure 3.2 and Figure 3.3 respectively. The light yellow shaded region in the bayesian network represents the

causal relationships among various entities and encloses representation layer $\phi(.)$ while the purple shaded region refers to the parameters used to model survival time.

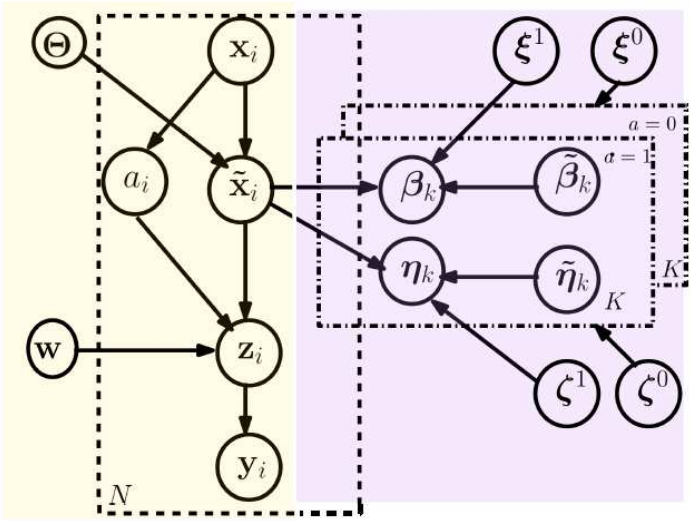

Figure 3.2: Bayesian Network of SurvCI

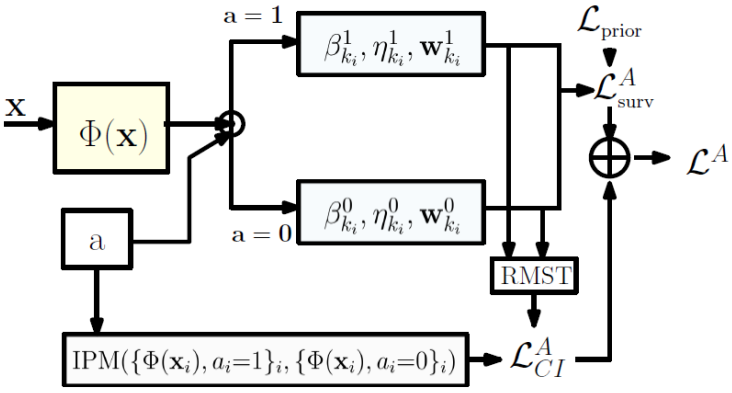

Figure 3.3: SurvCI Architecture

The primary goal in SurvCI is to find the shape (β_k), scale (η_k) and other model parameters $\{\theta, w\}$ such that the overall loss function is minimized. There are two components of overall loss function. One is survival loss function and other being causal inference loss function. The evidence lower bound (ELBO) based loss function for survival outcome corresponding to both treated and control population in random censoring is defined as:

$$\mathcal{L}^A_{surv} = \mathcal{L}^A_U + \mathcal{L}^A_C + \mathcal{L}_{prior} \tag{3.8}$$

where $\mathcal{L}^A_U$ is loss function for uncensored individuals, $\mathcal{L}^A_C$ is loss function for censored

individuals and $\mathcal{L}_{prior}$ is a prior loss term. These are further defined as:

$$\mathcal{L}_U^A = \sum_a \sum_{i=1}^{|D_U|} \sigma_K(ln(f^a(y_i^a|\beta_{k_i}^a, \eta_{k_i}^a))) \tag{3.9}$$

$$\mathcal{L}_C^A = \sum_a \sum_{i=1}^{|D_C|} \sigma_K(ln(S^a(y_i^a|\beta_{k_i}^a, \eta_{k_i}^a))) \tag{3.10}$$

$$\mathcal{L}_{prior} = \lambda \sum_{k=1}^{K} \|\beta_k - \beta_0\|_2^2 + \|\eta_k - \eta_0\|_2^2 \tag{3.11}$$

where $a \in \{0, 1\}$. For counterfactual inference, the loss function used is the same as used in CFRNet [10] and is termed as $\mathcal{L}_{CI}$.

$$\mathcal{L}_{CI} = \frac{\beta'}{N} \sum_{i=1}^{N} v_i L\left(h\left(\Phi\left(x_i\right), a_i\right), y_i^a\right) + \lambda' R(h) + \kappa \,\mathrm{IPM}(p_\Phi^{a_i=1}, p_\Phi^{a_i=0}) \tag{3.12}$$

where $\{\beta', \lambda', \kappa\}$ are hyperparameters, while $v_i = \frac{a_i}{2s} + \frac{1-a_i}{2(1-s)}$, a_i refers to the treatment, $R(.)$ is model complexity term introduced as regularization, $s = \frac{1}{N}\sum_{i=1}^{N} a_i$ and L(.,.) is squared error term defined as:-

$$L\left(h\left(\Phi\left(x_i\right), a_i\right), y_i^a\right) = \left\| h\left(\Phi\left(x_i\right), a_i\right) - y_i^a \right\|_2^2 \tag{3.13}$$

$$= \left\| \int_0^{T^*} \hat{S}^a\left(y_{T,i}^a \mid x_i\right) dy_T - y_{T,i}^a \right\|_2^2 \tag{3.14}$$

. The IPM term uses Maximum Mean Discrepancy distance metric.

The overall loss function is :-

$$\mathcal{L}^A = \rho \mathcal{L}_{surv}^A + \mathcal{L}_{CI} \tag{3.15}$$

Here, ρ is another hyperparameter.

CHAPTER 4

METHODOLOGY

4.1 MedCI

We propose an approach that reliably estimates the treatment effect for two outcome variables: time-to-event and associated medical costs for inpatient stays. Our work is an extension of CFRNet framework with an additional hypothesis layer used to model the associated medical costs. The assumptions of positivity [1.31] and strong ignorability [1.27] in CFRNet holds for our case also. Further, we assume that the censoring is random. Various terminologies and notations used in our work are defined as follows:-

- We define a representation mapping of covariates $\mathcal{X}$ to space $\mathcal{R}$ using one-to-one, twice differentiable function $\phi : \mathcal{X} \to \mathcal{R}$. Also, let $\psi : \mathcal{R} \to \mathcal{X}$ be the inverse function of ϕ, such that $\psi(\phi(x)) = x$ for all $x \in \mathcal{X}$
- The outcome space $\mathcal{Y}, \mathcal{C} \subset \mathbb{R}$ and the treatment a is binary in nature $\{0, 1\}$.
- We define two hypothesis:
  $h_1 : \mathcal{R} \times \{0, 1\} \to \mathcal{Y}$
  $h_2 : \mathcal{R} \times \{0, 1\} \to \mathcal{C}$
  where $h_1(\phi(x))$ is used to estimate time-to-event/censoring while $h_2(\phi(x))$ is used to estimate medical costs.
- The respective loss functions are :
  $L_1 : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$
  $L_2 : \mathcal{C} \times \mathcal{C} \to \mathbb{R}_+$
  where both the losses are root mean squared error (RMSE) loss functions.

Definition 4.1. Let $p^{a=1}(x) := p(x|a = 1)$ and $p^{a=0}(x) := p(x|a = 0)$ be the treatment and control distributions respectively.

Definition 4.2. Let p_ϕ be the distribution induced by ϕ over $\mathcal{R}$. Then $p_\phi^{a=1}(r) := p_\phi(r|a = 1)$ and $p_\phi^{a=0}(r) := p_\phi(r|a = 0)$ are the treatment and control distributions induced over $\mathcal{R}$

Definition 4.3. Given the hypothesis h_1 and h_2, the respective expected loss for an instance and treatment pair (x, a) is:

$$l_{1h,\phi}(x, t) = \int_{\mathcal{Y}} L(Y_t, h_1(\phi(x), a))p(Y_t|x)dY_t \tag{4.1}$$

$$l_{2h,\phi}(x,t) = \int_{\mathcal{C}} L(C_t, h_2(\phi(x),a))p(C_t|x,t)dC_t \tag{4.2}$$

Definition 4.4. ϵ_{F1} and ϵ_{CF1} denote the expected factual and counterfactual losses of h_1 and ϕ, defined as:

$$\epsilon_{F1}(h_1,\phi) = \int_{\mathcal{X}\times\{0,1\}} l_{1h,\phi}(x,a)p(x,a)dxda \tag{4.3}$$

$$\epsilon_{CF1}(h_1,\phi) = \int_{\mathcal{X}\times\{0,1\}} l_{1h,\phi}(x,a)p(x,1-a)dxda \tag{4.4}$$

Definition 4.5. ϵ_{F2} and ϵ_{CF2} denote the expected factual and counterfactual losses of h_2 and ϕ, defined as:

$$\epsilon_{F2}(h_2,\phi) = \int_{\mathcal{X}\times\{0,1\}} l_{2h,\phi}(x,a)p(x,a)dxda \tag{4.5}$$

$$\epsilon_{CF2}(h_2,\phi) = \int_{\mathcal{X}\times\{0,1\}} l_{2h,\phi}(x,a)p(x,1-a)dxda \tag{4.6}$$

Definition 4.6. On similar lines, the expected losses for treatment and control distributions are defined as:

$$\epsilon_{F1}^{a=1}(h_1,\phi) = \int_{\mathcal{X}} l_{1h,\phi}(x,1)p^{a=1}(x)dx \tag{4.7}$$

$$\epsilon_{F1}^{a=0}(h_1,\phi) = \int_{\mathcal{X}} l_{1h,\phi}(x,0)p^{a=0}(x)dx \tag{4.8}$$

$$\epsilon_{F2}^{a=1}(h_2,\phi) = \int_{\mathcal{X}} l_{2h,\phi}(x,1)p^{a=1}(x)dx \tag{4.9}$$

$$\epsilon_{F2}^{a=0}(h_2,\phi) = \int_{\mathcal{X}} l_{2h,\phi}(x,0)p^{a=0}(x)dx \tag{4.10}$$

$$\epsilon_{CF1}^{a=1}(h_1,\phi) = \int_{\mathcal{X}} l_{1h,\phi}(x,1)p^{a=0}(x)dx \tag{4.11}$$

$$\epsilon_{CF1}^{a=0}(h_1,\phi) = \int_{\mathcal{X}} l_{1h,\phi}(x,0)p^{a=1}(x)dx \tag{4.12}$$

$$\epsilon_{CF2}^{a=1}(h_2,\phi) = \int_{\mathcal{X}} l_{2h,\phi}(x,1)p^{a=0}(x)dx \tag{4.13}$$

$$\epsilon_{CF2}^{a=0}(h_2,\phi) = \int_{\mathcal{X}} l_{2h,\phi}(x,0)p^{a=1}(x)dx \tag{4.14}$$

Let $v := p(a=1)$ be the fraction of population that has been treated.

Since $p(x,a) = v \cdot p^{a=1}(x) + (1-v) \cdot p^{a=0}(x)$, we obtain the results as stated in Lemma 4.1 using Definition 4.4 and Definition 4.6.

Lemma 4.1.

$$\epsilon_F(h_1, h_2, \phi) = v \cdot (\epsilon_{F1}^{a=1}(h_1, \phi) + \epsilon_{F2}^{a=1}(h_2, \phi)) + (1 - v) \cdot (\epsilon_{F1}^{a=0}(h_1, \phi) + \epsilon_{F2}^{a=0}(h_2, \phi)) \tag{4.15}$$

$$\epsilon_{CF}(h_1, h_2, \phi) = (1 - v) \cdot (\epsilon_{CF1}^{a=1}(h_1, \phi) + \epsilon_{CF2}^{a=1}(h_2, \phi)) + v \cdot (\epsilon_{CF1}^{a=0}(h_1, \phi) + \epsilon_{CF2}^{a=0}(h_2, \phi)) \tag{4.16}$$

In Lemma 4.1, we postulate that the overall expected treated and control loss is a linear combination of both the respective hypothesis losses.

Lemma 4.2. *Suppose $\phi : \mathcal{X} \to \mathcal{R}$ is an invertible representation with ψ as the inverse function. The distributions $p_\phi^{a=1}$ and $p_\phi^{a=0}$ are as defined in Definition 4.1. Let $v := p(a = 1)$ be the fraction of population that has been treated. Assume G is a family of functions such that $g^{'} : \mathcal{R} \to \mathbb{R}$ and $IPM_G(.,.)$ refers to the integral probability metric induced by G. Let h_1 and h_2 are the two hypothesis as defined earlier. Also, let there exists a constant $B_\phi > 0$ such that for a given treatment, $a = \{0, 1\}$, the function $g^{'}_{\phi,h}(r, a) := \frac{1}{B_\phi} \cdot (l_{1h,\phi}(\psi(r), a) + l_{2h,\phi}(\psi(r), a)) \in G$. We obtain:*

$$\begin{aligned}\epsilon_{CF}(h_1, h_2, \phi) \leq (1 - v) \cdot \{\epsilon_{F1}^{a=1}(h_1, \phi) + \epsilon_{F2}^{a=1}(h_2, \phi)\} + v \cdot \{\epsilon_{F1}^{a=0}(h_1, \phi) + \epsilon_{F2}^{a=0}(h_2, \phi)\} \\ + B_\phi \cdot IPM_G(p_\phi^{a=0}, p_\phi^{a=1})\end{aligned} \tag{4.17}$$

Proof.

$$\epsilon_{CF}(h_1,h_2,\phi) - (1-v)\cdot\{\epsilon_{F1}^{a=1}(h_1,\phi)+\epsilon_{F2}^{a=1}(h_2,\phi)\} - v\cdot\{\epsilon_{F1}^{a=0}(h_1,\phi)+\epsilon_{F2}^{a=0}(h_2,\phi)\}$$

$$= (1-v)\cdot\{\epsilon_{CF1}^{a=1}(h_1,\phi)+\epsilon_{CF2}^{a=1}(h_2,\phi)\} + v\cdot\{\epsilon_{CF1}^{a=0}(h_1,\phi)+\epsilon_{CF2}^{a=0}(h_2,\phi)\}\}$$

$$- (1-v)\cdot\{\epsilon_{F1}^{a=1}(h_1,\phi)+\epsilon_{F2}^{a=1}(h_2,\phi)\} - v\cdot\{\epsilon_{F1}^{a=0}(h_1,\phi)+\epsilon_{F2}^{a=0}(h_2,\phi)\} \tag{4.18}$$

$$= (1-v)\cdot\{\epsilon_{CF1}^{a=1}(h_1,\phi)+\epsilon_{CF2}^{a=1}(h_2,\phi) - \epsilon_{F1}^{a=1}(h_1,\phi) - \epsilon_{F2}^{a=1}(h_2,\phi)\}$$

$$+ v\cdot\{\epsilon_{CF1}^{a=0}(h_1,\phi)+\epsilon_{CF2}^{a=0}(h_2,\phi) - \epsilon_{F1}^{a=0}(h_1,\phi) - \epsilon_{F2}^{a=0}(h_2,\phi)\} \tag{4.19}$$

$$= (1-v)\cdot\{\int_{\mathcal{X}}(l_{1h,\phi}(x,1)\cdot p^{a=0}(x) + l_{2h,\phi}(x,1)\cdot p^{a=0}(x))dx$$

$$- \int_{\mathcal{X}}(l_{1h,\phi}(x,1)\cdot p^{a=1}(x) + l_{2h,\phi}(x,1)\cdot p^{a=1}(x))dx\}$$

$$+ v\cdot\{\int_{\mathcal{X}}(l_{1h,\phi}(x,0)\cdot p^{a=1}(x) + l_{2h,\phi}(x,0)\cdot p^{a=1}(x))dx$$

$$- \int_{\mathcal{X}}(l_{1h,\phi}(x,0)\cdot p^{a=0}(x) + l_{2h,\phi}(x,0)\cdot p^{a=0}(x))dx\} \tag{4.20}$$

$$= (1-v)\cdot\{\int_{\mathcal{X}}(l_{1h,\phi}(x,1) + l_{2h,\phi}(x,1))\cdot(p^{a=0}(x) - p^{a=1}(x))dx\}$$

$$+ v\cdot\{\int_{\mathcal{X}}(l_{1h,\phi}(x,0) + l_{2h,\phi}(x,0))\cdot(p^{a=1}(x) - p^{a=0}(x))dx\} \tag{4.21}$$

$$= B_\phi\cdot(1-v)\{\int_{\mathcal{R}}\frac{1}{B_\phi}\cdot(l_{1h,\phi}\cdot(\psi(r),1) + l_{2h,\phi}(\psi(r),1))\cdot(p_\phi^{a=0}(r) - p_\phi^{a=1}(r))dr\}$$

$$+ B_\phi\cdot v\{\int_{\mathcal{R}}\frac{1}{B_\phi}\cdot(l_{1h,\phi}(\psi(r),0) + l_{2h,\phi}(\psi(r),0))\cdot(p_\phi^{a=1}(r) - p_\phi^{a=0}(r))dr\} \tag{4.22}$$

$$\leq B_\phi\cdot(1-v)\sup_{g'\in\mathcal{G}}\left|\int_{\mathcal{R}} g'(r)\cdot(p_\phi^{a=0}(r) - p_\phi^{a=1}(r))dr\right|$$

$$+ B_\phi\cdot v\sup_{g'\in\mathcal{G}}\left|\int_{\mathcal{R}} g'(r)\cdot(p_\phi^{a=1}(r) - p_\phi^{a=0}(r))dr\right| \tag{4.23}$$

$$\leq B_\phi\cdot IPM_G(p_\phi^{a=0}, p_\phi^{a=1}) \tag{4.24}$$

□

Here, equality (4.18) is as per Equation (4.16) of Lemma 4.1, while equality (4.20) is by Definition 4.6 of the expected losses. Further, equality (4.22) is the change of variables and inequalities (4.23) & (4.24) are by the definition of function g and term IPM_G respectively.

Definition 4.7. For a=0,1, we define:

$m_a(x) = \mathbf{E}[Y_a|x]$

$n_a(x) = \mathbf{E}[C_a|x]$

The treatment effect can be reformulated as:
$\tau(x) = m_1(x) + n_1(x) - m_0(x) - n_0(x)$

Recall that $f_1 : \mathcal{X} \times \{0,1\} \to \mathcal{Y}$ and $f_2 : \mathcal{X} \times \{0,1\} \to \mathcal{C}$ are the hypotheses such that $f_1(x,a) = h_1(\phi(x),a)$ and $f_2(x,a) = h_2(\phi(x),a)$ for a representation ϕ and hypotheses h_1, h_2 defined over the output of ϕ.

Definition 4.8. The treatment effect is estimated using:
$\hat{\tau} = f_1(x,1) + f_2(x,1) - f_1(x,0) - f_2(x,0)$

Definition 4.9. The expectation of square of difference between estimated and actual treatment effect, also called as expected Precision in Estimation of Heterogeneous Effect (PEHE) loss is calculated as:
$\epsilon_{PEHE}(f) = \int_{\mathcal{X}} (\hat{\tau}_f(x) - \tau(x))^2 p(x) dx$

Definition 4.10. The expected variance of Y_a and C_a with respect to distribution $p(x;a)$:
$\sigma^2_{Y_a}(p(x,a)) = \int_{\mathcal{X}\times\mathcal{Y}} (Y_a - m_a(x))^2 p(Y_a|x) p(x,a) dY_a dx$
$\sigma^2_{C_a}(p(x,a)) = \int_{\mathcal{X}\times\mathcal{C}} (C_a - n_a(x))^2 p(C_a|x) p(x,a) dC_a dx$

Lemma 4.3. *Given two functions, $f_1 : \mathcal{X} \times \{0,1\} \to \mathcal{Y}$ and $f_2 : \mathcal{X} \times \{0,1\} \to \mathcal{C}$, and distribution p(x,a) defined over $\mathcal{X} \times \{0,1\}$:*

$$\int_{\mathcal{X}\times\{0,1\}} (f_1(x,a) - m_a(x))^2 p(x,a) dx da + \int_{\mathcal{X}\times\{0,1\}} (f_2(x,a) - n_a(x))^2 p(x,a) dx da$$
$$= \epsilon_F - \sigma^2_{Y_a}(p(x,a)) - \sigma^2_{C_a}(p(x,a)) \tag{4.25}$$

$$\int_{\mathcal{X}\times\{0,1\}} (f_1(x,a) - m_a(x))^2 p(x,1-a) dx da + \int_{\mathcal{X}\times\{0,1\}} (f_2(x,a) - n_a(x))^2 p(x,1-a) dx da$$
$$= \epsilon_{CF} - \sigma^2_{Y_a}(p(x,1-a)) - \sigma^2_{C_a}(p(x,1-a)) \tag{4.26}$$

We have proved it for ϵ_F while the proof is identical for ϵ_{CF}.

Proof.

$$
\begin{aligned}
\epsilon_F &= \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (f_1(x,a) - Y_a)^2 p(Y_a|x)p(x,a)dY_a dxda \\
&+ \int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} (f_2(x,a) - C_a)^2 p(C_a|x)p(x,a)dC_a dxda \qquad (4.27) \\
&= \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (f_1(x,a) - m_a(x))^2 p(Y_a|x)p(x,a)dY_a dxda \\
&+ \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (m_a(x) - Y_a)^2 p(Y_a|x)p(x,a)dY_a dxda \\
&+ 2\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (f_1(x,a) - m_a(x)) \cdot (m_a(x) - Y_a) p(Y_a|x)p(x,a)dY_a dxda \\
&+ \int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} (f_2(x,a) - n_a(x))^2 p(C_a|x)p(x,a)dC_a dxda \\
&+ \int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} (n_a(x) - C_a)^2 p(C_a|x)p(x,a)dC_a dxda \\
&+ 2\int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} (f_2(x,a) - n_a(x)) \cdot (n_a(x) - C_a) p(C_a|x)p(x,a)dC_a dxda \qquad (4.28) \\
&= \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (f_1(x,a) - m_a(x))^2 p(Y_a|x)p(x,a)dY_a dxda \\
&+ \int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} (f_2(x,a) - n_a(x))^2 p(C_a|x)p(x,a)dC_a dxda \\
&+ \sigma^2_{Y_a}(p(x,a)) + \sigma^2_{C_a}(p(x,a)) \qquad (4.29)
\end{aligned}
$$

□

Equality (4.27) is by the definition of ϵ_F while equality (4.28) is simple mathematical manipulation. Equality (4.29) is due to Definition 4.10 and also because two integral terms approach zero since $m_a(x) = \int_{\mathcal{X}} Y_a\, p(Y_a|x)dx$ and $n_a(x) = \int_{\mathcal{X}} C_a\, p(C_a|x)dx$.

Theorem 4.1. *Assume $\phi : \mathcal{X} \to \mathcal{R}$ is a one-to-one representation function, and Ψ is its inverse. Further, assume that $p_\phi^{a=0}, p_\phi^{a=1}$ are defined as in Definition 4.2. Suppose $v = p(a = 1)$. Let $\mathcal{G}$ be a family of functions $g : \mathcal{R} \to \mathbb{R}$, and $IPM_G(\cdot,\cdot)$ indicate the integral probability metric induced by $\mathcal{G}$. Consider $h_1 : \mathcal{X} \times \{0,1\} \to \mathcal{Y}$, $h_2 : \mathcal{X} \times \{0,1\} \to \mathcal{C}$ to be two hypotheses. Let the loss $L(y_1, y_2) = (y_1 - y_2)^2$. Assume there exists a constant $B_\Phi > 0$, such that the functions $g_{\Phi,h}(r,a) := \frac{1}{B_\Phi} \cdot (l_{1h,\phi}(\psi(r), a) +$*

$l_{2h,\phi}(\psi(r), a)) \in \mathcal{G}$ *for* $a \in \{0, 1\}$. *We then have:*

$$\epsilon_{PEHE} \leq 4 \cdot \{\epsilon_F(h_1, h_2, \phi) + \epsilon_{CF}(h_1, h_2, \phi) - 2 \cdot (\sigma_Y^2 + \sigma_C^2)\} \tag{4.30}$$

$$\leq 4 \cdot \{\epsilon_{F1}^{a=1}(h_1, \phi) + \epsilon_{F2}^{a=1}(h_2, \phi) + \epsilon_{F1}^{a=0}(h_1, \phi) + \epsilon_{F2}^{a=0}(h_2, \phi)$$
$$+ B_\phi \cdot IPM_G(p_\phi^{a=0}, p_\phi^{a=1}) - 2 \cdot (\sigma_Y^2 + \sigma_C^2)\} \tag{4.31}$$

Proof.

$$\epsilon_{PEHE} = \int_{\mathcal{X}} (\hat{\tau}(x) - \tau(x))^2 p(x) dx \tag{4.32}$$

$$= \int_{\mathcal{X}} \{(f_1(x, 1) + f_2(x, 1) - f_1(x, 0) - f_2(x, 0))$$
$$- (m_1(x) + n_1(x) - m_0(x) - n_0(x))\}^2 p(x) dx \tag{4.33}$$

$$= \int_{\mathcal{X}} \{(f_1(x, 1) - m_1(x)) + (f_2(x, 1) - n_1(x))$$
$$+ (m_0(x) - f_1(x, 0)) + (n_0(x) - f_2(x, 0))\}^2 p(x) dx \tag{4.34}$$

$$\leq 2 \cdot \int_{\mathcal{X}} \{(f_1(x, 1) - m_1(x) + f_2(x, 1) - n_1(x))^2$$
$$+ (m_0(x) - f_1(x, 0) + n_0(x) - f_2(x, 0))^2\} p(x) dx \tag{4.35}$$

$$= 2 \cdot \int_{\mathcal{X}} (f_1(x, 1) - m_1(x) + f_2(x, 1) - n_1(x))^2 p(x, a = 1) dx$$
$$+ 2 \cdot \int_{\mathcal{X}} (m_0(x) - f_1(x, 0) + n_0(x) - f_2(x, 0))^2 p(x, a = 0) dx$$
$$+ 2 \cdot \int_{\mathcal{X}} (f_1(x, 1) - m_1(x) + f_2(x, 1) - n_1(x))^2 p(x, a = 0) dx$$
$$+ 2 \cdot \int_{\mathcal{X}} (m_0(x) - f_1(x, 0) + n_0(x) - f_2(x, 0))^2 p(x, a = 1) dx \tag{4.36}$$

$$= 2 \cdot \int_{\mathcal{X} \times \{0,1\}} (f_1(x, a) - m_a(x) + f_2(x, a) - n_a(x))^2 p(x, a) dx da$$
$$+ 2 \cdot \int_{\mathcal{X} \times \{0,1\}} (f_1(x, a) - m_a(x) + f_2(x, a) - n_a(x))^2 p(x, 1 - a) dx da \tag{4.37}$$

$$\leq 4 \cdot \{\int_{\mathcal{X} \times \{0,1\}} (f_1(x, a) - m_a(x))^2 p(x, a) dx da$$
$$+ \int_{\mathcal{X} \times \{0,1\}} (f_2(x, a) - n_a(x))^2 p(x, a) dx da\}$$
$$+ 4 \cdot \{\int_{\mathcal{X} \times \{0,1\}} (f_1(x, a) - m_a(x))^2 p(x, 1 - a) dx da$$
$$+ \int_{\mathcal{X} \times \{0,1\}} (f_2(x, a) - n_a(x))^2 p(x, 1 - a) dx da\} \tag{4.38}$$

$$\leq 4 \cdot (\epsilon_F(h_1, h_2, \phi) + \epsilon_{CF}(h_1, h_2, \phi) - 2 \cdot (\sigma_Y^2 + \sigma_C^2)) \tag{4.39}$$

□

Inequality (4.35) and (4.38) are arrived at using the mathematical identity $(a+b)^2 \leq 2 \cdot (a^2 + b^2)$. Rest of the proof is self explanatory in nature based on the definitions and lemmas provided so far. The second inequality in the theorem can be proved by combining Lemma 4.1 and Lemma 4.2.

The architecture of proposed MedCI method includes a representation layer and two hypothesis layers (each for medical cost and outcome) with respect to two treatment arms as shown in Figure 4.1.

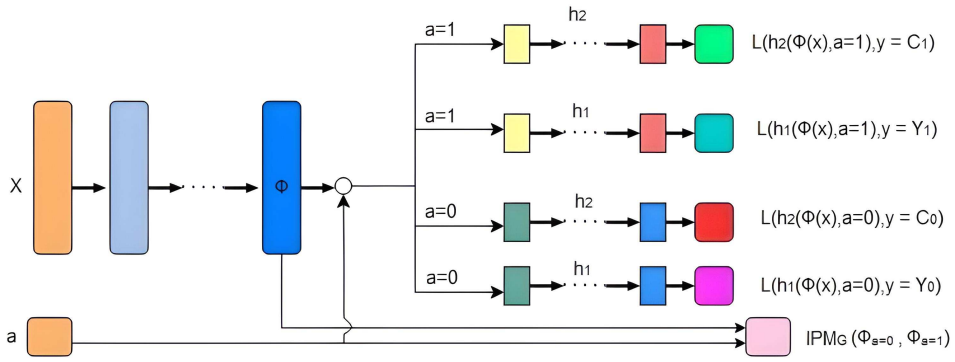

Figure 4.1: MedCI Neural Network Architecture

The optimal solution is obtained by minimizing loss function as depicted in equation 4.40 using Adam optimizer where the error is backpropagated via both hypotheses layers and representation networks.

$$\mathcal{L} = \frac{\beta}{n} \sum_{i=1}^{n} s_i \cdot (L(h_1(\phi(x_i), a_i), y_i) + L(h_2(\phi(x_i), a_i), c_i)) + \lambda \cdot \mathcal{R}(h) + \gamma \cdot IPM_G(\{\phi(x_i)\}_{i:a_i=0}, \{\phi(x_i)\}_{i:a_i=1}) \tag{4.40}$$

where $s_i = \frac{a_i}{2v} + \frac{1-a_i}{2(1-v)}$, $v = \frac{1}{n}\sum_{i=1}^{n} a_i$ and $\mathcal{R}(.)$ is a model complexity term. Also v refers to the proportion of treated instances ($a_i = 1$), $\{\beta, \lambda, \gamma\}$ are the hyperparameters and $L(.,.)$ is squared error loss. The IPM term utilized Maximum Mean Discrepancy (Sriperumbudur et al., 2012 [29]) distance metric.

The algorithm is parameterized using multiple fully connected deep neural networks. At each epoch only one treatment arm is switched on, and hence, the model

trains only that respective arm of both the hypotheses in that epoch. The time-to-event and medical costs for a patient are predicted in both current and counterfactual scenario and then various metrics such as average treatment effect, root mean squared error, PEHE etc. are estimated to evaluate the performance.

4.2 MedSCI

We propose MedSCI, an extension of SurvCI [11] approach by adding an additional hypothesis layer for the estimation of medical costs. The approach for predicting time-to-event is taken from SurvCI straightaway. The same methodology is used for estimating survival curve and then estimating survival time from it using RMST. However, SurvCI is designed only for single outcome problem(time-to-event). We extend it for multi-outcome scenario. We propose, for medical cost estimation, a deep neural network on top of existing architecture with the representation layer output as the input to the network. Mathematically, we define the hypothesis $\tilde{h} : \mathcal{R} \times \{0, 1\} \rightarrow \mathcal{C}$ such that $\tilde{h}(\phi(x), a)$ is used to model the associated medical costs based on patient data. The architecture of proposed MedSCI is shown in Figure 4.2.

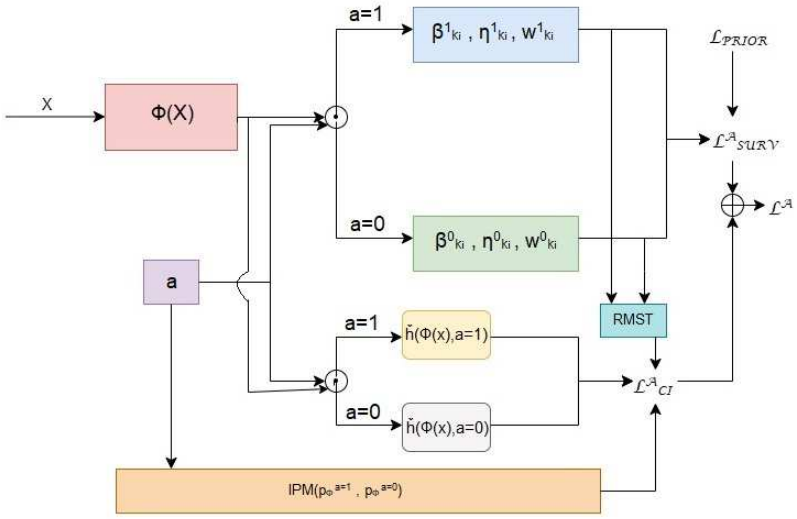

Figure 4.2: MedSCI Architecture

The survival loss remains same but there is a change in the counterfactual inference

loss term. The losses are defined as follows:-

$$\mathcal{L}^A_{surv} = \mathcal{L}^A_U + \mathcal{L}^A_C + \mathcal{L}_{prior} \tag{4.41}$$

where $\mathcal{L}^A_U$ is loss function for uncensored individuals, $\mathcal{L}^A_C$ is loss function for censored individuals and $\mathcal{L}_{prior}$ is a prior loss term. These are further defined as:

$$\mathcal{L}^A_U = \sum_a \sum_{i=1}^{|D_U|} \sigma_K(ln(f^a(y^a_i|\beta^a_{k_i}, \eta^a_{k_i}))) \tag{4.42}$$

$$\mathcal{L}^A_C = \sum_a \sum_{i=1}^{|D_C|} \sigma_K(ln(S^a(y^a_i|\beta^a_{k_i}, \eta^a_{k_i}))) \tag{4.43}$$

$$\mathcal{L}_{prior} = \lambda \sum_{k=1}^{K} \|\beta_k - \beta_0\|_2^2 + \|\eta_k - \eta_0\|_2^2 \tag{4.44}$$

where $a \in \{0, 1\}$, $|D_U|$ refers to total no. of individuals who experience the event, $|D_C|$ refers to total no. of censored individuals, f^a is the failure density function for treatment a, S^a is the survival density function for treatment a, σ_K is softmax activation over K distributions, β_k and η_k are the shape and scale parameters respectively while β_0, η_0 and λ are the prior parameters. The counterfactual inference loss term is modified as follows:

$$\mathcal{L}_{CI} = \frac{\beta'}{N} \sum_{i=1}^{N} v_i (L\left(h\left(\Phi\left(x_i\right), a_i\right), y^a_i\right) + L\left(\tilde{h}\left(\Phi\left(x_i\right), a_i\right), C^a_i\right)) + \lambda' R(h) + \kappa \, \mathrm{IPM}(p^{a_i=1}_\Phi, p^{a_i=0}_\Phi) \tag{4.45}$$

where $\{\beta', \lambda', \kappa\}$ are hyperparameters, while $v_i = \frac{a_i}{2s} + \frac{1-a_i}{2(1-s)}$, a_i refers to the treatment, $R(.)$ is model complexity term introduced as regularization, $s = \frac{1}{N}\sum_{i=1}^{N} a_i$ and L(., .) is squared error term defined as:-

$$L\left(h\left(\Phi\left(x_i\right), a_i\right), y^a_i\right) = \left\| h\left(\Phi\left(x_i\right), a_i\right) - y^a_i \right\|_2^2 \tag{4.46}$$

$$= \left\| \int_0^{T^*} \hat{S}^a\left(y^a_{T,i} \mid x_i\right) dy_T - y^a_{T,i} \right\|_2^2 \tag{4.47}$$

$$L\left(\tilde{h}\left(\Phi\left(x_i\right), a_i\right), C^a_i\right) = \left\| \tilde{h}\left(\Phi\left(x_i\right), a_i\right) - C^a_i \right\|_2^2 \tag{4.48}$$

The IPM term uses Maximum Mean Discrepancy distance metric. The overall loss function is :-

$$\mathcal{L}^A = \rho \mathcal{L}^A_{surv} + \mathcal{L}_{CI} \tag{4.49}$$

Here, ρ is another hyperparameter.

The objective is to find the best parameters for both time-to-event as well as medical cost prediction networks such that the overall loss is minimized. Note that in MedSCI, a nested loop is created where the medical cost estimation network trains inside the survival network. Once the training is complete, the algorithm is evaluated using multiple metrics such as RMSE, Concordance Index, PEHE etc.

CHAPTER 5

EXPERIMENTATION & RESULTS

5.1 Dataset

We have used two datasets for simulation of experiments: a semi-synthetic dataset and a synthetic dataset. Note that in case of semi-synthetic dataset, apart from medical cost, the other outcome is not time-to-event. Rather, we have used this dataset to show the generalizability of our MedCI approach i.e. the algorithm works well for any kind of outcome variable. Nevertheless, we have tested both of our approach on synthetic dataset which has time-to-event and medical costs as two outcome variables.

5.1.1 Semi-Synthetic Dataset

The dataset is constructed by adapting the strategy of the Atlantic Causal Inference Conference (ACIC) Data Analysis Challenge 2017 (Hahn et al. [30]) where the outcome variable and treatment is generated via a data generating process (DGP) with covariates taken from Infant Health and Development Program, or IHDP (Gross et al. [31]). The outcome variable in the DGP is synthesized depending on four different types of errors:-

1. additive, independent and identically distributed (IID)
2. additive, group correlated
3. additive, heteroskedastic
4. non-additive, independent and identically distributed (IID)

Further the DGP has "high" or "low" settings depending on:

- magnitude of the causal effect, ξ (which can take value either 0.33 or 2)
- strength of the confounding, κ (which can take value either (0.5,0) or (-1,3))
- noise level in the response variable, η (which can take value either 0.25 or 1.25)

In the subsequent section we explain the generation process of outcome and cost variable. It should be noted that in the DGP, only 8 features are utilized among a total of 58 from the IHDP data.

Outcome Synthesis

The outcome synthesis is done using the DGP as follows:-

$$f(x) = x_1 + x_{43} + 0.3(x_{10} - 1) \tag{5.1}$$

$$\pi(x) = Pr(Z_i = 1) \tag{5.2}$$

$$= \frac{1}{1 + exp(\kappa_1 f(x) + \kappa_2)} \tag{5.3}$$

$$\mu(x) = -\sin(\phi(\pi(x))) + x_{43} \tag{5.4}$$

$$\tau(x) = \xi(x_3 x_{24} + (x_{14} - 1) - (x_{15} - 1)) \tag{5.5}$$

$$\sigma(x) = 0.4 + \frac{x_{21} - 1}{15} \tag{5.6}$$

where x_i is the i^{th} covariate in the IHDP data, and $\phi(.)$ refers to a normal random variable's cumulative distribution function. It is assumed that the errors(ϵ), have an independent, identical standard normal distribution. Let

$$\sigma_y = \eta \sqrt{Var(\mu(x) + \pi(x)\tau(x))}$$

where variance is taken over observed samples. The outcome variable is then computed using:-

$$Y_i = \mu(x_i) + \tau(x_i) Z_i + \sigma_y \epsilon_i \tag{5.7}$$

where Z_i is the treatment value corresponding to i^{th} instance. We have used the settings such that $\kappa_1 = 3$, $\kappa_2 = -1$, $\eta = 0.25$ and $\xi = 2$. For counterfactual outcome generation, the same process is followed except that in equation 5.7, counterfactual treatment value is used such that the outcome generation equation becomes:-

$$Y_i^{cf} = \mu(x_i) + \tau(x_i)(1 - Z_i) + \sigma_y \epsilon_i \tag{5.8}$$

The range of values for outcome generated by equation 5.7 is {min: -6.138, max: 81.854} while for counterfactual outcome generated by equation 5.8, it is {min:-6.43,

max: 87.171}.

Medical Cost Synthesis

For medical cost synthesis, the DGP is slightly modified as follows:-

$$f(x) = x_1 + x_{43} + 0.3(x_{10} - 1) \tag{5.9}$$

$$\pi(x) = Pr(Z_i = 1) \tag{5.10}$$

$$= \frac{1}{1 + exp(\kappa_1 f(x) + \kappa_2)} \tag{5.11}$$

$$\mu(x) = |\log(\pi(x))| + x_{43} \tag{5.12}$$

$$\tau(x) = \alpha - \xi(x_3 x_{24} + |(x_{14} - 1) - (x_{15} - 1)|) \tag{5.13}$$

$$\sigma(x) = 0.4 + \frac{x_{21} - 1}{15} \tag{5.14}$$

where x_i is the i^{th} covariate in the IHDP data and α is a parameter to regulate the overlapping of control and treatment cost distributions. Here, the value of α is kept equal to 400. It is assumed that the errors(ϵ) in medical cost have an independent, identical exponential distribution.

Then the factual cost and counterfactual cost is generated using equation 5.7 and equation 5.8 respectively with the identical parametric settings as in outcome synthesis.

The range of values for generated cost is {min: 74.74, max: 545.80} while for counterfactual cost it is {min: 50.18, max: 637.41}. The histogram plots of factual and counterfactual cost with respect to treatment and control population are depicted in Fig. 5.1 and Fig. 5.2.

The final dataset consists of 4302 instances, 58 IHDP attributes, treatment column, factual and counterfactual outcome column as well as factual and counterfactual medical cost column.

5.1.2 Synthetic Dataset

The process of generating this dataset is a mixture of synthetic simulation experiment used in SurvITE [27] and the generation process of medical cost in our semi-synthetic

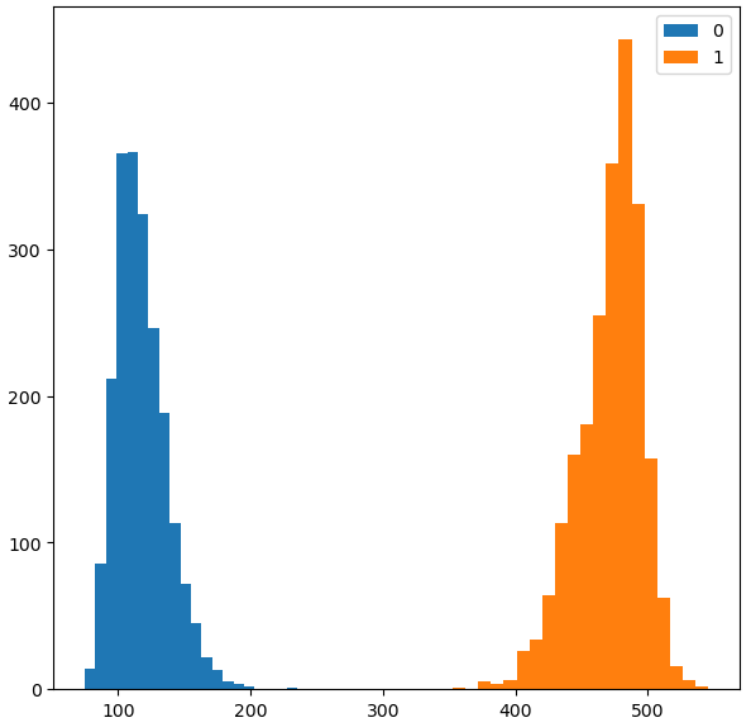

Figure 5.1: Histogram plot of Factual Cost in Semi-Synthetic Dataset (Treatment=1, Control=0)

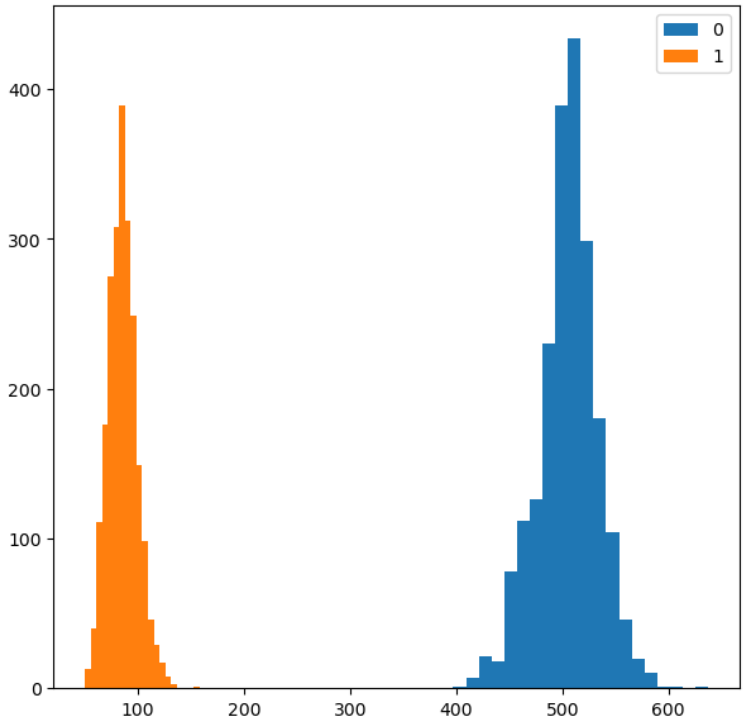

Figure 5.2: Histogram plot of Counterfactual Cost in Semi-Synthetic Dataset (Treatment=1, Control=0)

dataset. The covariates generated from a 10-dimensional multivariate normal distribution are directly taken from SurvITE dataset. The treatment values are sampled from bernoulli distribution: a $\sim Bern(\xi \cdot \sigma(\sum_{p \in \mathcal{P}} x_p))$, where σ is sigmoid function, ξ (selection strength) is taken as 0.33 while $\mathcal{P} = \{0, 1\}$.

Time-to-event Synthesis

Hazard function for event time is formulated as follows:

$$\lambda^a(t|x) = \left\{ \begin{array}{ll} 0.1\sigma(-5x_1^2 - a \cdot (\mathbb{1}\{x_3 \geq 0\} + 0.5)), & \text{for } (t \leq 10) \\ 0.1\sigma(10x_2 - a \cdot (\mathbb{1}\{x_3 \geq 0\} + 0.5)), & \text{for } (t > 10) \end{array} \right\} \tag{5.15}$$

where $\mathbb{1}$ is an indicator function that checks the inside condition and returns 1 if it is true, else returns 0.
The hazard function for censoring time is defined as follows:-

$$\lambda_C(t|x) = 0.01\sigma(10x_4^2) \tag{5.16}$$

The time $t = 30$ is marked as the end of study i.e. $\lambda_C(30|x) = 1$. Once, the hazard function is known, the survival function is estimated using the equation 5.17.

$$S(\tau|a, x) = \prod_{t \leq \tau} (1 - \lambda(t|x)) \tag{5.17}$$

The area under the survival curve up to a specific time point, also called as restricted mean survival time (RMST) represents the factual event time or censoring time depending on the whether the survival function is derived from event hazard function or censoring hazard. Then, the time-to-event for an instance is estimated as minimum of (event time,censoring time) as depicted in equation 1.13.

Identical process is followed for the generation of counterfactual time-to-event except that in this case, the treatment values used in the DGP are the counterfactual treatment values.

The range of values for generated factual and counterfactual time is {min: 11.84, max: 26.42}.

Medical Cost Synthesis

The cost generation procedure used in semi-synthetic dataset is modified for the synthesis of medical cost here as follows:

$$f(x) = x_1 + x_2 + 0.3(x_3 - 1) \tag{5.18}$$

$$\pi(x) = Pr(Z_i = 1) \tag{5.19}$$

$$= \frac{1}{1 + exp(\kappa_1 f(x) + \kappa_2)} \tag{5.20}$$

$$\mu(x) = |\log(\pi(x))| + |x_2| \tag{5.21}$$

$$\tau(x) = \alpha - \xi(|x_3 x_6| + |(x_8 - 1) - (x_9 - 1)|) \tag{5.22}$$

$$\sigma(x) = 0.4 + \frac{|x_{10} - 1|}{15} \tag{5.23}$$

where x_i is the i^{th} covariate in the data and α is a parameter to regulate the overlapping of control and treatment cost distributions. Here, the value of α is kept equal to 1000. It is assumed that the errors(ϵ) in medical cost have an independent, identical exponential distribution.

Then the factual cost and counterfactual cost is generated using equation 5.7 and equation 5.8 respectively with the identical parametric settings as that in semi-synthetic DGP.

The range of values for generated cost is {min: 0.280, max: 1931.77} while for counterfactual cost it is {min: 0.168, max: 2043.76}. The histogram plots of factual and counterfactual cost with respect to treatment and control population are depicted in Fig. 5.3 and Fig. 5.4.

The final dataset consists of 10000 instances, 10 covariates, treatment column, factual and counterfactual time-to-event/censoring column as well as factual and counterfactual medical cost column.

5.2 ACIC Semi-Synthetic Experiments

A range of experiments were carried out on semi-synthetic dataset using the proposed MedCI method. The values of factual outcome is in the range [-6.13,81.85] while for

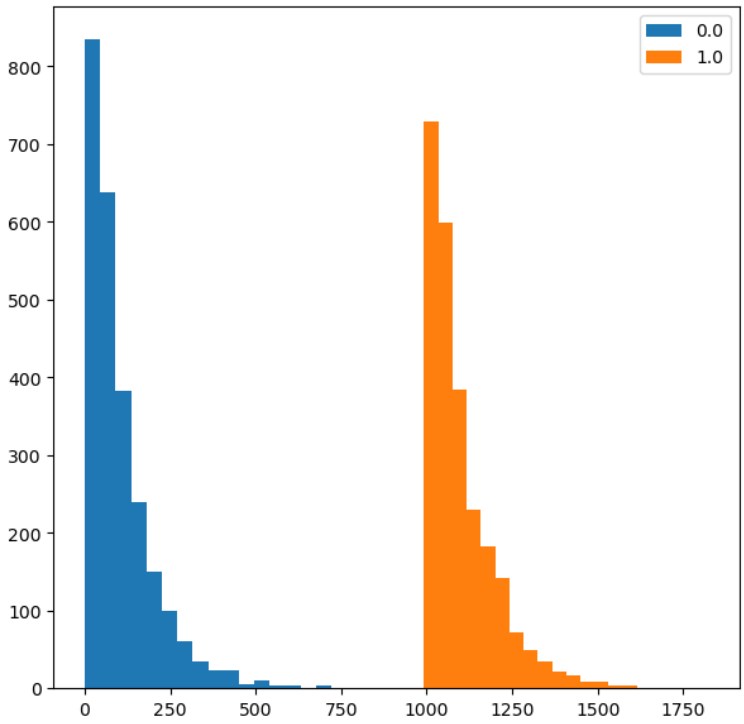

Figure 5.3: Histogram plot of Factual Cost in Synthetic Dataset (Treatment=1, Control=0)

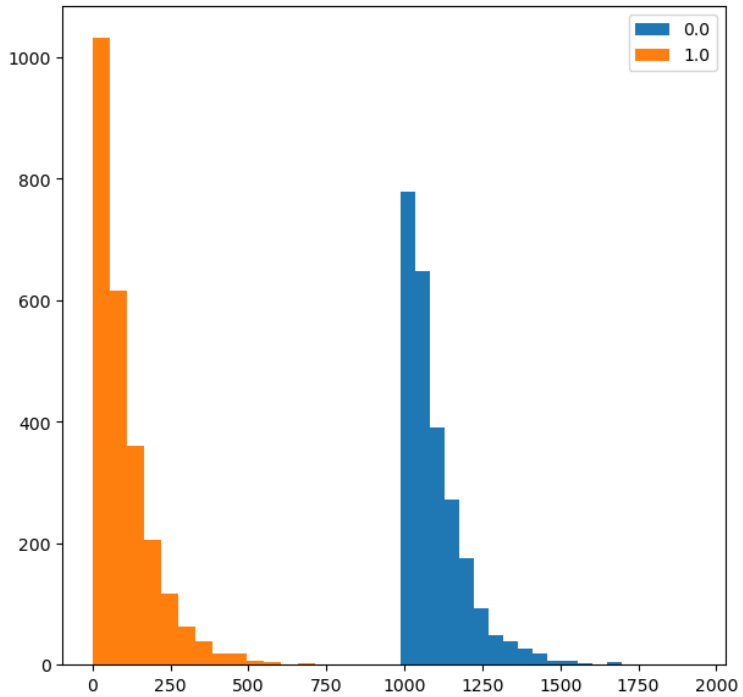

Figure 5.4: Histogram plot of Counterfactual Cost in Synthetic Dataset (Treatment=1, Control=0)

counterfactual outcome, it is [-6.43,87.17]. On the other hand the range of values for factual medical cost is [74.74,545.80] while for counterfactual cost, it is [50.18,637.41].

The representation network is multi-layer perceptron with 3 hidden layers, 48 hidden nodes per layer, and 48 nodes in the output layer. The dropout rate is chosen as 0.145 with Rectified Linear Unit (ReLU) activation function for hidden layers. Further, the neural network architecture for outcome hypothesis as well as medical cost hypothesis also consists of multilayer perceptron network, with 3 hidden layers and 32 nodes per hidden layer. The dropout rate is again chosen to be 0.145 and ReLU is chosen as activation function for hidden layers. Identical network is defined for both the control and treatment arms in both the categories. Adam optimizer is used with weight decay of 0.5. The step size for learning rate decay is chosen as 100 while the multiplicative factor for learning rate decay is selected as 0.97. The scatter plot of input data and representation network transformed data across two dimensions obtained by t-SNE is shown in Figure 5.5. The Maximum Mean Discrepancy value of zero signifies that the control and treated group are comparable enough in the representation layer output data and there is no treatment bias.

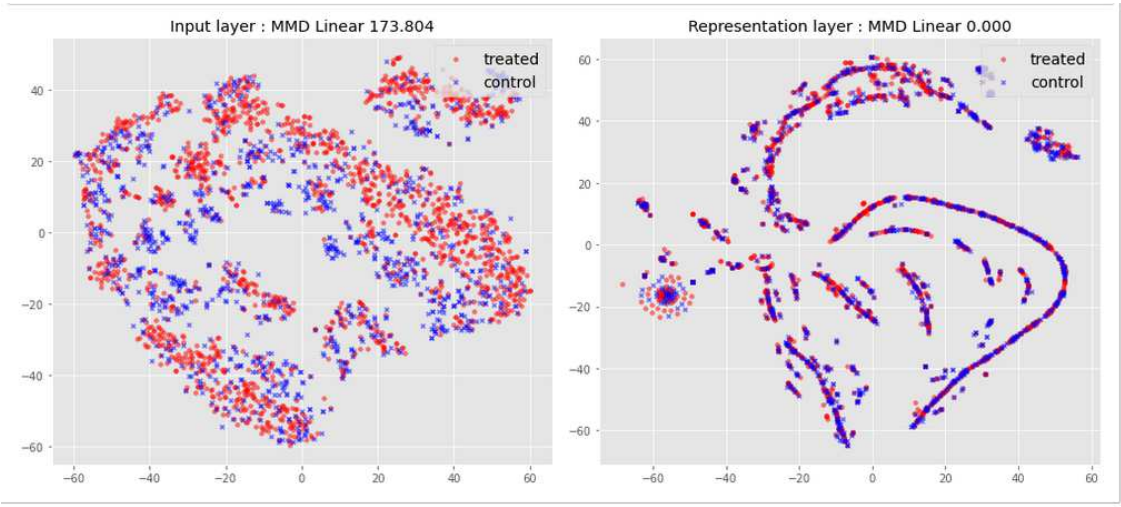

Figure 5.5: Scatter Plot comparison of Input and Representation Network Transformed Data

The dataset is splitted into train-test ratio of 80:20. For 8000 epochs, grid search is carried out with the values of hyperparameter γ taken $\{10^5, 100\}$ and learning rate values used as $\{0.1, 0.01, 0.001\}$. The results are specified in Table 5.1 and Table 5.2. The hyperparameter γ signifies the importance of IPM loss term. Higher the value of γ, higher is the contribution of IPM loss in overall loss.

Table 5.1: MedCI results for Outcome on semi-synthetic ACIC dataset

	γ=100			γ=10^5		
	lr=0.1	**lr=0.01**	**lr=0.001**	**lr=0.1**	**lr=0.01**	**lr=0.001**
ATE	3.556	1.258	3.397	1.918	1.626	3.043
ϵ_{ATE}	0.515	1.782	0.357	1.121	1.414	0.002
ATT	3.588	1.194	2.703	1.889	1.487	2.302
ϵ_{ATT}	0.924	1.469	0.039	0.774	1.176	0.361
ATC	3.522	1.324	4.116	1.948	1.770	3.810
ϵ_{ATC}	0.092	2.106	0.686	1.481	1.659	0.380
RMSE	5.532	5.329	3.937	5.523	5.409	3.797
$\sqrt{\epsilon_{PEHE}}$	6.788	6.942	5.244	6.706	6.825	4.748

Table 5.2: MedCI results for Medical Cost on semi-synthetic ACIC dataset

	γ=100			γ=10^5		
	lr=0.1	**lr=0.01**	**lr=0.001**	**lr=0.1**	**lr=0.01**	**lr=0.001**
ATE	462.16	474.92	477.48	463.85	470.64	479.85
ϵ_{ATE}	77.023	89.78	92.34	78.728	85.50	94.71
ATT	463.19	460.04	463.13	463.35	462.73	465.44
ϵ_{ATT}	80.05	76.90	79.99	80.21	79.59	82.30
ATC	461.08	490.33	492.33	464.37	478.83	494.77
ϵ_{ATC}	73.88	103.12	105.13	77.17	91.62	107.57
RMSE	87.00	84.13	84.29	86.56	85.09	83.44
$\sqrt{\epsilon_{PEHE}}$	87.21	101.22	101.54	87.89	94.84	103.89

Note that in the Tables:

- lr: Learning Rate
- ATE: Average Treatment Effect
- ϵ_{ATE} : Error in Average Treatment Effect
- ATT : Average Treatment Effect on Treated Population
- ϵ_{ATT} : Error in Average Treatment Effect on Treated Population
- ATC : Average Treatment Effect On Control Population
- ϵ_{ATC} : Error in Average Treatment Effect on Control Population
- RMSE : Root Mean Square Error
- ϵ_{PEHE}: Error in Precision in Estimation of Heterogeneous Effect

Also, note that the lowest RMSE value for both outcome ($RMSE = 3.797$) and medical cost ($RMSE = 83.44$) is achieved using learning rate of 0.001 and γ = 10^5. Also, the minimum $\sqrt{\epsilon_{PEHE}}$ loss of outcome ($\sqrt{\epsilon_{PEHE}} = 4.748$) and minimum ϵ_{ATE} loss of outcome ($\epsilon_{ATE} = 0.002$) is obtained by the same configuration. The variation of overall loss and medical cost loss with epochs for this configuration is shown in Figure 5.6 and Figure 5.7 respectively. Further, the minimum $\sqrt{\epsilon_{PEHE}}$ loss for medical cost is 87.21 and minimum ϵ_{ATE} loss is 77.023, achieved by learning rate = 0.1 and γ = 100.

5.3 Synthetic Experiments

A range of experiments were carried out on synthetic dataset using the proposed MedCI and MedSCI methods. The values of factual as well as counterfactual time-to-event is in the range [11.84,26.42]. On the other hand the range of values for factual medical cost is [0.28, 1937.77] while for counterfactual cost, it is [0.168, 2043.76].

5.3.1 MedCI

In case of MedCI, the representation network is simple neural network with input data mapped to 200 dimensional output layer via Rectified Linear Unit (ReLU) activation. This representation network output data is then fed as an input to both the hypotheses

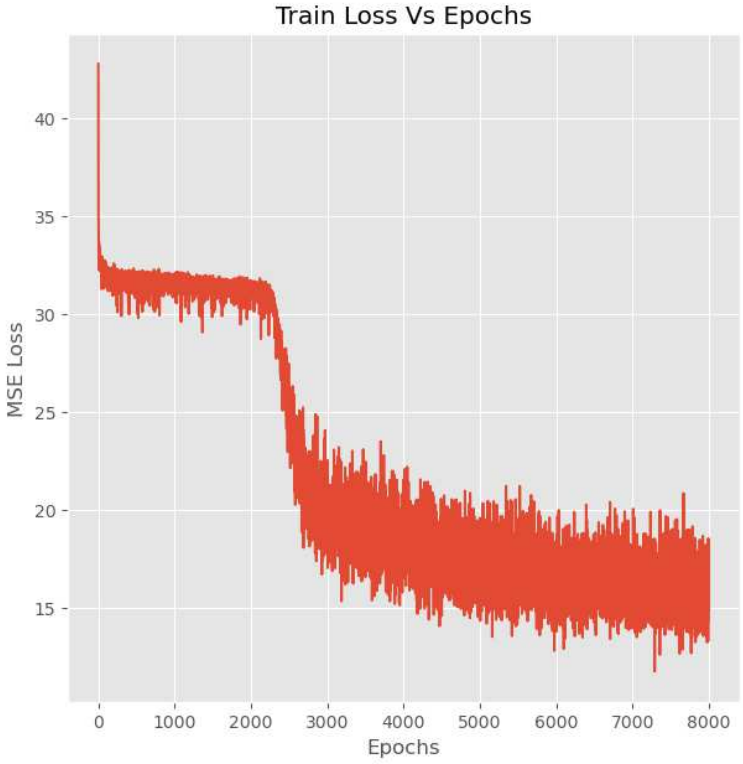

Figure 5.6: Training Loss vs Epochs (learning rate =0.001, γ=10^5)

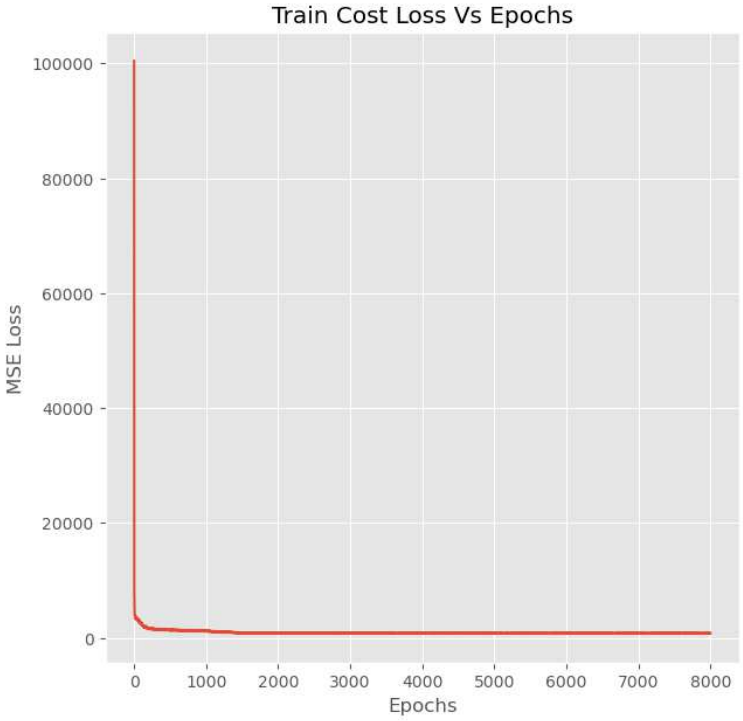

Figure 5.7: Medical Cost Training Loss vs Epochs (learning rate =0.001, γ=10^5)

networks. The medical cost hypothesis is parameterized via a deep neural network with two hidden layers consisting of 500 and 1000 nodes respectively. A dropout rate of 0.1 is used while the activation function is ReLU6 for hidden layers. On the other hand, for time-to-event modelling, the neural network is two hidden layer deep with 500 and 200 nodes respectively. Again, the dropout rate is kept at 0.1 and the activation function is ReLU6 for hidden layers. Identical network is defined for both the control and treatment arms for the respective output variable. Adam optimizer is used with weight decay of 0.1. The step size for learning rate decay is chosen as 50 while the multiplicative factor for learning rate decay is selected as 0.97.

The dataset is splitted into train-test ratio of 80:20. For 10000 epochs, grid search is carried out with the values of hyperparameter γ taken $\{1000, 100, 10\}$ and learning rate values used as $\{0.01, 0.001\}$. The results are specified in Table 5.3 and Table 5.4. Again, the hyperparameter γ signifies the importance of IPM loss term.

Table 5.3: MedCI results for Time-To-Event (Duration of Stay) on Synthetic Dataset

	lr=0.01			**lr=0.001**		
	γ=10	**γ=100**	**γ=1000**	**γ=10**	**γ=100**	**γ=1000**
ATE	1.145	1.314	0.681	1.226	1.296	1.175
ϵ_{ATE}	0.263	0.094	0.727	0.182	0.112	0.232
ATT	2.307	2.647	1.372	2.469	2.611	2.368
ϵ_{ATT}	0.835	1.176	0.098	0.998	1.139	0.897
ATC	0.02	0.01	0.01	0.028	0.0	0.0
ϵ_{ATC}	1.346	1.346	1.346	1.346	1.346	1.346
$RMSE_T$	3.933	4.264	2.982	3.699	3.827	3.508
$RMSE_C$	4.462	4.427	4.763	4.391	4.372	4.430
$\sqrt{\epsilon_{PEHE}}$	2.184	2.329	2.122	2.370	2.382	2.323

Note that in the Tables:

- lr: Learning Rate
- ATE: Average Treatment Effect
- ϵ_{ATE} : Error in Average Treatment Effect
- ATT : Average Treatment Effect on Treated Population
- ϵ_{ATT} : Error in Average Treatment Effect on Treated Population
- ATC : Average Treatment Effect On Control Population
- ϵ_{ATC} : Error in Average Treatment Effect on Control Population

- $RMSE_T$: Root Mean Square Error w.r.t. Treated Population
- $RMSE_C$: Root Mean Square Error w.r.t. Control Population
- ϵ_{PEHE}: Error in Precision in Estimation of Heterogeneous Effect

Table 5.4: MedCI results for Medical Cost on Synthetic Dataset

	lr=0.01			**lr=0.001**		
	γ=10	**γ=100**	**γ=1000**	**γ=10**	**γ=100**	**γ=1000**
ATE	1000.41	1000.51	997.98	1000.69	993.87	1000.57
ϵ_{ATE}	82.82	2.92	0.39	3.10	3.72	2.97
ATT	1012.22	1001.71	1037.49	997.85	985.34	1004.52
ϵ_{ATT}	14.49	3.98	39.76	0.11 8	12.38	6.78
ATC	988.75	999.32	959.02	1003.50	1002.28	996.67
ϵ_{ATC}	8.69	1.868	38.43	6.04	4.82	0.77
$RMSE_T$	25.59	24.39	33.01	52.87	59.93	48.95
$RMSE_C$	94.13	92.85	97.02	93.73	100.64	93.94
$\sqrt{\epsilon_{PEHE}}$	135.96	131.68	141.80	140.15	144.21	138.17

Note that the minimum value of $RMSE_T$ and $RMSE_C$ for time-to-event ($RMSE_T$=2.982, $RMSE_C$=4.372) arises at the configuration {Learning Rate = 0.01, γ = 1000} and {Learning Rate = 0.001, γ = 100} respectively. On the other hand, the minimum value of $RMSE_T$ and $RMSE_C$ for medical cost ($RMSE_T$=24.39, $RMSE_C$=92.85) arises at the configuration {Learning Rate = 0.01, γ = 100}. Further the minimum $\sqrt{\epsilon_{PEHE}}$ loss for time-to-event ($\sqrt{\epsilon_{PEHE}} = 2.122$) occurs at {Learning Rate=0.01, γ=1000} while for medical cost ($\sqrt{\epsilon_{PEHE}} = 131.68$), it occurs at {Learning Rate=0.01, γ=100}. Also, the minimum ϵ_{ATE} for time-to-event ($\epsilon_{ATE} = 0.094$) occurs at configuration {Learning Rate=0.01, γ=100} while the minimum ϵ_{ATE} for medical cost ($\epsilon_{ATE} = 0.39$) occurs at configuration {Learning Rate=0.01, γ=1000}. By observing the overall results, we conclude that the optimal hyperparameters are as follows:-

- Learning Rate = 0.01
- γ = 100

The variation of overall training loss and medical cost training loss with epochs for optimal hyperparameters is shown in Figure 5.8 and Figure 5.9.

We compare our results for time-to-event modelling on baseline CFRNet using the same synthetic dataset. The results are tabulated in Table 5.5.

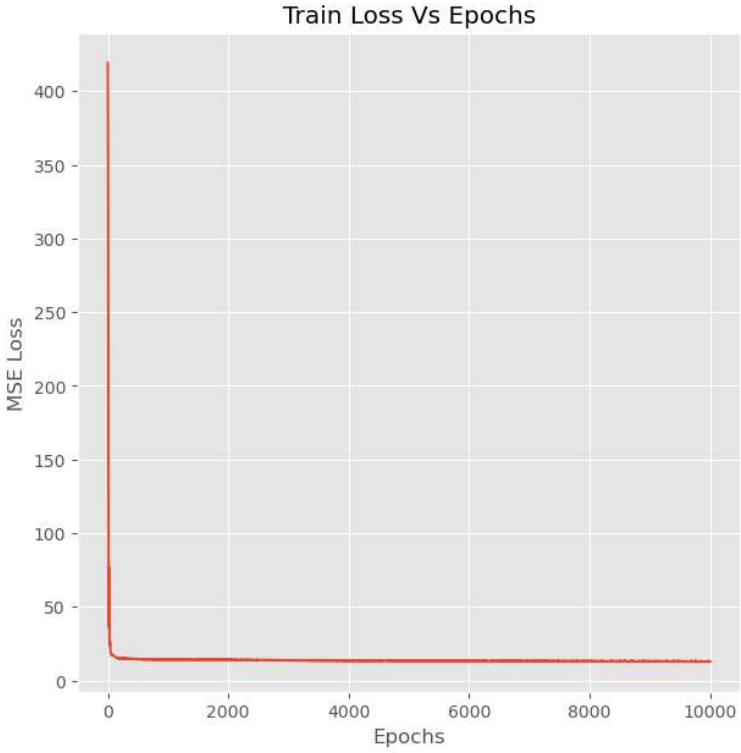

Figure 5.8: Training Loss vs Epochs (learning rate =0.01, γ=100)

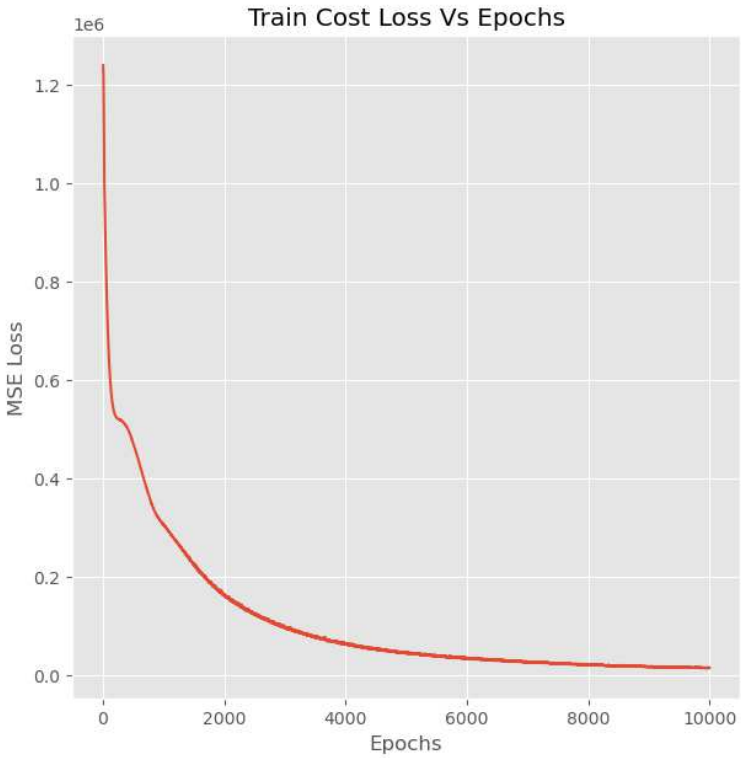

Figure 5.9: Medical Cost Training Loss vs Epochs (learning rate =0.01, γ=100)

Table 5.5: CFRNet results for Time-To-Event on Synthetic Dataset

ATE	3.063
ϵ_{ATE}	1.596
ATT	2.878
ϵ_{ATT}	1.469
ATC	3.234
ϵ_{ATC}	1.714
RMSE	2.703
$\sqrt{\epsilon_{PEHE}}$	3.346

From Table 5.3 and Table 5.5, it is evident that our framework has a lower $\sqrt{\epsilon_{PEHE}}$ as well as ϵ_{ATE} and it outperforms the baseline.

5.3.2 MedSCI

We evaluated our proposed MedSCI method on the synthetic dataset. The added medical cost hypothesis layer is parameterized using deep neural network consisting of two hidden layers with 1200 and 200 nodes respectively. A dropout rate of 0.1 is used with ReLU6 activation function for hidden layers. Further, Adam optimizer is used with learning rate of 0.001 and weight decay of 0.1. The step size for learning rate decay is chosen as 100 while the multiplicative factor for learning rate decay is selected as 0.96. The dataset consists of 50:50 train-test ratio with 5000 instances for each category. The train set is further divided into train and validations sets with split ratio of 70:30.

For estimation of time-to-event, we use the methodology as defined in SurvCI [11]. The shape, scale and logits parameters are modelled using a sequential neural network. Then these parameters are used to predict time-to-event from area under the curve of survival probability distribution. Adam optimizer is used and the learning rate is chosen as 0.0003 while the batch size is kept at 350 for time-to-event hypothesis network.

For training the model, a nested loop of 200 iterations of time-to-event hypothesis network is run where inside each training iteration,the medical cost hypothesis network is trained alongside for 8000 epochs. The final results are presented in Table 5.6.

Here, in the table :

- ϵ_{ATE} : Error in Average Treatment Effect

Table 5.6: MedSCI results on Synthetic Dataset

Loss	Value
ϵ_{ATE}	0.52
$CI_{a=0}$	0.75
$CI_{a=1}$	0.78
$\sqrt{\epsilon_{PEHE}}$	2.20
IPM	0.123
$ELBO$	3.93
$RMSE$	1.69
$RMSE_{Cost}$	509.79

- CI : Concordance Index
- ϵ_{PEHE} : Error in Precision in Estimation of Heterogeneous Effect
- IPM : Integral Probability Metric Loss
- $ELBO$: Binary treatment based evidence lower bound loss
- $RMSE$: Root mean squared error in duration of stay (time-to-event)
- $RMSE_{Cost}$: Root mean squared error in Medical Cost

5.4 Conclusion

We proposed two novel counterfactual inference frameworks, MedCI and MedSCI for the joint estimation of medical cost and time-to-event analysis. The frameworks are evaluated on a semi-synthetic and synthetic dataset and the results are promising. For MedCI, in case of semi-synthetic medical cost that varies from 50 to 650, the root mean square error in estimation is as low as 83 while for the synthetic cost that lies in range [0,2000], the error is as low as 25. Further, the $\sqrt{\epsilon_{PEHE}}$ in medical cost for semi-synthetic experiments touches the lower bound of 87 while for synthetic experiment, it is as low as 132. Similarly, for time-to-event outcome i.e. the duration of stay that varies from 11 to 26 days, the root mean square error in estimation is roughly 4 while the $\sqrt{\epsilon_{PEHE}}$ is around 2.3. Even for MedSCI, we observe low values of ϵ_{ATE}, $\sqrt{\epsilon_{PEHE}}$, IPM loss, $ELBO$ loss and $RMSE$ loss.

This highlights that the proposed methods perform fairly well for the evaluation of treatment effect in a multi-outcome prediction problem with one of the outcome being

time-to-event. As stated earlier, to the best of our knowledge, there existed no algorithm that accurately estimates the inpatient duration of stay in hospital with associated medical costs, for each alternative therapy while evaluating the treatment effect of each intervention on both the outcome of interests. Our research work addresses this gap and provides a solution for the same.

5.5 Future Work

This work can be extended for:-

- We assumed random censoring throughout. This work can be extended for informative censoring.
- We assumed binary treatment scenario. This work can be extended for multiple treatment case.
- The work can be explored for a competing risk scenario where multiple events can influence the duration of stay.

REFERENCES

[1] "Survival data analysis." [Online]. Available: https://www2.karlin.mff.cuni.cz/~pesta/NMFM404/survival.html#References

[2] Michael, "Wikiwand - Survival function." [Online]. Available: https://wikiwand.com/en/Survival_function

[3] 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, publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1958.10501452. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452

[4] J. In and D. Lee, "Survival Analysis: Part II - Methods reducing a gap between statistics and real world," *Korean Journal of Anesthesiology*, vol. 72, May 2019.

[5] W. Nelson, "Hazard Plotting for Incomplete Failure Data," *Journal of Quality Technology*, vol. 1, no. 1, pp. 27–52, 1969, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00224065.1969.11980344. [Online]. Available: https://doi.org/10.1080/00224065.1969.11980344

[6] S. Kärkkäinen, K. Silventoinen, P. Svedberg, K. Alexanderson, A. Huunan-Seppälä, K. Koskenvuo, M. Koskenvuo, J. Kaprio, and A. Ropponen, "Health-Related and Sociodemographic Risk Factors for Disability Pension due to Low Back Disorders A 30-Year Prospective Finnish Twin Cohort Study," *Journal of occupational and environmental medicine / American College of Occupational and Environmental Medicine*, vol. 53, pp. 488–96, May 2011.

[7] L. J. Wei, "The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis," *Statistics in Medicine*, vol. 11, no. 14-15, pp. 1871–1879, 1992.

[8] 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, _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1972.tb00899.x. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1972.tb00899.x

[9] B. Neal, *Introduction to Causal Inference*, 2020.

[10] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," May 2017, number: arXiv:1606.03976 arXiv:1606.03976 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1606.03976

[11] M. Gupta, G. Kannan, R. Prasad, and G. Gupta, "Learn to Live Longer: Counterfactual Inference using Balanced Representations for Parametric Deep Survival Analysis," preprint, Oct. 2022. [Online]. Available: https://www.

techrxiv.org/articles/preprint/Learn_to_Live_Longer_Counterfactual_Inference_using_Balanced_Representations_for_Parametric_Deep_Survival_Analysis/19105451/2

[12] M. Chen, X. Wu, J. Zhang, and E. Dong, "Prediction of total hospital expenses of patients undergoing breast cancer surgery in Shanghai, China by comparing three models," *BMC Health Services Research*, vol. 21, no. 1, p. 1334, Dec. 2021. [Online]. Available: https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-021-07334-y

[13] W. E. Muhlestein, D. S. Akagi, A. R. McManus, and L. B. Chambless, "Machine learning ensemble models predict total charges and drivers of cost for transsphenoidal surgery for pituitary tumor," *Journal of Neurosurgery*, vol. 131, no. 2, pp. 507–516, Aug. 2019. [Online]. Available: https://thejns.org/view/journals/j-neurosurg/131/2/article-p507.xml

[14] L.-L. Tong, J.-B. Gu, J.-J. Li, G.-X. Liu, S.-W. Jin, and A.-Y. Yan, "Application of Bayesian network and regression method in treatment cost prediction," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, p. 284, Dec. 2021. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01647-y

[15] S.-M. Lee, J.-O. Kang, and Y.-M. Suh, "Comparison of Hospital Charge Prediction Models for Colorectal Cancer Patients: Neural Network vs. Decision Tree Models," *Journal of Korean Medical Science*, vol. 19, no. 5, p. 677, 2004. [Online]. Available: https://jkms.org/DOIx.php?id=10.3346/jkms.2004.19.5.677

[16] A. Taha, S. Taha-Mehlitz, V. Ochs, B. Enodien, M. D. Honaker, D. M. Frey, and P. C. Cattin, "Developing and validating a multivariable prediction model for predicting the cost of colon surgery," *Frontiers in Surgery*, vol. 9, p. 939079, Nov. 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fsurg.2022.939079/full

[17] Luciana Bertocco de Paiva Haddad, Luana Regina Baratelli Carelli Mendes, Liliana Ducatti, Vinicius Rocha-Santos, Wellington Andraus, and Luiz Augusto Carneiro D'Albuquerque, "Machine Learning in the Prediction of Costs for Liver Transplantation," *Quality in Primary Care*, vol. 25, no. 5, p. 9, Oct. 2017. [Online]. Available: https://www.primescholars.com/articles/machine-learning-in-the-prediction-of-costs-forliver-transplantation-100448.html

[18] M. A. Abd-Elrazek, A. A. Eltahawi, M. H. Abd Elaziz, and M. N. Abd-Elwhab, "Predicting length of stay in hospitals intensive care unit using general admission features," *Ain Shams Engineering Journal*, vol. 12, no. 4, pp. 3691–3702, Dec. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2090447921001349

[19] E. Rocheteau, P. Liò, and S. Hyland, "Predicting Length of Stay in the Intensive Care Unit with Temporal Pointwise Convolutional Networks," Nov. 2020, number: arXiv:2006.16109 arXiv:2006.16109 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2006.16109

[20] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU Collaborative Research Database, a freely available multi-center database for critical care research," *Scientific Data*, vol. 5, no. 1, p. 180178, Sep. 2018, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/sdata2018178

[21] S. Sheikhalishahi, V. Balaraman, and V. Osmani, "Benchmarking machine learning models on multi-centre eICU critical care dataset," *PLOS ONE*, vol. 15, no. 7, p. e0235424, Jul. 2020. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0235424

[22] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, p. 24, Feb. 2018. [Online]. Available: https://doi.org/10.1186/s12874-018-0482-1

[23] C. Lee, W. Zame, J. Yoon, and M. v. d. Schaar, "DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, number: 1. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11842

[24] C. Nagpal, X. Li, and A. Dubrawski, "Deep Survival Machines: Fully Parametric Survival Regression and Representation Learning for Censored Data With Competing Risks," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3163–3175, 2021.

[25] C. Shi, D. Blei, and V. Veitch, "Adapting Neural Networks for the Estimation of Treatment Effects," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/8fb5f8be2aa9d6c64a04e3ab9f63feee-Paper.pdf

[26] P. Chapfuwa, S. Assaad, S. Zeng, M. J. Pencina, L. Carin, and R. Henao, "Enabling counterfactual survival analysis with balanced representations," in *Proceedings of the Conference on Health, Inference, and Learning*, ser. CHIL '21. New York, NY, USA: Association for Computing Machinery, Apr. 2021, pp. 133–145. [Online]. Available: https://doi.org/10.1145/3450439.3451875

[27] A. Curth, C. Lee, and M. van der Schaar, "SurvITE: Learning Heterogeneous Treatment Effects from Time-to-Event Data," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 26 740–26 753. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/e0eacd983971634327ae1819ea8b6214-Paper.pdf

[28] J. L. Hill, "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011, publisher: Taylor & Francis _eprint: https://doi.org/10.1198/jcgs.2010.08162. [Online]. Available: https://doi.org/10.1198/jcgs.2010.08162

[29] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet, "On the empirical estimation of integral probability metrics," *Electronic Journal of Statistics*, vol. 6, no. none, pp. 1550 – 1599, 2012, publisher: Institute of Mathematical Statistics and Bernoulli Society. [Online]. Available: https://doi.org/10.1214/12-EJS722

[30] P. R. Hahn, V. Dorie, and J. Murray, "Atlantic Causal Inference Conference (ACIC) Data Analysis Challenge 2017," *arXiv: Methodology*, 2019.

[31] Gross, R. T., and others, "Infant Health and Development Program (IHDP): Enhancing the Outcomes of Low Birth Weight, Premature Infants in the United States, 1985-1988," 1993.