Universiteit Leiden

Master Computer Science

Exploring the presence and effect of treatment selection bias for localized prostate cancer data from Netherlands Cancer Registry

Name: Shixuan Xie
Student ID: S2171112

Date: 17/11/2020

1st supervisor: Prof.dr. P.J.F. Lucas
2nd supervisor: Drs.ir. M. Sieswerda

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Acknowledgements

I would like to thank my supervisor Prof. Peter Lucas and Drs. Melle Sieswerda for their support and advice. I would also like to thank another intern in the group, Ruby van Rossum, who gave me some more ideas and suggestions when we compared our results. Finally, I also want to thank my parents and my friends for all their support and strength they bring to me.

Contents

Abstract

Observational data have been used increasingly in the study of cancer for the evaluation of the efficacy of therapies. However, without proper experimental design especially randomization, all kinds of bias may appear when exploring observational data.

In this thesis, we focused on treatment selection bias. We explored the impact of selection bias on therapeutic effect of localized prostate cancer using data from the Netherlands Cancer Registry between 2005 and 2014. To begin with, Cox proportional-hazards models were constructed before and after the adjustment of propensity score. Both of them indicated active treatment decreased the risk of death (hazard ratio = 0.79, 95% confidence interval = [0.73,0.85]). We also found the role of adjustment was limited: the Cox model after the adjustment was almost the same as the unadjusted one. Then we constructed a hybrid Bayesian Network (BN) combining our prior knowledge with structure learning. After that, the parameter learning was finished with the EM algorithm. Active treatment was associated with a survival benefit in both the Cox model and the BN model. However, this positive association was much weaker in the BN model compared with Cox proportional-hazards models. We also found that age was a factor that could partly explain the difference in outcome between active treatment group and non-active treatment group. Older patients had a higher risk of death.

In comparison between the survival rate of male residents in the Netherlands, survival rate of non-active treatment group visualized by Kaplan–Meier estimator and survival rate of active treatment group visualized by Kaplan–Meier estimator, we got improbable results: after adjusting for age and year of diagnosis for active treatment group, the survival rate of the general male population was lower than the survival rate of active treatment group. If adjusted for age and year of diagnosis for non-active treatment group, the survival rate of general male population was lower than non-active treatment group. It was a glance of the strong effect of selection bias. An explanation was patients who underwent active treatment had better underlying health, which could partly be indicated by a younger age. We also suspected that there were some other unmeasured confounders that related to baseline health.

Our experiment was a reproduction of Giordano's paper [32]. Our experiment was not the only one to find selection bias in observational dataset and the adjustment of propensity score had little impact on these irrational results. In the paper of Giordano, they also found that localized prostate cancer patients who received active treatment had a lower risk of death compared with noncancer control and the propensity score adjustment played little rule to alter the findings [32]. So when analysing observational data, researchers have to be cautious and the results should be viewed critically.

Keywords: Selection bias, Propensity score adjustment, Cox proportional-hazards model, Bayesian Network

1 Introduction

Clinical research can be divided into two categories: *observational research* and *experimental research*. Data obtained by them are known as *observational data* and *experimental data*, respectively. The main difference between the two types of data is whether or not study design and intervention are applied in the data collection step [31]. In an experimental study, the data is produced by an experimental design and test method. Researchers will set some intervention factors artificially to acquire the right type of data which is available to answer their research questions as clearly as possible [28]. The missing step of experimental design may introduce *bias*. *Bias* is a systematic error in which the results obtained by analysing the samples do not necessarily reflect the true results of the target population.

A *randomized controlled trial* (RCT) is one of the common methods of experimental study. A well-organized RCT is the gold standard in treatment evaluation in clinical trails [39]. RCT is accomplished by randomization, which means dividing patients into two or more groups randomly. One of the groups is called the *control group* and the others are the *experimental groups*. The former receives no intervention or a standard intervention with a known effect and the latter undergo different therapies, one for each group. If the sample size is big enough, the random allocation will ensure that there is no relation between the covariates and treatment allocation. In other words, the distribution of covariates between different groups are similar. To further reduce the bias, RCT can be blinded or double-blinded. In a *blinded RCT*, patients in control and treatment groups are unaware of the group they are in. Giving a placebo to patients in the control group is a common way to do that. This is to reduce *performance bias*, which is introduced by the awareness of the applied intervention [59]. In the article by Henry K. Beecher [16], he concluded that the symptoms of 35% patients were relieved by placebos compared with no intervention. Other papers from Turner *et al.* and Roberts *et al.* also proved that placebos played an incredible role in symptoms alleviation [68] [61].

In addition to the use of the placebo, the information that may influence the participants (both subjects and researchers) will be withheld until the end of the experiment in *a double-blinded RCT*. This is to reduce the risk that researchers may know which intervention was received. This prior knowledge might make researchers behave differently and affect outcome measurement [41].

Nevertheless, RCT is not always feasible due to the high cost of time and money. Compared with an observational study, it is more expensive to acquire a dataset with large sample sizes. Take our experiment as an example: the 5-year survival rate of patients who diagnosed with localized prostate cancer is about 90%, and the 10-year survival rate is about 75% [11]. It is expensive and unrealistic to design an experiment for such a long time. Regardless of the high cost, RCT is still not perfect. The ethical issues ranging from privacy protection informed consent and risk minimization [60] make RCT not always possible. Information about the procedures, the risk and benefits of the treatment should be informed to participants. This is sometimes contradictory to blindness. Besides, according to the Declaration of Helsinki, "the well-being of the individual research subject must take precedence over all other interests" [13]. Ethically, patients should receive the best-proven therapy if it is feasible [13]. But for some patients, the well-performed randomization and blinding approach may lead to the missing opportunity for the most suitable treatment. Another issue is that in many cases, patients in control group would be more painful [54]. Finally, except the biases mentioned above, there are many other kinds of biases in reality. It is hard to reduce or eliminate all of them. All of these make RCT not always possible.

On the other hand, *Observational data*, which refer to the dataset that collected directly by observation without manipulation or *intervention*, have the advantages such as large sample sizes, low cost and long-term follow-up. Thus they are increasingly used to evaluate the effectiveness of therapies and cancer outcomes [40] [32]. One of the biases that would be introduced by the missing steps of randomization is known as *Selection bias*. It will cause the differences in sample's characteristics between groups. When assessing the effectiveness of

therapies, the discrepancies of outcomes may be caused by these difference in characteristics rather than the therapies itself. If researchers compare the outcome of different treatment groups directly, they would be misled and come to the wrong conclusion. *Treatment selection bias* is a type of selection bias that introduced when doctors are predisposed to select a specific treatment based on different characteristics of patients without realizing it. For instance, when comparing two treatments with potential difference in effect, patients who are in better health (considered more likely to tolerate the treatment) would have a higher tendency to receive the more toxic treatment [32]. If that is the case, it is difficult to determine if and how much the difference in outcome and survival time is caused by the treatment.

In summary, a well-organized RCT is the gold standard for effectiveness research in clinical traits [39]. The experimental designs for RCTs aim to reduce bias and guarantee that the different outcome will more relate to different therapies, rather than patient's characteristics like their age, health, the prior reasoning or any other unrelated stuff. But the limitations make RCTs not always feasible. And sometimes, the advantages of observational data including large sample sizes and long time follow-up are considered more important. So observational data is sometimes a good choice to assess the effectiveness of different therapies. However, if observational data is used, researchers should be careful about bias.

1.1 Related work

Investigators have already been aware of impact of the potential bias on observational dataset. In the paper of Wong *et al.* [69], they concluded that for patients who have low- or intermediate-risk prostate cancer and aged between 65 and 80 years, active treatment may have a positive effect on survival time. However, under the assumption that multivariate methods like propensity score adjustment can not fully eliminate selection bias and confounding, they recommended that these results should be validated in RCTs. Under a similar assumption that selection bias can not be adjusted for covariates like stage, grade, region and no. of PSA tests completely, Giordano *et al.* [32] reanalysed three published datasets separately using Cox proportional-hazards model after propensity score adjustment. The three cases were: **1.** Androgen-deprivation therapy versus none after primary radiation therapy for locally advanced prostate cancer, **2.** Active treatment versus none for men with localized prostate cancer, which is a reproduction of Wong's experiment, and **3.** 5-FU based adjuvant chemotherapy versus none for lymph node-positive colon cancer. Giordano *et al.* concluded that for all the three cases, little changes can be found in the outcome of Cox model before and after reducing the effect of confounders with propensity score methods. What's more, two out of the three cases showed that the observational data produced improbable and incorrect results, which was a small glance of the strong effect of selection bias. Realizing this, Giordano *et al.* suggested that the conclusion from observational data should be viewed critically as they may result from an interaction of a selection bias [32]. When assessing the effectiveness of therapies using observational data, researchers should be cautious and modest.

1.2 Project overview

This thesis aims to check the presence of treatment selection bias and how it would influence the outcome. Therefore, we reanalysed NCR data of localized prostate cancer using propensity score adjustment and Cox proportional-hazards models just as described in Giordano's paper [32] to compare the outcome of active treatment group and non-active treatment group for men aged 65 to 80 years. Because our experiment used observational data rather than RCT data, the *average treatment effect* (ATE) can not be recognised from the dataset due to the effect of confounders. To get an unbiased estimate of ATE, propensity score adjustment was applied and the dataset was stratified into 5 subgroups according to the quintiles of propensity score. The stratification can remove 90% of the bias caused by the confounders [22]. After that, we used Cochran-Mantel-Haenszel (CMH) test to test if the adjustment works. CMH test is a

hypothesis testing method to test the association between two variables when stratifying data with a third variable [48]. If the third variable is not taken into account, *Simpson's paradox* may be observed. *Simpson's paradox* is a phenomenon in which the association between a pair of variables may disappear, increase, decrease or change direction in these two cases: **1.** Study the whole population and **2.** Study the subpopulation that stratified according to the categories of the third variable. There are some other methods for test like calculating the *standardized mean difference* (SMD). But as a reproduction of Giordano's paper, we applied the CMH test.

After the adjustment, Cox proportional-hazard model was applied. It is one of the commonly used models in survival analysis. Survival analysis is a branch of statistics that analyses the correlation between survival time, the outcome and a number of variables. In an ideal situation, if the dataset is non-censored, a series of analysis methods are available. But the exact survival time of patients are sometimes unknown due to the death of the patient, the refusal of follow-up or the event of interest does not happen at the end of the experiment. Compared with other methods, survival analysis is able to make the most use of the right-censored data instead of removing them. We constructed two Cox proportional-hazard models, one performed propensity score adjustment, and the other did not. By comparing the two models, we could explore how much the propensity score adjustment helped to eliminate the effect of confounders. After the adjustment, another five Cox models were established, one for each age group to examine the relations between age and outcome associated with treatment.

Next, as an extension, we constructed a *Bayesian Network* (BN) to study treatment selection bias from a perspective of causal inference. The BN model is a type of probabilistic graphical model (GM) [17], which combines principles from probability theory and graph theory. Each node in this network acts as a random variable and the edges are the relations between the nodes. The quantitative parameter are represented by the *conditional probability distribution* (CPD) of the nodes. The graphic structure makes it a good tool for visualizing the relations among variables. It is easy to follow every step and understand how the change of one or more variables will influence other ones. This transparency distinguishes the BN model from other "black box" machine learning approaches such as neural network [58]. What's more, compared with the Cox model, some prior knowledge can be added to make the results easier to explain.

To evaluate our BN model, we applied sensitivity analysis with AUC-ROC curves. It is a commonly used approach to test the performance of a binary classifier. But apart from the capability of classification, what we are more interested in when studying the association between survival and a series of covariates is the probability distributions estimated by our model. We would like to test the relations between the predicted survival probabilities and the observed survival probabilities. For instance, doctors will ask: what is the probability that a patient with certain characteristics still being alive two years after diagnosis? Therefore, in our experiment, a model that fits the dataset well, or in other words, provides predicted probabilities close to the observed probabilities, is more meaningful in practical application for predicting and decision making. This comparison between predicted and observed probabilities is visualized by a calibration plot.

Finally, we compared the survival rate of the active treatment group, non-active treatment group and the weighted survival rate of general male residents in the Netherlands [1] adjusted for age and year of diagnosis. If the survival rate of active treatment group was even higher than general population, as the same with Giordano's paper [32], we could reasonably suspect that selection bias can be found in our dataset.

1.3 Thesis structure

The remaining part of the paper is as follows: Chapter 2 is about the background knowledge of prostate cancer and the cancer staging system. The mathematical principles of propensity score approaches, survival analysis, Cox proportional-hazards model and BN model will be

comprehensively explained in chapter 3. Chapter 4 describes the steps we took in detail. The results will be illustrated in chapter 5. Discussion, conclusion and future work are shown in chapter 6.

2 Clinical background

2.1 Development, detection and prognosis of prostate cancer

Prostate cancer is one of the most common cancers diagnosed in men; prostate cancers are often classified as belonging to *adenocarcinomas* ("adeno" means gland-like) histological type [50] as they originate in a glandular cell or the cancer cells look similar to gland cells, or both. As often in cancer, the cumulative exposure to particular risk factors, such as specific hormones (for example, testosterone) and unrepaired DNA damage due to wear and tear, implies to elderly people have cancer more frequently. This rule also applied to prostate cancer. Therefore, the incidence of prostate cancer will increase with age. Eight out of ten prostate cancers are diagnosed in men over their 65s [36], and they are seldom diagnosed in men under their 40s. Other primary risk factors of prostate cancer include race and family history, i.e., genetic factors may be involved. There are also some less clear factors, such as diet and obesity (probably also through hormonal influences) that will affect the incidence of prostate cancer.

Early prostate cancer is symptomless [38]. As it develops further, symptoms such as frequent urination, urgency of urination and a slow urine stream may appear. Most of the patients are only diagnosed with prostate cancer after the appearance of symptoms [38]. But there are also some screening tests to find prostate cancer in its early stage before the symptoms have been developed. *Prostate-specific antigen* (PSA) blood test is one of an important screening tests and is usually the first step in diagnosing prostate cancer. It can decrease the risk of diagnosing with advanced prostate cancer and lead to a lower prostate cancer mortality [43] [14]. PSA is a protein produced by epithelial cells of the prostate gland, whose function is believed to be liquefying the seminal fluid [47]. An abnormal increase of PSA level in blood may indicate prostate diseases. However, PSA level can also rise due to some benign prostate diseases. So PSA test by itself can not indicate the presence of prostate cancer. Generally speaking, a traditional threshold PSA level of 4 ng/ml is applied since the level of PSA in blood is under 4 ng/ml for most of the men without prostate cancer [2]. Other important thresholds of PSA value including 10 ng/ml and 20 ng/ml, as described in the 7th edition of the American Joint Committee on Cancer (AJCC) cancer staging manual [29], which is an international standard that decides the extent of tumor. Additionally, another screening method is *digital rectal exam* (DRE). In a DRE, the doctor will insert a lubricated, gloved finger into one's rectum [3]. By feeling for any hard spots or bumps, the doctor will determine the state of one's prostate.

If a man is suspected to have prostate cancer after the screening tests, a *prostate biopsy* may be needed to confirm the diagnosis. In a prostate biopsy, a core needle is used to collect a very small sample from the prostate gland, after which the sample will be looked at under a microscope by a pathologist. Normally, doctors will take more than six samples from different places of prostate to increase the detection rate and decrease the false negative rate [49].

In the advanced stage, prostate cancer cells may metastasize to other parts of the body, through the lymphatic route to the lymph nodes (common is that the inguinal lymph node is involved), and the hematogenous route to bones (in particular the vertebral column). However, aggressive or metastatic prostate cancer is less common. Most prostate cancers grow slowly. The 5-year, 10-year and 15-year relative survival rates for patients diagnosed with regional or local (non-metastatic) prostate cancer is 99%, 95% and 82% respectively, whereas only 33% men with metastatic prostate cancer survive longer than five years [20].

2.2 TNM classification of malignant tumors

The TNM classification of malignant tumors (TNM) is a globally recognized standard classification method developed by the Union for International Cancer Control (UICC). Today the TNM staging system is fully described in the American Joint Committee on Cancer (AJCC)

manual [29]. It is used to characterize the prognosis of cancer in terms of its spread to other, often distant parts of the body. In this staging system, tumor stage is described by means of three aspects: T, N and M, where [4]:

- **T** describes the size and extent of the primary (i.e., where the tumor started) tumor;
- **N** describe whether or not nearby lymph nodes are involved;
- **M** describes if the cancer is metastasized, i.e., whether cancer has spread to other parts of the body, such as bone, liver, and lungs.

The categories that can be assigned to each descriptors T, N and M traditionally start with a number and are written behind the descriptors, eg., T1, N0, M0. Sometimes the categories are subdivided into subcategories like T1a and T1b to tell more detailed information. These categories are tumor-specific and defined in the AJCC staging manual [29]. They are usually ordered from good prognosis (limited disease) to bad prognosis (extensive disease). For prostate cancer, the following category of descriptor T can be 1,1a,1b,1c,2,2a,2b,2c,3,3a,3b and 4. For descriptor N, the category is either 0 (lymph nodes not affected) or 1 (lymph nodes affected). For descriptor M, the category is also 0 (no metastasis) and 1 (distance metastasis present). The smaller the number, the less aggressive the tumor is. When the numbers are the same, the aggressiveness is in alphabetical order. To characterize tumors that can not be measured or assessed, the letter "X" is used. In the end, cases of cancer with similar prognosis are grouped into a roman numeral according to the categories of T, N and M [29]. This roman numeral is known as aggregated stage group or anatomic stage group and ranges from I to IV. Depending on the tumor and histologic type [29], the grouping rule can be different. The grouping rule of prostate cancer can be found in AJCC manual [29].

Depending on whether or not radical prostatectomy has been performed, the staging can be either clinical or pathological. They are distinguished with a prefix to the descriptor, eg., cT or pT. Some less common prefixes include yc, yp and a. Prefix y denotes clinical (yc) or pathological (yp) state after radiation therapy and a represents cancer found at autopsy [29]. *Clinical staging* is decided before surgery using the information of physical examinations, lab results, medical imaging, and tumor biopsies. Thus clinical staging is possible for all the patients. However *pathological staging* (*histological staging*), in most cases, is only possible for those who have undergone prostatectomy. This is because the sample of the tumor and some regional lymph nodes removed from the area around the tumor are needed for pathologic classification and will be examined under a microscope. These tissues can only be resected by the operator [29]. The test results of the tissues, combined with the clinical information are used to determine the pathological stage. Therefore, pathological staging is more precise than clinical stage.

2.3 Gleason scoring system

Apart from the TNM staging system, some other grading systems may be available for special types of cancer. For prostate cancer, the *Gleason score* indicates how aggressive the prostate cancer is. It distinguishes between *primary* and *secondary Gleason grade*. The Gleason grade ranges from 1 to 5, where a lower grade means the cancer cells are better differentiated and resemble normal prostate tissue [56], thus less aggressive. On the contrary, a higher grade means that the cancer cells are worse differentiated and looks more abnormal. Figure 1 [23] shows the different patterns in Gleason grade.

To determine the *primary Gleason grade* and *secondary Gleason grade*, the tissue removed from the biopsy will be looked at under a microscope. The *primary grade* refers to the most frequent pattern observed in the tissue. This pattern should be larger than 50% of the whole pattern. The *secondary grade* refers to the second most common pattern and should be less than 50%. The sum of the two Gleason grades is the final Gleason score. Thus the Gleason

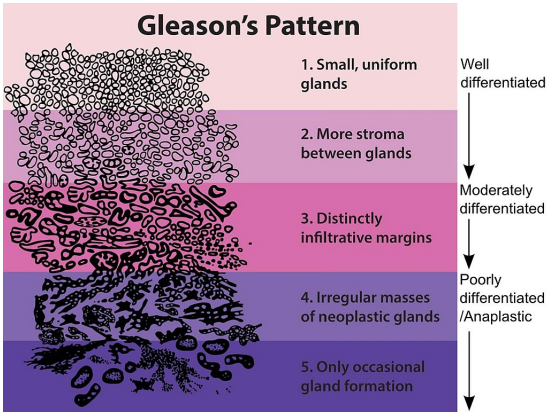


Figure 1: Different patterns in *Gleason grade* [23]

score has a range of two to ten. However, there will be some information lost if doctors only record the final Gleason score. For instance, the aggressiveness of cancer may be different between the two cases: one patient has a primary grade of four and a secondary grade of three. The other one has a primary grade of three and a secondary grade of four. But they share a value of 7 in the final Gleason score.

3 Methodology

3.1 Propensity score adjustment

3.1.1 Estimation of Average Treatment Effect (ATE) in an RCT dataset

Suppose there are n units or cases in a dataset $D = \{(x_i, t_i, y_i) \mid i = 1, \ldots, n\}$, where $X = x_i = (x_i^1, \ldots, x_i^m)$ represents *covariates* or *baseline characteristics* (pretreatment variables), $T = t_i$ is a binary variable indicates treatment, with $t_i = 0$ if the i-th unit received control treatment and $t_i = 1$ if the i-th unit received experimental treatment under study [63]. $Y = y_i$ stands for treatment *effect* with $y_i = 0$ the effect is *absent* and $y_i = 1$ the effect is *present*.

Let $P(X, T, Y)$ be the joint probability distribution defined using data from dataset D, with $P(X = x_i, T = t_i, Y = y_i)$ the probability of the occurrence of a particular instance in the dataset. If D is obtained from an RCT of high quality, the following properties will hold:

First, treatment allocation T is *independent* of X and will not be affected by X:

$$X \perp\!\!\!\perp_p T \mid \varnothing$$

hence

$$P(T \mid X) = P(T)$$

It should be noted that this statement only holds for pretreatment patient characteristics X because treatment may affect particular physiological mechanisms in the patient.

Second, the randomization also guarantee a similar distribution of all covariates X between groups

$$P(X \mid T = 1) \approx P(X \mid T = 0)$$

This corresponds to the condition of lack of discimination between patients irrespective of treatment, but now after marginalisation out the effect variable:

$$\sum_Y P(X, Y \mid T = 0) \approx \sum_Y P(X, Y \mid T = 1)$$

Finally, treatment allocation is said to be *strongly ignorable* given a subset of variables $V \subseteq X$ if

$$Y \perp\!\!\!\perp_p T \mid V \tag{1}$$

This means the effect of T to Y should be mediated through V. According to the properties of conditionally independent, we have

$$P(Y \mid V) = P(Y \mid T, V) \tag{2}$$

Note that $P(Y \mid T, V)$ is what we can observed from the dataset D. Under the strongly ignorable assumption, it is possible to estimate the *average treatment effect (ATE)* as the measure of the effect of different therapies when comparing outcomes of different treatment groups directly, i.e.,

$$\mathrm{ATE_k} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in \{0,1\}} P(Y = k \mid X = x_i, T = j)$$

with $k \in \{0, 1\}$ and

$$\mathrm{ATE} = \mathrm{ATE}_1 - \mathrm{ATE}_0$$

However, in an observational dataset, the strongly ignorable assumption is usually not known to hold without the randomization step. But if the assumption holds in a non-randomized dataset and V is a discrete variable with M levels, ATE can be estimated. For instance, we can divide the data into M subgroups according to the level of V and estimate ATE_m in each level m, then the overall ATE can be written as the weighted average of ATE_m.

Although we can do the same stratification when V is a high dimensional variable, the number of subclasses will explode. Suppose V consists of p binary variables, the number of subclasses will be 2^p. Therefore, the cases allocated to each stratum will be reduced. Most strata will not contain both treated and control cases [62]. Subclassification based on the propensity score is a dimensionality reduction approach to avoid that.

3.1.2 Using propensity score adjustment to estimate ATE in observational dataset

Defined by Rosenbaum and Rubin [62], the propensity score $e(x_i)$ is the conditional probability of the intended treatment $T = 1$ given the evidence $X = x_i$, i.e,

$$e(x_i) = P(T = 1 \mid X = x_i) = \mathrm{E}(T \mid X = x_i)$$

The treatment variable T is Bernoulli distributed, Therefore, by definition $P(T = 1 \mid X = x)$ equals to the mean or expectation, $\mathrm{E}(T \mid x)$. In observational studies, the variables T and X are not independent and the true propensity score is generally unknown. A practical and common way to estimate the propensity score $P(T = 1 \mid X = x)$ is by logistic regression:

$$e(x) = e(x_1, \ldots, x_m) = \left(1 + \exp - \sum_{j=0}^{m} a_j x_j\right)^{-1} \tag{3}$$

Where a_j $(j = 0, \ldots, n)$ are the coefficients computed by fitting the equation to the data D, a_0 is the baseline and $x_0 = 1$.

A propensity score $e(X)$ is a type of balancing score $b(X)$, which is defined as a function that has the property

$$X \perp\!\!\!\perp_p T \mid b(X)$$

with $b(X) = X$ is the finest balancing score and $e(X) = f(b(X))$ is the coarsest balancing score for some function f, and we have the following theorems [62]:

Theorem 1. *Let b(X) be a function of the random variable X, then b(X) is a balancing score, i.e.,*

$$X \perp\!\!\!\perp_p T \mid b(X)$$

iff b(X) is finer than the propensity score e(X) if e(X) = f(b(X)) for some function f.

The most important point to prove theorem 1 is that $b(X)$ does not contain more information than X, therefore,

$$P(T \mid b(X), X) = P(T \mid X))$$

and by the definition of balancing score $b(X)$, we have

$$P(T \mid b(X), X) = P(T \mid b(X))$$

Therefore,

$$P(T \mid b(X)) = P(T \mid b(X), X) = P(T \mid X))$$

It implies that allocating treatment based on patient characteristics X or its function $b(X)$ will lead to the same results.

Theorem 2. *If treatment allocation is strongly ignorable given X, it is strongly ignorable given any balancing score b(X) [62], i.e,*

$$Y \perp\!\!\!\perp_p T \mid X \;\; \textit{given that } 0 < P(T = 1 \mid X) < 1$$

similar for

$$Y \perp\!\!\!\perp_p T \mid b(X)$$

As propensity score $e(X)$ is a type of balancing score $b(X)$, theorem 2 is also hold for $e(X)$. Therefore, adjusted for a balancing score $b(X)$ can produce an unbiased estimate of ATE.

After the computation of propensity score, cases are stratified into 5 equal-size subgroups. The cutting points are quintiles of the estimated propensity score. In the paper of [22], they concluded that stratifying the cases into 5 groups is sufficient to remove 90% of the bias caused by confounders. A larger number of subgroups may reduce the number of cases in each subgroup and increase variance, while a smaller number of subgroups would be insufficient to remove the bias.

3.2 Cochran-Mantel-Haenszel (CMH) test

Begin with the simplest case. Suppose two binary variables A and B are stratified based on the third variable C, which has r categories. The stratification will result in r contingency tables of 2×2. The i-th contingency table can be written as

	$B=1$	$B=0$
$A=1$	a_i	b_i
$A=0$	c_i	d_i

In the paper of Richard Landis *et al.* [45], the null hypothesis H_0 of the CMH test is expressed as follows. H_0: For each of the i-th table ($1 \le i \le r$), the variable is distributed at random with respect to the another variable [45], i.e., no association can be found between the two variables (after taking the confounders into account). The CMH test actually transformed the problem into a sampling problem.

Conditional on the row and column totals, the distribution of the frequencies of a_i follow the *hypergeometric distribution*. So it is sufficient to compare the observed and expected counts in one cell per table [37]. To construct the CMH test statistic, one of the four cells (normally a_i) from each stratum is taken as pivotal. Choosing other cells will not affect the value of statistics or the conclusion.

According to the properties of the hypergeometric distribution, the expectation E of a_i is

$$\mathrm{E}(a_i) = \frac{(a_i + b_i)(a_i + c_i)}{a_i + b_i + c_i + d_i}$$

the variance V can be written as

$$V(a_i) = \frac{(a_i + b_i)(a_i + c_i)(b_i + c_i)(b_i + d_i)}{(a_i + b_i + c_i + d_i)^2(a_i + b_i + c_i + d_i - 1)}$$

The statistics is constructed as follows:

$$X_0^2 = \frac{(|\sum_{i=1}^{r} a_i - \sum_{i=1}^{r} \mathrm{E}(a_i)| - \frac{1}{2})^2}{\sum_{i=1}^{r} \mathrm{V}(a_i)} \tag{4}$$

The statistics X_0^2 follows a χ^2 distribution with 1 degree of freedom.

For a more complex case, when the contingency table has a $2 \times n$ shape, where n is the number of categories of variable B (n>2). If the number of strata is still r, the i-th contingency table can be written in matrix form as:

$$\begin{pmatrix} a_{11i} & a_{12i} & a_{13i} & \dots & a_{1ni} \\ a_{21i} & a_{22i} & a_{23i} & \dots & a_{2ni} \end{pmatrix}$$

To estimate the association between $[a_{11i}, a_{12i}, a_{13i}, \dots, a_{1ni}]$ and $[a_{21i}, a_{22i}, a_{23i}, \dots, a_{2ni}]$ across the r strata ($1 \le i \le r$), the null hypothesis is the same, and the pivotal cells for the i-th strata a_i is a $(n-1)$ row vector (n>2):

$$a_i^T = (a_{11i}, a_{12i}, a_{13i}, \ldots, a_{1(n-1)i})$$

Where a_i^T is the transpose of a_i, which is assumed to follow the *multivariate hypergeometric distribution*. If we let

$$r_i^T = (a_{\cdot 1i}, a_{\cdot 2i}, a_{\cdot 3i}, \ldots, a_{\cdot(n-1)i})$$

be the column sum of the i-th contingency table,

$$(a_{1\cdot i}, a_{2\cdot i})$$

be the row sum of the i-th contingency table, and $a_{\cdot\cdot i}$ be the sum of the i-th contingency table. According to the properties of *multivariate hypergeometric distribution*, the expectation of a_i can be written as

$$\mathrm{E}(a_i) = \frac{a_{1\cdot i} r_i}{a_{\cdot\cdot i}}$$

and the covariance matrix v_i can be written as

$$v_i = a_{1\cdot i} a_{2\cdot i} \frac{a_{\cdot\cdot i}\mathrm{diag}(r_i) - r_i r_i^T}{(a_{\cdot\cdot i} - 1)a_{\cdot\cdot i}^2}$$

with $\mathrm{diag}(r_i)$ is an $(n-1) \times (n-1)$ diagonal matrix with element r_i. The statistics can be written as

$$(a - e)^T v^{-1} (a - e) \tag{5}$$

Where $a = \sum_{i=1}^{r} a_i$, $e = \sum_{i=1}^{r} e_i$ and $v = \sum_{i=1}^{r} v_i$

Statistics 5 follows a χ^2 distribution with $(n-1)$ degrees of freedom.

Note that CMH test can also be used for the most general case when the contingency table has a $m \times n$ shape $(m, n > 2)$, where m and n are the number of categories of variable A and B, respectively. In this situation, the statistics is much more complex and follows a χ^2 distribution with $(m-1)(n-1)$ degree of freedom.

3.3 Survival analysis

In survival analysis, The outcome is usually known as the event of interest. The survival time, which is always positive, reflects the time from the beginning of observation to the occurrence of the event of interest. Dataset can be classified into two categories according to the outcome of populations. They are:

- **Complete data** The specific time of the event (the event of interest) is known and recorded at the time of follow-up.
- **Censored data** The event did not (yet) happen at time of follow-up, which means the actual survival time is not the same as the recorded value. There are three types of censored data:
  - *Right censored* is the most common type of censored data. In right censored data, survival time is underestimated compared with the record value, but it is unknown by how much [5]. Reasons for this can be: the loss/refusal of follow up or the event of interest does not happen at the end of the experiment.
  - *Left censored* means survival time is overestimated compared with the record value, but it is unknown by how much [5]. In this way, the event of interest happened before the time of record.

– *Interval censored* Similar, interval censored means the event of interest happens between a period, but it is impossible to figure out the exact time. One common example that will cause interval censored is the periodic follow-up [42]. Researchers only know the event of interest happened between two follow-ups, but the actual time is unknown. Right censored data is a special case of interval-censored data [42].

We will focus on survival analysis for right censored data. Suppose the event of interest is death. Let $T \geq 0$ be a non-negative variable which has density function $f(t)$ and distribution function $F(t)$, the *survival function* $S(t)$ can be written as [51]

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^{\infty} f(x)dx \tag{6}$$

which represents the probability being alive by duration t.

Mortality at any time t (instantaneous death rate) m_t is the slope of survival function $S(t)$ at time t, denoted as

$$m_t = \lim_{\delta t \to 0}(S(t) - S(t + \delta t)) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t)}{\delta t} \tag{7}$$

The hazard function $h(t)$ at any time t is the mortality m_t divided by patients at risk at time t

$$h(t) = \frac{m_t}{S(t)} = \frac{\lim_{\delta t \to 0}(S(t) - S(t + \delta t))}{S(t)} \tag{8}$$

Epidemiologically, the hazard rate $h(t)$ describes the instantaneous death rate of participant who survived at time t. In survival analysis, different models will be constructed to represent $h(t)$ as a combination of time t and some functions of covariates x.

3.3.1 Kaplan-Meier estimator

The Kaplan-Meier estimator is an nonparametric approach to visualize the survival probability. Let d_i be the number of deaths at time t_i, n_i be the number of cases *at risk* at time t_i, where *at risk* means individuals being alive at time t_i [21]. The cumulative survival rate $S(t_i)$ can be written as

$$S(t_i) = S(t_{i-1})\left(1 - \frac{d_i}{n_i}\right) \tag{9}$$

Where $t_0 = 0$, $S(0) = 1$. The variance of the survival function is calculated as:

$$V(S(t)) = S(t)^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

and the pointwise 95% confidence interval of the survival function is $S(t) \pm 1.96 \times \sqrt{V(S(t))}$.

3.3.2 Cox proportional-hazards model

In the Cox model, the distribution of the hazard function $h(t \mid x)$ is represented as a combination of the baseline hazard at time t and a function of the covariates x:

$$h(t \mid x) = h_0(t)\exp(b_1x_1 + b_2x_2 + \cdots + b_px_p) \tag{10}$$

where $x_1, x_2, \ldots, x_p$ represents the covariates, $b_1, b_2, \ldots, b_p$ are coefficients that measure the effect size of the covariates and $h_0(t)$ is the baseline hazard function; it is the risk function when all covariates equal to 0.

Moving $h_0(t)$ to the left, equation 10 can be written as:

$$\frac{h(t \mid x)}{h_0(t)} = \exp(b_1 x_1 + b_2 x_2 + \cdots + b_p x_p) \tag{11}$$

To comparing the hazards between different groups. The *hazard ratio* (HR) is introduced. It is the effect estimate of Cox model [66]. The HR of two participants i, j is the ratio of their hazard functions

$$\mathrm{HR} = \frac{h_i(t \mid x)}{h_j(t \mid x)} = \frac{h_0(t)\exp(b_1 x_{11} + b_2 x_2 + \cdots + b_p x_p)}{h_0(t)\exp(b_1 x_{12} + b_2 x_2 + \cdots + b_p x_p} \tag{12}$$

If all the covariates $x_2, x_3, \ldots x_n$ are the same except for x_1, they can be eliminated. Then we have HR of covariate x_1 to estimate the association between x_1 and mortality risk. HR > 1 indicates a increasing risk of death. On the contrary, HR < 1 indicates a decreasing risk of death.

There are three basic assumptions in the model. First, survival time of every participant should be independent. Second, the relationship between hazard and predictors should be multiplicative. And most importantly, the proportional Hazard assumption: for each of the covariate, HR should stay constant over time. In other words, if the death risk of patient k in treatment group is twice that of patient k' in control group at time t, the death risk of k at any other time should be twice of k' as well. This makes the survival curve of different groups can never cross.

3.4 Bayesian networks

Bayesian network (BN) is a type of probabilistic graphical model (PGM) [17], which combines principles from probability theory and graph theory. Formally, a BN is defined as a pair $B = (G, P)$, with $G = (V(G), A(G))$ a directed acyclic graph, where $V(G)$ is a set of nodes corresponding to random variables, and $A(G) \subseteq V(G) \times V(G)$ a set of directed edges or arcs, representing dependence and independence information amongst variables. The network structure G models the relations among multiple variables, and the relations are quantified by a *joint probability distribution* (JPD) P. Though usually used for representing causality, the BN model is not necessarily causal. Mathematically, the dependency $X \to Y$ is equivalent to $Y \to X$ for two random variables X and Y. But the direction of the arrow can not be reversed in causal inference, which aims to analyze the response of the *effect* as the *cause* changed.

3.4.1 Basics of probability theory

The joint probability distribution P follows the basic rules of probability theory, except that the graph structure G enforces particular independence constraints on P, as discussed below.

- **Conditional independence**

  Let $X, Y, Z \subseteq V(G)$ be a set of variables. X and Y are said to be conditionally independent given Z:

  $$X \perp\!\!\!\perp_p Y \mid Z$$

  if and only if

  $$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

  or equivalently,

  $$P(X \mid Y, Z) = P(X \mid Z)$$

- **Chain rule**

  The conditional probability of X_1 given $X_2, X_3, \ldots, X_n$ is defined as:

$$P(X_1 \mid X_2, \ldots, X_n) = \frac{P(X_1, X_2, \ldots, X_n)}{P(X_2, \ldots, X_n)} \tag{13}$$

  which can be also written as

$$P(X_1, X_2, \ldots, X_n) = P(X_1 | X_2, \ldots, X_n) P(X_2, \ldots, X_n) \tag{14}$$

  This formula (14) can be applied recursively, giving rise to the *chain rule*:

$$\begin{aligned} P(X_1, X_2, \ldots, X_n) &= P(X_1 \mid X_2, \ldots, X_n) P(X_2 \mid X_3, \ldots, X_n) \cdots P(X_n) & (15) \\ &= \prod_{i=1}^{n} P(X_i \mid X_{i+1}, \ldots, X_n) & (16) \end{aligned}$$

  In a BN model, The *joint probability distribution* (JPD) P can be expressed by the chain rule. When taking into account the conditional independence, each individual variable $X_v \in V(G)$ only needs to be conditioned on its associated parents $X_{\pi(v)}$. So the JPD can be written as:

$$P(X_{v(G)}) = \prod_{v \in V(G)} P(X_v \mid X_{\pi(v)})$$

  Specifically, if X_v has no parent, $P(X_v \mid X_{\pi(v)}) = P(X_v)$

  Finally, if all variables X_v are mutually independent, we have that

$$P(x_v \mid X_{\pi(v)}) = P(X_v)$$

  hence the joint probability distribution equals the product of the individual probabilities

$$P(X_{v(G)})) = \prod_{v \in V(G)} P(X_v)$$

  However, this is a completely uninteresting situation.

Take figure 2 as an simple example of the BN model again. In this model,

$$V(G) = \{Difficulty, Intelligence, Grade, SAT, Letter\}$$

and

$$E(G) = \{Difficulty \to Grade, Intelligence \to Grade, Intelligence \to SAT, Grade \to Letter\}$$

and the joint probability distribution is defined as:

$$\begin{aligned} &P(Difficulty, Intelligence, Grade, SAT, Letter) = \\ &\quad P(Difficulty) P(Intelligence) P(Grade \mid Difficulty, Intelligence) \\ &\qquad \cdot P(SAT \mid Intelligence) P(Letter \mid Grade) \end{aligned} \tag{17}$$

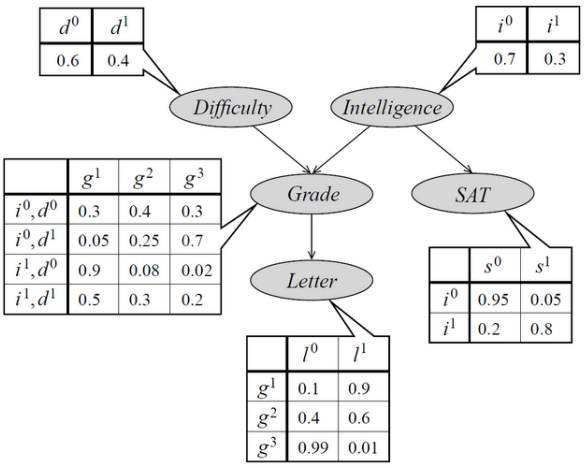


Figure 2: An example Bayesian network. This network is a reproduction of student network in [44]

3.4.2 Basics of causal diagrams

- **D-separation** is a criterion to identify conditional independence from a causal graph. The independence that follows from the graph structure G is written as $\perp\!\!\!\perp_G$. Given three nodes A, B and C with associated variables X_A, X_B and X_C, we have that if

$$A \perp\!\!\!\perp_G B \mid C$$

then

$$X_A \perp\!\!\!\perp_p X_B \mid X_C$$

But the other way around does not always hold.

In the causal diagram, there are three kind of basic junctions:

(1) ***mediation*: $(A \to B \to C)$**: In this junction, B is referred to as *mediator* that transmit the effect of A to C [58]. When the state of B is known, B *blocks* A and C, thus no information could pass from A to C. An example of this junction in figure 2 is *Difficulty* $\to$ *Grade* $\to$ *Letter*. The difficulty of a course by itself does not influence the probability of getting a reference letter, but it does have an effect on the score. Therefore, if the student's grade is not known, the difficulty of the course will affect the probability of getting a reference letter by changing the state of grade. However, if the grade is known, the probability of obtaining a reference letter will only depend on the grade rather than course difficulty. Formally, it can be expressed by conditional dependency: $A \perp\!\!\!\perp_G C \mid B$

(2) ***fork*: $(A \leftarrow B \to C)$**: In this junction, B, the common cause of A and C is also known as a confounder. A confounder B can make A and C statistically correlated even if there are no direct causal relations between them [58]. In figure 2, $Grade \leftarrow Intelligence \to SAT$ is a fork. Sometimes, experience tells us that the grade and the SAT score are positively correlated. But this is not a causal relationship, i.e., a higher grade will not cause a higher SAT score. Instead, both of the two variables can be explained by *intelligence*, the third variables. When the state of intelligence is decided, the spurious correlation between grade and SAT score will be eliminated. So the conditionally independent relation here is $A \perp\!\!\!\perp_G C \mid B$

(3) ***Collider*: $(A \to B \leftarrow C)$**: In this junction, B is a common effect of A and B. If A and C are independent in the beginning, conditioning on B will make them dependent [58]. One

of the example of this junction in figure 2 is $Difficulty \rightarrow Grade \leftarrow Intelligence$. Both the course difficulty and the student's intelligence will have an effect on the student's grade. If the grade is not given, the course difficulty and the intelligence of the student seems to be unrelated. But when the grade of the student is known, the difficulty of the course and the student's intelligence will be correlated. i.e., Given two students with the same grade, the one who took a more difficult course tend to be smarter than the other one. The conditionally independent relation here is $A \perp\!\!\!\perp_G C \mid B$

We can explore causal relations and non-causal relations using a causal diagram. Suppose we want to investigate the causal relation from *Grade* to *Letter* in figure 2. In this case, we don't care how *difficulty* and *intelligence* are associated with getting a letter. In contrast, these potential statistically correlations ($Difficulty \rightarrow Grade \rightarrow Letter$, $Intelligence \rightarrow Grade \rightarrow Letter$) are what we want to get rid of. Therefore, we need to set the evidence $grade = g$. After that, edges pointing towards *Grade*, i.e., $Difficulty \rightarrow Grade$, $Intelligence \rightarrow Grade$ need to be removed as well.

On the other hand, statistically associations can be allowed if we want to explore non-causal relations. Thus all we need to do is to set evidence $Grade = g$, then check how the changes in *Grade* will influence the probability of getting a letter.

3.4.3 Structure learning and parameter learning for Bayesian networks

Learning a BN model from a dataset consists of two steps. The first step is called structure learning, which aims to identify the graph structure of the network. The next step is referred to as parameter learning. It is to estimate the conditional probability distribution of the nodes in the BN model. Both of the two steps can be done manually with the domain knowledge of an expert [53]. If researchers have not done that, some algorithms are available to learn the structure as well.

3.4.4 Structure learning algorithm

Strategies of structure learning There are two types of methods for structural learning algorithm: constrained-based methods and score-based methods, where:

- **Constrained-based methods** focus on identifying DAG structure that can represent a set of conditional independence in the best way.
- **Score-based methods** treat structure learning as an optimization problem. Score-based approaches aim to search and select the best BN that fits or explains the dataset. The searching step is usually achieved by general heuristic optimization algorithms like hill climbing or tabu search. The goodness of fit can be reflected with a network score.

Constraint-based approaches may be more efficient when given a large sample size as the power of identifying conditional independence relationships may be limited by small sample sizes. Also, constraint-based approaches define the direction of edges by conditional independence relations. Some edges may remain undirected in this step [67]. So score-based approaches are more preferred in general, especially for a small or noisy dataset. For the reasons discussed above, we chose the score-based approach for structure learning.

Network scores The network score is a criterion used for model selection in score-based approach. The accuracy of the model can be continuously improved when the model gets closer to fit the training set. But raising accuracy may also lead to overfitting. It may result in a much more complex model that is less generalized. Therefore, network scores are introduced to find a balance between the accuracy and complexity of the model. Many information criteria have been developed, two of them are commonly used: The *Akaike Information Criterion* (AIC) and *Bayesian Information Criterion* (BIC). Closely related, both of them work by introducing a penalty term to balance model's accuracy and complexity.

- **Akaike Information Criterion (AIC)** Developed by Hiruji Akaike [10] is defined as:

$$AIC = 2k - 2\ln(L) \tag{18}$$

  Where k represents the number of parameters in the model and L represents the maximum value of the likelihood function. It is based on the concept of information entropy.

- **Bayesian Information Guidelines (BIC)** was developed by Gideon Schwarz [64]. The formula of BIC is

$$BIC = k\ln(n) - 2\ln(L) \tag{19}$$

  Where L and k are similarly defined as AIC, and n denotes the number of samples in the dataset. Slightly different from AIC, it takes the sample size into account as well.

Formula 18 and 19 demonstrate that model with lower value of AIC/BIC is preferred.

Tabu search Tabu search algorithm [33] was applied to learn the structure of the network. It is a variant of hill-climbing algorithm [46]. The tabulist in Tabu search algorithm stores the shortly visited solutions, which are forbidden to access in a short time. But as the iteration goes on, solutions in a tabulist can be popped up and accessed again. This algorithm can prevent repeated search around the local optimum and perform more extensive exploration [18]. Pseudocode of Tabu search is shown in algorithm 1.

Algorithm 1 Tabu search algorithm

```
Generate an initial network structure G randomly
G_best ← G
Calculate score_G
i = 0
Tabulist ← φ
while Score_G decreasing do
    Generate a modified network G*
    Calculate score_G*
    if G* ∉ Tabulist then
        if score_G* <score_G then
            G_best ← G*
        end if
    end if
    i = i + 1
    Update Tabulist
end while
return G_best
```

Model averaging Model averaging is an approach to avoid a local optimum. To start with, Tabu search algorithm will be repeated n times to build n (globally suboptimal) networks [53], among which the frequency of each edge is regarded as the *strength* of the edge. The averaged network, which is believed to be more robust, contains only the edges that have a higher strength than a threshold t.

3.4.5 Parameter learning algorithm

After learning the structure of BN from the data, parameter learning will be applied to estimate and update the conditional probability distributions of each node. *Expectation-maximization* (EM) algorithm is the most commonly used algorithm for parameter learning

when the data is incomplete. This algorithm aims at computing maximum-likelihood of parameters by iteration. It alternates between *E-step*: guess a probability distribution given the current model [27] and calculate the expectation of the log-likelihood and *M-step*: update the model parameter that can increase the expected log-likelihood until parameter converge [6]. Specifically, it can deal with incomplete data or latent variables [26]. The missing value in the dataset will be filled with possible values when the EM algorithm is running.

3.5 BN model validation

3.5.1 Sensitivity analysis

The sensitivity analysis is performed using *Receiver Operating Characteristics* (ROC) and *Area-under-the-curve* (AUC). In a binary classification model, the 2×2 confusion matrix consists of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN), as shown in figure 3.

		Actual class	
		Positive	Negative
Predicted class	Positive	TP	FP
	Negative	FN	TN

Figure 3: Example of a confusion matrix

The *true positive rate* (TPR) is defined as

$$TPR = \frac{TP}{TP + FN}$$

which is also referred to as *sensitivity* and the *false positive rate* (FPR) is defined as

$$FPR = \frac{FP}{FP + TN}$$

Both TPR and FPR have a value between 0 and 1. But they will change when varying the classifier threshold. The ROC curve is used to show TPR and FPR at different classifier thresholds. It is summarized by AUC, which is a value from 0 to 1. AUC indicates how well the measure of separability a classifier has.

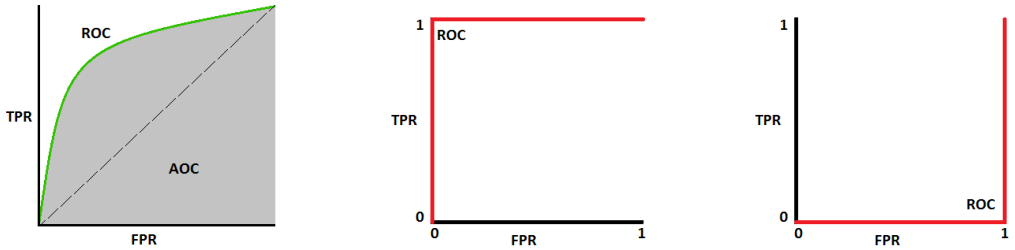
Shown in figure 4(a) [55] is an example of the ROC curve. The x-axis is FPR and the y-axis is the TPR. ROC is the green curve and AUC is the grey area under ROC. In a perfect classifier, TPR = 1, FPR = 0 and AUC = 1, as shown in figure 4(b). A random guess will lead to a point along the dotted line with slope 1 (also AUC = 0.5). Therefore, a curve close to the dotted line indicates the classifier fails to separate the two classes. Generally speaking, points above the dotted line represent good classification results as they are better than a random guess. In contrast, points that below the dotted line are bad classification. For most cases, a higher AUC indicates a better predictor, but this is not always true. If most of the points are below the line and the value of AUC is quite low, it is a good classifier to some extent because we can reverse the predicted results to obtain a good prediction as well, as shown in figure 4(c).

3.5.2 Calibration plot

In a calibration plot, the x-axis represents the predicted probability and the y-axis represents the observed conditional probability.

Let $C = i$ be the i-th state of class variable C and x are the patient's characteristics that have been observed, the observed conditional probability of $C = i$ can be written as

$$P_{obs}(C = i \mid x) = \frac{P(C = i, x)}{P(x)}$$



(a) (b) (c)

Figure 4: *left:* an example of ROC curve; *middle:* a perfect classifier (AUC = 1); *right:* AUC = 0 but it is still a good predictor as we can reverse the predicted value [55]

Each dot (P_{pred}, P_{obs}) in this plot represents the predicted and observed probabilities that a subpopulation with certain characteristics x classified into i–th state of class variable C. A line with a slope of 1 will denote a perfect prediction (predicted probability = observed probability). Therefore, dots closer to the line are preferred.

4 Experiments

4.1 Data

Prostate cancer data were obtained from the Netherlands Cancer Registry (NCR), which covers the information of more than 95% of Dutch cancer patients [7]. In the dataset, patients diagnosed with prostate cancer between 2005 and 2014 were included. It consists of 104,324 cases. After removing the cases with missing values, the number of cases were 101,467. To reproduce Giordano's second experiment, which was a reproduction of Wong's paper [69], we also read Wong's paper [69] for more detail. Wong *et al.* investigated the association between active treatment and survival for localized prostate cancer patients [69].

The survival rate of the general male population for men aged 65-80 from 2005 to 2014 in the Netherlands was found on the website of Statistics Netherlands (CBS) [1], which is an independent government-funded organization established in 1899 responsible for collecting and maintaining population statistics of the Netherlands. People can acquire the data for free by viewing the CBS website and clicking on the filter button. The explanation of the data is also available on this website.

4.2 Preprocessing

The population being studied included men aged between 65 and 80 years and diagnosed with localized prostate cancer between 2005 and 2014. They received either active treatment or observation (non-active treatment).

Active treatment was defined as being either radical prostatectomy, radiation implants, or external-beam radiation therapy. *Non-active treatment* was defined as consisting of only observing a patient, i.e., no treatment. Although one of the therapies defined as active treatment was surgery (radical prostatectomy), we still chose to use clinical staging (cT, cN and cM) in our research rather than pathological staging (pT, pN and pM), since pathological staging was usually not available if a patient was only observed, although pathological staging when it is done offers usually more reliable information. The definition of localized prostate cancer was: tumor stage cT = cT1 or cT2, well to moderately differentiated (Gleason score of 2-7), cN = cN0 (no lymph nodes affected) and cM = cM0 (no metastasis).

In the paper of Wong [69], patients that died within one year of diagnosis were excluded. This was to remove patients died of other causes considering the low mortality rate of localized prostate cancer (the 5-year mortality rate of localized prostate cancer is 10% [11]).

Table 1 shows detailed information about the exclusion criteria and the number of samples left after each step. In the end, The dataset contained 8 features and 8480 cases, among which 6197 patients received active treatment and 2283 patients were observed only.

Next, we mapped some of the variables as mentioned in Wong's paper [69], Firstly, they used the differentiation grade provided by The Surveillance, Epidemiology, and End Results (SEER) Program, which is an authoritative source for cancer statistics run by the United States [8]. This differentiation grade mapping system is different from what normally used in the Netherlands: it is decided from the *Gleason score*. *Gleason score* of 2-4, 5-7, 8-10 were classified as well-differentiated, moderately differentiated and poorly differentiated, respectively. We did the same mapping to our dataset. The mapped value was recorded in a new column tumor_grade, where "0" denoted well-differentiated and "1" denoted moderately-differentiated. Poorly-differentiated cases were not included since we excluded cases with a score of 8-10 previously. In Wong's paper [69], they also reclassified the cT into two categories: tumors that classified as T2a or lower were one category. Tumors that are classified as T2b and T2c were classified as another category. We followed their description and created a new column tumor_stage, in which "0" denoted cT $\leq$ cT2a. "1" denoted cT $\in$ {cT2b, cT2c}. Furthermore, they discretized the age at diagnosis into five bins: [65,67], [68,70],[71,73],[74,77] and [78,80]. Finally, the discretization was applied to feature PSA, which resulted in four

Table 1: Selection criteria for localized prostate cancer

Criteria	Number of Cases Left
age: 65 - 80	61539
Gleason score : 2 - 7	44414
tumor stage cT are classified as cT1 or cT2	36341
nonmetastatic (cM = cM0)	28050
not a lymph node-positive tumor (cN = cN0)	12077
Died after one year of diagnosis	11833
treatment that fit in criteria (active treatment and observation)	8480

Table 2: The variables in the dataset after preprocessing

Variable	Values	Descriptions
tumor stage	≤ T2a	size and extension of tumor
	T2b and T2c	
tumor grade (differentiation)	well-differentiated (Gleason score of 2-4)	differentiation grade of tumor
	moderately-differentiated (Gleason score of 5-7)	
incjr(incidence year)	integer	year of diagnosis
age	65-67	age at diagnosis
	68-70	
	71-73	
	74-77	
	78-80	
PSA value	[0,4)	PSA value when diagnosis
	[4,10)	
	[10,20)	
	[20,)	
treatment	active treatment	indicating if active treatment was applied
	observation (non-active treatment)	
vitstat	0	indicating whether patient were dead at the end of the follow-up
	1	
vitfup	integer	time of following up

bins: [0,4), [4,10), [10,20) and [20, $+\infty$). The cut-off points were decided according to the 7th edition of AJCC manual [29], which has been mentioned in the introduction section.

After the selection process, the features were discretized. The features left are listed in table 2, together with short descriptions and their variable's domains. The distribution of the variables in the *active treatment group* and *non-active treatment group* are visualized in figure 5.

4.3 Propensity score adjustment

After preprocessing, we applied Pearson's chi-square test for statistical independence between treatment allocation and the other covariates. Next, propensity scores were calculated using logistic regression. We followed the suggestion of Jill *et al.* [9] to include all the variables that related to the outcome even though they were unrelated to the exposure (treatment). Therefore, incidence year, age, tumor stage, tumor grade and PSA were all included to compute the propensity score. Finally, the dataset was subdivided into five strata according to the quintiles of the computed propensity score. The stratum that each case belonged to was stored in a new column *quintiles*. The CMH test was applied to check whether the baseline characteristics were similar within the strata.

4.4 Cox model

The last step of reproducing Giordano's experiment was the Cox model. This was done using the Python package *lifelines* [25]. Two Cox model were constructed to make a comparison before and after the adjustment of propensity score. (With or without the column *quintiles*

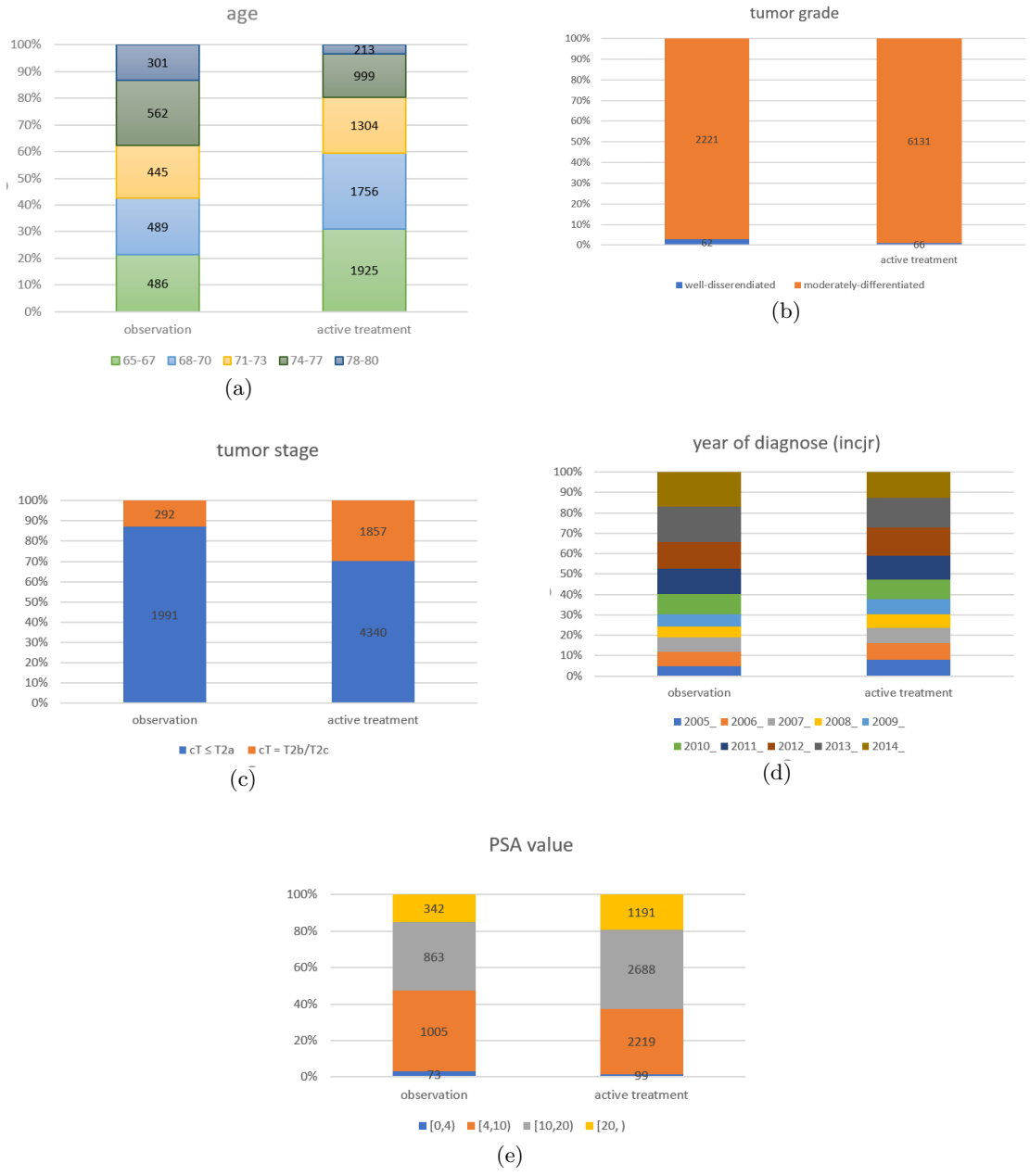


Figure 5: Baseline characteristics of patients: age, tumor grade, tumor stage, incidence year and PSA. For all the variables, distinct difference can be found between active treatment group and non-active treatment group (P <0.05) according to the Pearson's chi-square test of statistically independence between treatment allocation and the variables as shown in table 4

as one of the input variables). Additionally, reproducing the result of Wong's paper [69], we constructed another five Cox models, one for each age group.

4.5 Bayesian network

The BN model were produced using the *bnlearn R* package [65]. All the variables have been discretized except for survival time in preprocessing. As the discrete network has better interpretability, We then mapped the survival time into three new features: survival_less_than_2_years, survive_GE_2_years (survival greater than 2 years or 2-year survival) and survive_GE_5_years (survival greater than 5 years or 5-year survival) before applying structure learning to construct a discrete BN model. These features would be the nodes indicating survival in the BN model. Values of them were binary, where "0" means "no" and "1" means "yes". Noted that for censored data, missing values, i.e., *NA* would be introduced in the mapping step. For instance, if a patient was last observed to be alive after four years of diagnosis, value of the feature survive_GE_2_years would be "1" and value of survive_GE_5_years would be missing. The same issue would not happen for non-censored data, since the survival time was known.

In the BN model, there was no node survive_GE_10years as no enough data were available. If this node had been added to the network, it would introduce too many missing values: for all the patients who did not die at the end of follow-up and whose time of following up was less than 10 years, the value of the node would be missing. Nearly three-quarters of the cases were as this type.

The structure of the BN model was identified by the hybrid approach, i.e., a combination of structure learning and prior knowledge. First, our knowledge and our understanding of this experiment helped to define a blacklist. Edges in the blacklist could never be included in the model. They were: **(1)**. Edges started from the nodes that indicating survival to the rest of the nodes and **(2)**. edges from a longer survival to a shorter one. For example, edge survive_GE_5_years → survive_GE_2_years was forbidden, but the reversed one was not. For **(1)**, the nodes indicating survival were the outcome of the model. We would like to investigate the causality from other variables to survival, but not the other way around.

Second, model averaging with Tabu search algorithm was applied to learn the structure of the network from the data. The number of candidate networks was set to 2000 to use the methods of model averaging. The network score we used in Tabu search was *BIC*.

After the structure of the BN model have been identified by model averaging, some edges were added manually. This was the second application of prior knowledge. We have tried to add the same edges before structure learning, but it led to a worse AUC value when we evaluate the BN model. Therefore edges were added after structure learning. By adding an edge $A \rightarrow B$, we assume there was a causal relation from A to B based on our understanding for prostate cancer. Edges that we added were listed in table 3.

Once the structure of the network has been decided, the EM algorithm was used for parameter learning, which resulted in a set of conditional probability distribution tables, one for each node. Five-fold cross-validation and calibration were then applied to evaluate the performance of the network. To do the calibration, 80% of the cases were selected randomly as the training set, and the rest of 20% cases were the test set. In the training set, missing values were allowed as they would be filled by EM algorithm. But missing values were not allowed in the test set. So the cases with missing values were removed from the test set.

Finally, to do the causal and non-causal reasoning, our BN model was imported into *SamIam* [24], which provides a graphical user interface (GUI) for uses to visualize, edit and analyse the BN model. In causal reasoning, edges pointed toward to treatment were removed and the evidence of treatment $= i$ ($i \in \{0, 1\}$) was set to investigate the causal relations between treatment and survival. In non-causal reasoning, no edges would be removed, only the evidence treatment $= i$ ($i \in \{0, 1\}$) was set.

Table 3: Edges added to the BN model that obtained by model averaging

From	To
tumor_stage	survive_GE_2_years
tumor_stage	survive_GE_5_years
tumor_stage	survival_less_than_2 years
PSA	survive_GE_2_years
PSA	survive_GE_5_years
PSA	survival_less_than_2 years
PSA	treatment group
treatment	survive_GE_2_years
PSA	tumor_stage

4.6 Visualization and comparison between survival rates of general male population, active treatment group and non-active treatment group

In the visualization step, to make it comparable, the weighted survival rates of general male residents adjusted for age and year were used. As the patients who died within the first year of diagnosis were excluded in preprocessing, we also computed the one to n year survival rate of general male population in the Netherlands $P(1\_to\_n\_year\_survival)$ as

$$\begin{aligned} P(1\_to\_n\_year\_survival) &= P(0\_to\_n\_year\_survival \mid 1\_year\_survival) \\ &= \frac{P(0\_to\_n\_years\_survival)}{P(1\_year\_survival)} \end{aligned} \tag{20}$$

Then the one to n year survival rate of male resident in the Netherlands was visualized, together with the Kaplan-Meier curve of active treatment group and non-active treatment group.

Table 4: Tests for independence between treatment allocation and other variables

Name of Variate	p-value before adjustment *	p-value after adjustment **
year of diagnosis	<0.005	0.99
tumor_stage	<0.005	0.28
age	<0.005	<0.05
tumor_grade	<0.005	0.30
PSA	<0.005	0.66

* Pearson's chi-square test
** Cochran Mantel-Haenszel Chi-square test

5 Results

5.1 Description of patient's baseline characteristics

Figure 5 shows the baseline characteristics of patient after preprocessing. For all the variables (age, tumor stage,tumor grade, incidence year and PSA value), distinct difference can be found between active treatment and observation group (p-value <0.05 as shown in table 4). Generally speaking, patients in treatment group were younger, had a worse tumor stage or a higher PSA value compared with patients in non-active treatment group.

5.2 Propensity score adjustment

Table 5 shows the quintiles of propensity score and the distribution of every variable. It indicates that patients who were younger, had a higher PSA value or worse tumor stage were more likely to receive active treatment as they are in higher propensity score quintile compared with patients who were older, had a lower PSA value or better tumor stage.

Table 4 shows results of CMH tests between treatment allocation and other covariates. When stratified by quintile of propensity score, the p-value for covariates incidence year, tumor_stage, tumor_grade and PSA were 0.99, 0.28, 0.30 and 0.66, respectively. Therefore, these four covariates were balanced within strata after propensity score adjustment. However, the covariate age could not be balanced even after the adjustment (p-values <0.05).

5.3 Cox model

The association between variables in the dataset and survival is shown in table 6 and 7, respectively. Both of the Cox models were statistically significant (p-value <0.05) and they were similar: the variables in the two models shared similar parameters. The two models only differed in whether or not the covariate age met the proportional hazard assumption, i.e., if the HR for age stayed constant over time. Before the adjustment, HR of all the covariates stayed constant except for age, and the adjustment made the HR for age stayed constant.

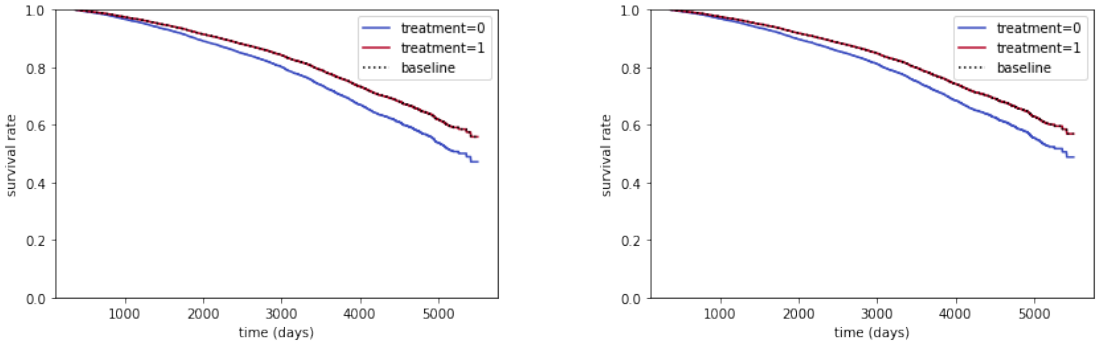
Considering the two Cox models are too similar, the Cox model we are referring to in further interpretation and discussion is the Cox model after the propensity score adjustment if not specifically mentioned.

Shown in figure 6(a) and 6(b) are the survival curves for the Cox models before and after the propensity score adjustment, respectively. The red and blue curve represent survival rates of active treatment group and non-active treatment group, respectively.

In figure 6(b), the red curve is above the blue curve, indicating that active treatment decreases the risk of death. The gap between the two survival rates becomes larger over time. In the first three years (1095 days), the active and non-active treated groups share high survival rate of about 95%. In the 10-th year (3650 days), the survival rates of active and non-active treated groups are about 80% and 75%, respectively. At the end of the x-axis (5000 days), the survival rate of the two groups are about 58% and 45%.

Table 5: Distribution of covariates stratified by propensity score quintiles. All data in the table are percentages.

	1st_quintile		2nd_quintile		3rd_quintile		4th_quintile		5th_quintile		total
	0.139-0.628		0.628-0.718		0.718-0.789		0.789-0.857		0.857-0.964		
	Observation	Treatment	Observation	Treatment	Observation	Treatment	Observation	Treatment	Observation	Treatment	
	n = 832	n = 915	n = 585	n = 1083	n = 463	n = 1274	n = 278	n = 1406	n = 125	n = 1519	
tumor_stage											
≤ T2a	0.969	0.950	0.942	0.946	0.911	0.948	0.701	0.725	0.136	0.144	0.747
T2b and T2c	0.031	0.050	0.058	0.054	0.089	0.052	0.299	0.275	0.864	0.856	0.253
tumor_grade											
well-differentiated	0.065	0.051	0.009	0.010	0.004	0.002	0.004	0.003	0.000	0.001	0.015
moderately-differentiated	0.935	0.949	0.991	0.990	0.996	0.998	0.996	0.997	1.000	0.999	0.985
Year of diagnosis											
2005	0.036	0.026	0.017	0.012	0.065	0.051	0.083	0.095	0.168	0.165	0.071
2006	0.066	0.038	0.068	0.076	0.032	0.065	0.112	0.125	0.112	0.095	0.080
2007	0.082	0.075	0.053	0.056	0.058	0.067	0.126	0.112	0.040	0.049	0.072
2008	0.035	0.027	0.034	0.052	0.052	0.054	0.112	0.090	0.096	0.101	0.064
2009	0.036	0.037	0.062	0.054	0.080	0.075	0.083	0.106	0.112	0.072	0.069
2010	0.087	0.090	0.118	0.116	0.104	0.107	0.101	0.090	0.088	0.090	0.098
2011	0.127	0.151	0.169	0.139	0.080	0.105	0.108	0.097	0.080	0.111	0.119
2012	0.107	0.118	0.109	0.114	0.201	0.194	0.137	0.116	0.120	0.130	0.134
2013	0.188	0.180	0.236	0.222	0.153	0.147	0.061	0.091	0.120	0.116	0.152
2014	0.237	0.257	0.133	0.158	0.175	0.135	0.079	0.078	0.064	0.071	0.139
age											
65-67	0.013	0.021	0.321	0.302	0.337	0.294	0.349	0.399	0.272	0.423	0.284
68-70	0.091	0.106	0.260	0.247	0.294	0.343	0.295	0.310	0.344	0.342	0.265
71-73	0.165	0.170	0.197	0.224	0.240	0.268	0.209	0.195	0.192	0.191	0.206
74-77	0.413	0.498	0.185	0.208	0.097	0.091	0.147	0.096	0.192	0.044	0.184
78-80	0.317	0.204	0.038	0.019	0.032	0.004	0.000	0.000	0.000	0.000	0.061
PSA											
[0,4)	0.069	0.052	0.014	0.022	0.004	0.007	0.022	0.011	0.000	0.001	0.020
[4,10)	0.477	0.482	0.699	0.678	0.266	0.290	0.187	0.223	0.192	0.238	0.380
[10,20)	0.333	0.381	0.176	0.183	0.594	0.541	0.543	0.499	0.456	0.494	0.419
≥20	0.121	0.084	0.111	0.117	0.136	0.162	0.248	0.266	0.352	0.267	0.181



(a) (b)

Figure 6: Survival curves obtained by Cox models *left*: before the adjustment of propensity score and *right*: after the adjustment of propensity score Blue curve (treatment = 0) denotes non-active treatment group and red curve (treatment = 1) denotes active treatment group

As shown in table 7, active treatment decreased 21% risk of death (HR = 0.79, 95% CI = [0.73,0.85], p <0.005) compared with observation. Furthermore, age, PSA and tumor stage were also statistically associated to survival (p-value <0.05). Patients who were older or had a worse tumor stage would have a higher risk of death (HR > 1). Risk of death was not associated with the incidence year, except for the year of 2006 and 2013 (p-value <0.05).

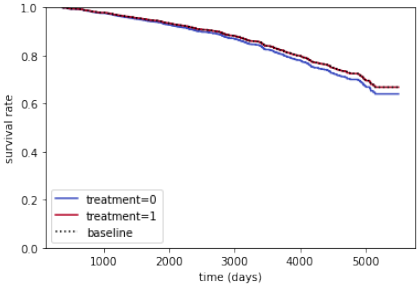
Figure 7 shows the survival curves stratified by treatment for each of the age groups ([65,67], [68,70], [71,73], [74,77], [78,80]). Overall speaking, figure 7 indicates a decreasing trend of survival rate with increasing age both for active treatment group and non-active treatment group. Take the active treatment group (red curve) as an example. Compare the five red curves, at any time t, the survival rate of younger patients were higher than older patients. This was the same when comparing the five blue curves (observation). When focusing on the differences in survival rates between the treatment and the observation within each age group, the HR of treatment for the five age groups were 0.91, 0.93, 0.82, 0.72 and 0.84, respectively, as shown in table 8. Therefore, active treatment decreased the mortality among all age groups. But the effect of treatment was only statistically significant in the age group of 71-73 and 74-77 years (with p-value <0.05).

5.4 Bayesian network

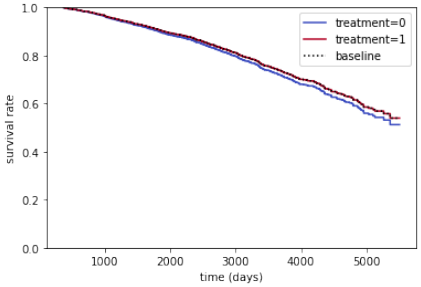
The network obtained by model averaging is shown in figure 8(a). It consists of 9 nodes and 10 edges. All of the edges presented at least 80% among the 2000 candidate networks. The frequencies of edges among the candidate networks were also known as *strength*. The width of the edges represents the strength of the relations: the edge becomes thicker if it occurs in more of the models generated from the data obtained from sampling with replacement from the original dataset. There are two undirected edges in figure 8(a): incjr – PSA and age – PSA. This was because the frequencies of the forward and reverse edges were about the same. We defined the direction of the two undirected edges as incidence year → PSA and age → PSA. It was odd to set an edge PSA → incjr, so the direction of the first undirected edge was set to incjr → PSA. For the second undirected edge, age → PSA was set as it was more likely to indicate a meaningful causal relation compared with the reversed direction.

The loglikelihood of the network obtained by model averaging was -54008.5. The average AUC were 0.59, 0.62 and 0.66 for survive_GE_2_years, survive_GE_5_years and treatment, respectively. The parents of treatment were age and tumor stage. The parents of survival were age, tumor stage and group.

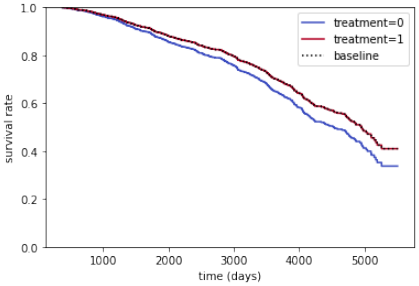
The structure of the final network, which has 19 edges, is shown in figure 9. The edges that we added were mentioned in the previous section and listed in table 3. The parents of node treatment were PSA, age and tumor stage, the parents of survival were treatment,tumor



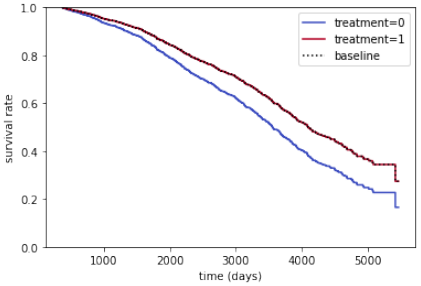
(a) 65-67 years old



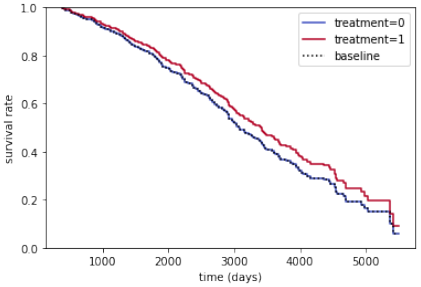
(b) 68-70 years old



(c) 71-73 years old



(d) 74-77 years old



(e) 78-80 years old

Figure 7: Survival curves (obtained by Cox model) stratified by different therapies across five age groups, where blue curve (treatment = 0) denotes non-active treatment group and red curve (treatment = 1) denotes active treatment group

Table 6: Association between variates in the dataset and mortality obtained by Cox regression before propensity score adjustment. All the covariates met the proportional hazards assumption except for age.

	Coef	HR for Death	95% CI lower bound	95% CI upper bound	z	p
treatment	-0.25	0.77	0.72	0.84	-6.22	<0.005
diffgr_map	0.07	1.07	0.84	1.37	0.58	0.56
tumor_stage	0.08	1.08	0.99	1.17	1.82	0.07
incjr_2005	-	1(Ref)	-	-	-	-
incjr_2006	0.11	1.12	1	1.26	1.93	0.05
incjr_2007	0.08	1.09	0.96	1.23	1.34	0.18
incjr_2008	0.05	1.05	0.92	1.2	0.7	0.49
incjr_2009	0.06	1.06	0.92	1.21	0.82	0.41
incjr_2010	-0.04	0.96	0.85	1.09	-0.62	0.53
incjr_2011	-0.05	0.95	0.84	1.07	-0.84	0.4
incjr_2012	-0.11	0.9	0.79	1.02	-1.66	0.1
incjr_2013	-0.14	0.87	0.77	0.99	-2.12	0.03
incjr_2014	-0.09	0.92	0.8	1.05	-1.25	0.21
age_65-67	-	1(Ref)	-	-	-	-
age_68-70	0.06	1.06	0.97	1.16	1.29	0.2
age_71-73	0.21	1.23	1.12	1.35	4.23	<0.005
age_74-77	0.55	1.74	1.59	1.91	11.7	<0.005
age_78-80	0.8	2.23	1.96	2.55	11.91	<0.005
PSA_0-4	-	1(Ref)	-	-	-	-
PSA_4-10	-0.14	0.87	0.79	0.96	-2.85	<0.005
PSA_10-20	0	1	0.91	1.1	0	1
PSA_>20	0.22	1.25	1.12	1.39	4.12	<0.005

Abbreviations: HR = hazard ratio; CI = confidence interval, Ref = reference level

stage, age and PSA. Loglikelihood of this network was -53787.8, a little higher than the previous one obtained by model averaging. The average AUC was 0.63, 0.62 and 0.70 for survive_GE_2_years, survive_GE_5_years and treatment. The average AUC increased slightly after adding the edges.

Figure 10 shows the ROC curve of five-fold cross-validation of the final network. The blue, red and green curve represent the ROC curve of survive_GE_2_years, survive_GE_5_years and treatment, respectively. To make a better comparison, table 9 shows the characteristics of the two networks (network obtained by model averaging and the final network) such as average AUC, the number of edges and the loglikelihood.

Calibration of the final network for survive_GE_2_years, survive_GE_5_years and treatment were visualized in figure 11(a), figure 11(b) and 11(c), respectively. The x-axis denotes the predicted probability and y-axis denotes the observed probability. The line with slope 1 indicates a perfect calibration (predicted = observed). In figure 11(c), each dot defines a unique combination of the parents of *treatment* (i.e., age, PSA and tumor stage). Similarly, in figure 11(a) and figure 11(b), each dot defines a unique combination of the parents of dots that indicate survival. The size of the dot denotes the size of the subpopulation. Although some of the small dots were far from the line, most of the dots were close to it. The calibration plot indicates the estimated probabilities are close to the observed frequencies. The BN model fit the dataset well.

The calibration plot of treatment (figure 11(c)) shows that patients between ages 78 and 80 were least likely to receive active treatment compared with younger patients: less than half of them received active treatment. Patients between 74 and 77 years were the second least likely group to receive active treatment. However, the effect of age one treatment allocation was weaker in younger groups, in which tumor stage seemed to be a stronger factor instead. Dots that represent age group of 65-67 (red), 68-70 (brown-green) and 71-73 (green) are mixed at

Table 7: Association between variates in the dataset and mortality obtained by Cox regression after propensity score adjustment. All the covariates met the proportional hazards assumption.

	Coef	HR for Death	95% CI lower bound	95% CI upper bound	z	p
treatment	-0.24	0.79	0.73	0.85	-5.83	<0.005
diffgr_map	0.14	1.16	0.9	1.48	1.15	0.25
tumor_stage	0.12	1.13	1.04	1.24	2.76	0.01
incjr_2005	-	1(Ref)	-	-	-	-
incjr_2006	0.12	1.13	1	1.26	2	0.05
incjr_2007	0.08	1.08	0.96	1.22	1.25	0.21
incjr_2008	0.06	1.06	0.92	1.21	0.8	0.42
incjr_2009	0.05	1.06	0.92	1.21	0.78	0.44
incjr_2010	-0.05	0.95	0.84	1.08	-0.77	0.44
incjr_2011	-0.07	0.94	0.83	1.06	-1.02	0.31
incjr_2012	-0.12	0.89	0.78	1.01	-1.79	0.07
incjr_2013	-0.15	0.86	0.75	0.97	-2.38	0.02
incjr_2014	-0.11	0.9	0.78	1.03	-1.54	0.12
quintiles_lv_1	0.2	1.22	1.09	1.37	3.45	<0.005
quintiles_lv_2	0.07	1.07	0.97	1.19	1.35	0.18
quintiles_lv_3	0.02	1.02	0.93	1.13	0.44	0.66
quintiles_lv_4	0	1	0.91	1.1	-0.01	0.99
quintiles_lv_5	-	1(Ref)	-	-	-	-
age_65-67	-	1(Ref)	-	-	-	-
age_68-70	0.07	1.07	0.98	1.17	1.44	0.15
age_71-73	0.19	1.21	1.09	1.33	3.79	<0.005
age_74-77	0.5	1.64	1.49	1.81	9.85	<0.005
age_78-80	0.71	2.04	1.77	2.35	9.88	<0.005
PSA_0-4	-	1(Ref)	-	-	-	-
PSA_4-10	-0.15	0.86	0.78	0.94	-3.12	<0.005
PSA_10-20	0.01	1.01	0.92	1.11	0.26	0.8
PSA_>20	0.24	1.27	1.15	1.42	4.48	<0.005

Abbreviations: HR = hazard ratio; CI = confidence interval, Ref = reference level

Table 8: HR estimates of the association between treatment and survival in different age groups

	coef	exp(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p
65-67	-0.1	0.91	-0.31	0.11	0.73	1.12	-0.93	0.35
68-70	-0.07	0.93	-0.26	0.11	0.77	1.12	-0.79	0.43
71-73	-0.2	0.82	-0.38	-0.01	0.69	0.99	-2.12	0.03
74-77	-0.32	0.72	-0.47	-0.18	0.62	0.84	-4.32	<0.005
78-80	-0.17	0.84	-0.4	0.06	0.67	1.06	-1.49	0.14

Table 9: Characteristics of the BN model obtained by model averaging and the final BN model

Model	AUC for 2-year survival rate	AUC for 5-year survival rate	AUC for treatment	Edges	Loglikelihood
BN model obtained by model averaging	0.59	0.62	0.66	10	-54008.5
Final network	0.62	0.62	0.7	19	-53787.8

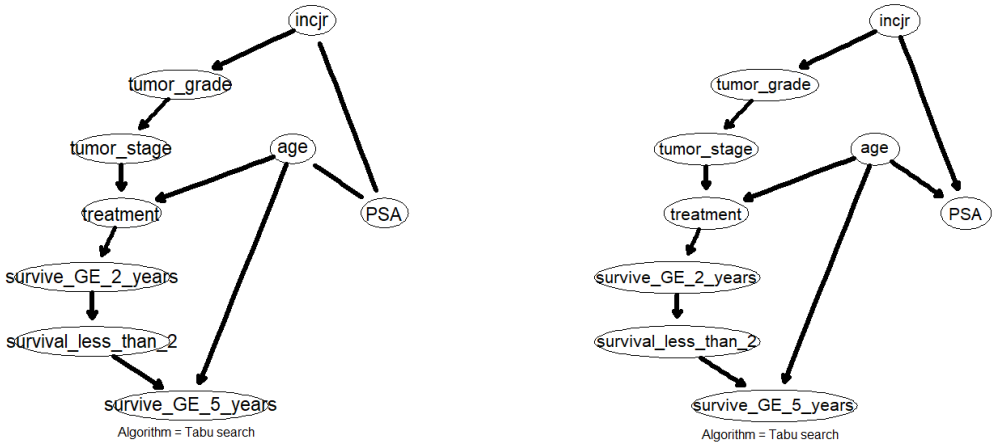


Figure 8: *left*: Network generated by the methods of model averaging, in which the candidate networks were generated by Tabu search. Note that there are two undirected edges and *right*: After define the directions of the undirected edges manually.

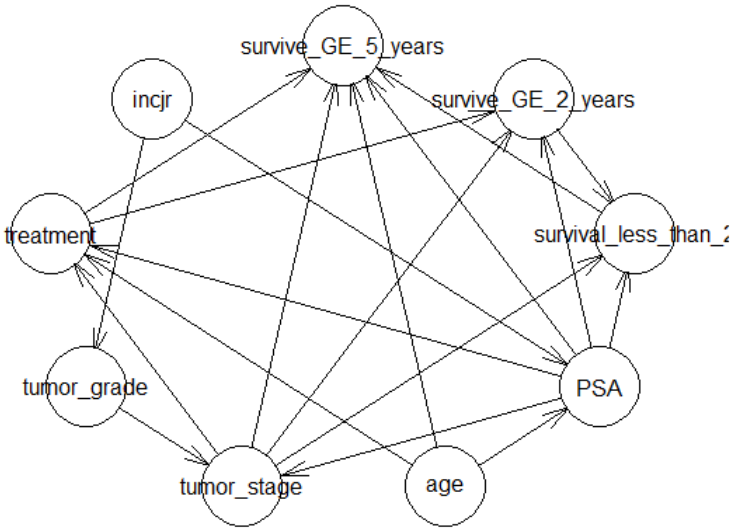


Figure 9: The structure of the final BN model

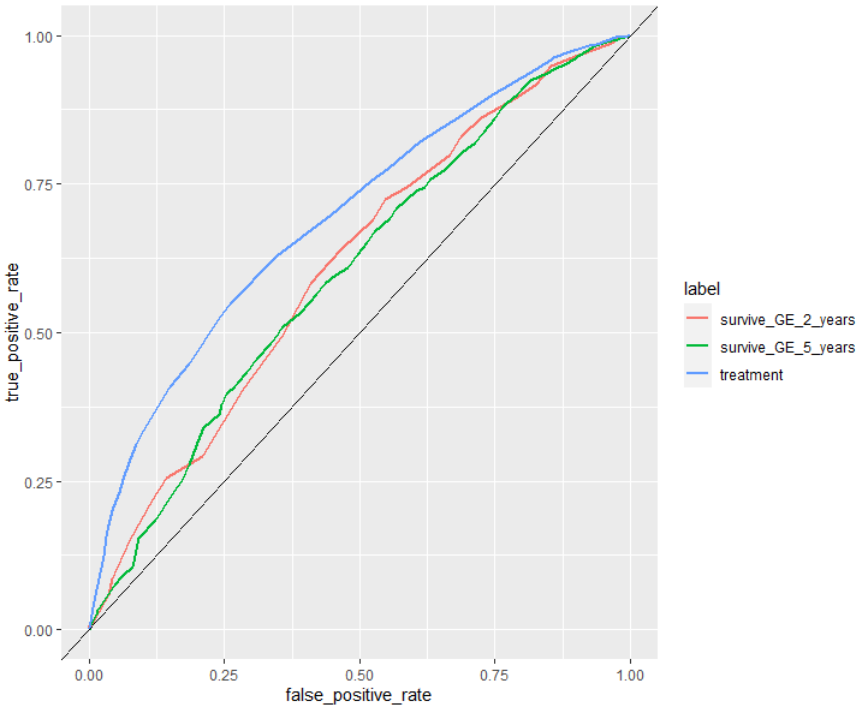


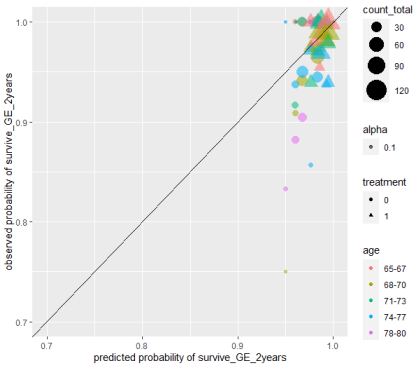
Figure 10: Average ROC curve for treatment allocation, 2-year survival rate and 5-year survival rate of the final network after 5-fold cross validation. The red, blue and green curve represent ROC curve for 2-year survival rate, 5-year survival rate and treatment, respectively.

the top and the middle of the line. But subpopulations represented by triangular dots (worse tumor stage) are concentrated at the top of the line, which means they have a higher observed probability of getting active treatment compared with circular dots (better tumor stage) in the middle of the diagonal.

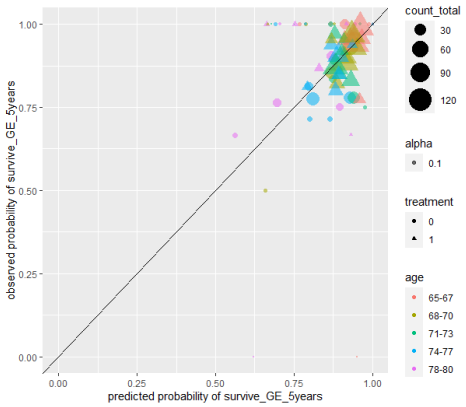
In figure 11(a), most of the dots are on the top-right. It indicates that most of the patients were alive after two years regardless of their characteristics. Even the elderly (age group of 78-80) had an observed 2-year survival rate larger than 80%.

But things are slightly different when it comes to the 5-year survival rate (figure 11(b)). In this figure, active treatment groups (triangular dots) and younger groups (65-67 that presented in red, 68-70 that presented in brown-green) have a better 5-year survival rate ($\geq$ 85%) at the very top of the line. Furthermore, the survival rate of younger age groups were better predicted. Most dots that less close to the line were purple or blue.

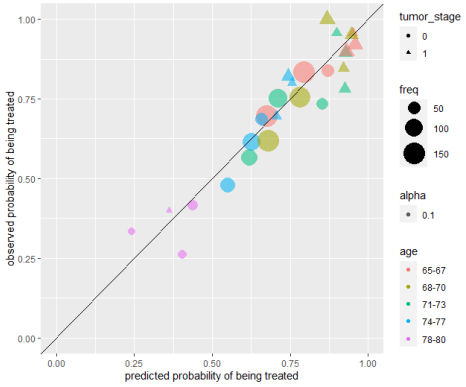
The changes of survival rate when setting difference treatment are listed in table 10. Doing non-causal reasoning using the final network by setting the evidence *treatment = 1*, the 2-year and 5-year survival rate was 98.7% and 91.3%, respectively. Changing the evidence to *treatment = 0*, the 2-year and 5-year survival rate would be 97.0% and 87.7%. Doing causal reasoning by removing all the edges that pointed towards treatment (tumor stage $\rightarrow$ treatment, age $\rightarrow$ treatment and PSA $\rightarrow$ treatment) and setting an evidence *treatment = 1*, the predicted 2-year and 5-year survival rate would be 98.8% and 91.5%, respectively. Alternatively, setting the evidence *treatment = 0*, the 2-year and 5-year survival rate would change to 97.2% and 87.6%. Doing causal and non-causal reasoning, the effect of treatment on survival presented by the BN model was quite similar, and the effect was much weaker than presented in Cox model (HR = 0.79).



(a) Calibration plot for 2-year survival rate



(b) Calibration plot for 5-year survival rate



(c) Calibration plot of treatment

Figure 11: Calibration plot of survival and treatment of the final BN model (figure 9). To better investigate the relations between predicted and observed probabilities in detail, the axis of figure (a) starts from 0.7 rather than 0, which is different from figure (b) and (c)

Table 10: Results of causal and non-causal reasoning when setting difference evidence of treatment in the final BN model

Survival time	non-causal reasoning		causal-reasoning	
	treatment = 1	*treatment = 0*	*treatment = 1*	*treatment = 0*
survive_GE_2_years	98.7	97	98.8	97.2
survive_GE_5_years	91.3	87.7	91.5	87.6

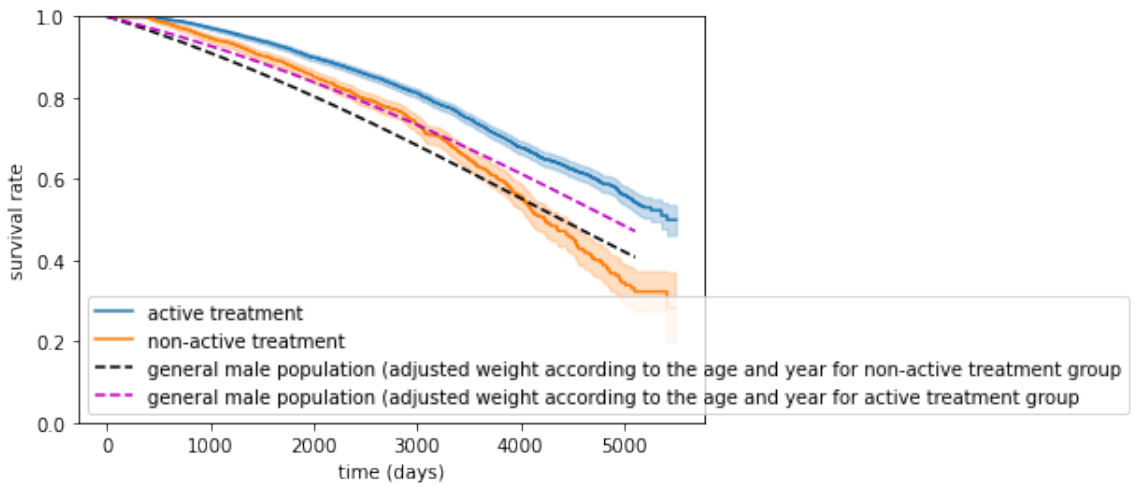


Figure 12: Survival curves for man with localized prostate cancer and for the general male population in the Netherlands. The survival rates of active and non-active treatment group are visualized by Kaplan–Meier estimator.

5.5 Visualizing the survival rate general male population in the Netherlands and the survival rate of active and non-active treatment group

The survival curves of active treatment group and non-active treatment groups were visualized by Kaplan–Meier estimator in figure 12. Figure 12 also includes the survival rates of the general male population that was treatment-corrected for age and year of diagnosis. The correction results in two dotted curves in figure 12: the magenta curve that represents the survival rate of general male population adjusted for age and year of diagnosis for active treatment group and the black curve that represents the survival rate of general male population adjusted for age and year of diagnosis for non-active treatment group. Note that all curves in the figure are kept constant in the first 365 days, during which the survival rates were 1. For the cancer-specific survival rate, this was due to the exclusion of patients that died within the first year of diagnosis. For survival rate of general male population, this was because we were using $P(1\_to\_n\_year\_survival)$.

At any time t, patients in active treatment group had a lower risk of death compared with patients in non-active treatment group and the general male population adjusted for age and year of diagnosis for active treatment group. Before 4000 days (about 11 years), the survival rate of general male population adjusted for non-active treatment group is slightly lower than survival rate of non-active treatment group. But the latter decline sharply and is lower than the former after 4000 days.

Adjusting the survival rate of general male population for age and year of diagnosis for active treatment or non-active treatment group would not cause a big difference as the magenta and black curves are close to each other.

To make it more detailed, table 11 shows the 2-year (730 days), 5-year (1825 days) and 10-year (3650 days) survival rates of **(1.)**general male population that was treatment-corrected by age and year of diagnosis, **(2.)** active and non-active treatment group visualized by Kaplan–Meier estimator and **(3.)** active and non-active treatment group estimated by Cox model and the BN model. Please note that the BN model didn't predict the 10-year survival rate as there were no enough cases.

Table 11: The 2-year, 5-year and 10-year survival rates obtained by different approaches. Numbers in the table are percentages. Please note that the BN model didn't predict the 10-year survival rate as there were no enough cases to make a prediction

Treatment	Model	2-year survival rate	5-year survival rate	10-year survival rate
Active treatment group	data from CBS (adjusted weight based on age and years distribution of active treatment group)	97.5	88.8	70
	Kaplan-Meier	98.7	91.4	72.6
	Cox model	98.5	92	74
	BN model(causal reasoning)	98.8	91.5	-
	BN model(non-causal reasoning)	98.7	91.3	-
Non-active treatment group	Data from CBS (adjusted weight based on age and years distribution of non-active treatment group)	96.9	86.3	64.6
	Kaplan-Meier	97.0	87.7	62.4
	Cox model	97.5	88.0	65.0
	BN model(causal reasoning)	97.2	87.6	-
	BN model(non-causal reasoning)	97.0	87.7	-

6 Discussion and conclusion

6.1 Summary

During the experiment, we reanalysed an NCR dataset: active treatment versus observation for men who were diagnosed with localized prostate cancer. We suspected that the outcome of patients would be influenced by treatment selection bias. A longer survive time may be explained by active treatment and better baseline health condition. In our dataset, the only variable that could partly indicate baseline health was age. So do not rule out other unmeasured confounders like self-rated health or measure physical function, though they were not provided in the dataset we had.

Our analysis showed that active treatment (radical prostatectomy, radiation implants or external-beam radiation therapy) decreased the risk of death for localized prostate cancer patients. In the Cox model after propensity score adjustment, active treatment decreased 20% risk of death compared with observation (HR = 0.79, 95% CI = [0.73, 0.85], p <0.05) for patients with localized prostate cancer. This conclusion, which was similar to the results obtained by Wong [69] and Giordano [32], was also comparable to the only published randomized trial investigating the outcome of radical prostatectomy versus observation for localized prostate cancer (relative risk: 0.74, 95% CI = [0.56, 0.99]) [19]. Furthermore, a series of Cox models for different age groups proofed the benefit of active treatment in all the age groups (table 8). This treatment survival advantage was more pronounced in age groups of 71-73 and 74-77 years (p <0.05). In the final BN model shown in figure 9, although both causal and non-causal inference indicate a survival benefit for active treatment as shown in table 10, this positive association was much weaker than the results of Cox model (HR = 0.79). Finally, the calibration plot for 5-year survival rate (figure 11(b)) also indicates actively treated patients had a higher probability of being alive 5 years after diagnosis.

Our study also found age was a factor that could partly explained the difference in outcome between active treatment group and non-active treatment group. In the Cox model, older patients, especially patients in the two oldest age groups, had a higher risk of death compared with the youngest age group (p <0.005, HR = 1.64 for ages 74-77 years, HR = 2.04 for ages 78-80 years as shown in table 7). In the BN model, the calibration plot of 2-year survival rate (figure 11(a)) shows that except for patients ages 78-80 years, all other patients had a 2-year survival rate larger than 90%. The calibration plot of 5-year survival rate (figure 11(b)) also indicates that younger patients had a better 5-year survival rate. The two youngest age groups (65-67 years, 68-70 years) had a similar 5-year survival rate (≤ 85%). Patients in the oldest age group had the worst 5-year survival rate. Though age in itself is not a risk factor, older age may be linked to a higher risk of other chronic diseases like hypertension, asthma and congestive heart failure [30], thus poorer health and a worse prognosis. In the paper of Seth *et al.*, they found that the age at diagnosis would have a stronger influence on treatment compared with cancer risk for prostate cancer. If the clinical characteristics of the patient

have been carefully selected to be similar, the life expectancy of the older man who received active treatment will increase to a level that is comparable with younger patients [15].

In figure 12, the survival rate of the general male population in the Netherlands was treatment-corrected for age and year of diagnosis. We compared **(1.)**the survival rate of active treatment group visualized by Kaplan–Meier estimator and the survival rate of general male population corrected for active treatment group for age and year of diagnosis and **(2.)** the survival rate of non-active treatment group visualized by Kaplan–Meier estimator and the survival rate of general male population corrected for non-active treatment group for age and year of diagnosis. Both of the two cases produced incredible results: general male population in the Netherlands had a lower survival rate than localized prostate cancer patients even after the treatment-correlation for age and year of diagnosis. An explanation is that unmeasured confounders are responsible for this effect that we observed. The survival rate of general male population corrected for non-active treatment for age and year of diagnosis (black curve) was lower than the survival rate of non-active treatment group (orange curve) before 4000 days. But after 4000 days, the latter decreased rapidly and became lower than the former. However, the survival rate of general male population corrected for active treatment group for age and year of diagnosis (magenta curve) was always lower than survival rate of active treatment group (blue curve). They never crossed. Therefore some unmeasured confounders that would affect the 10-year survival were likely to persist in the dataset.

In Cox model, other important factors related to mortality were tumor stage (HR = 1.13, 95% CI = [0.9,1.48]). Men who had a better tumor stage had a lower mortality. In the BN model, we believed there were causal relations from variables like tumor stage, PSA value towards survival rate and treatment allocation. Therefore, we added some edges manually to the BN model which was obtained by model averaging (shown in figure 8(b)) and came up with the final BN model (shown in figure 9). After adding the edges, the increasing of loglikelihood indicated the final network fit the dataset better. The sensitivity analysis using AUC-ROC curve indicated our final BN model performed poor in predicting the survival with an average AUC of 0.62. It was better in predicting treatment with an average AUC of 0.7, but still not competitive for classification. However, this model was not simply used as a classifier. The calibration plot shown in figure 11 indicates that the most of the probabilities predicted by the final model are close to observed frequencies. So our BN model is meaningful in practical application for predicting and decision making when investigating the association between survival and a series of covariates.

In conclusion, all the three models (Kaplan-Meier estimator, Cox model and BN model) indicated that active treatment had a positive impact on survival. They also indicated that age could be a factor that partly explained the difference in outcome between active treatment group and non-active treatment group. Younger age was linked to a lower risk of death. Finally, the comparison between the weighted survival rate of general male population and the survival rate of cancer patients (active treatment group and non-active treatment group) implied the presence of unmeasured confounders which may affect the outcome of patients. If researchers simply attribute the differences to treatment when accessing the effectiveness of therapies using observational data, it will lead to misguided or even wrong conclusion. Our experiment was not the only one to report this issue [32] [15] [35]. Therefore when assessing the effectiveness of different therapies using observational data, researchers need to be cautious, and the results should be viewed critically.

6.2 Limitations and future work

Limits of propensity score adjustment In our experiment, the propensity score adjustment did not substantially modify the findings of Cox model. Although the adjustment removed the imbalance in all the variables except for age, it had little effect on the HRs. the Cox model before and after the adjustment were almost the same.

The reason could be that in a strongly ignorable assumption

$$Y \perp\!\!\!\perp_p T \mid X$$

X should be a set of measured covariates [57]. However, when we made a comparison between the survival rate of active and non-active treatment group obtained by Kaplan–Meier estimator and the adjusted survival rate of general male population, we have suspected unmeasured confounder persisted in the prostate cancer dataset.

If the strongly ignorable assumption does not hold, there is no way to say the treatment assignment is strongly ignorable after adjusted for propensity score, nor can we estimate the unbiased ATE. However, it is very difficult to prove this assumption in an observational dataset. We can keep finding new covariates that have distinct differences between treatment and control group, but we can hardly know if there are still some unmeasured confounders. Therefore, when propensity score adjustment would be used to balance the dataset, researchers need to be cautious and aware that the adjustment may not always effective as some unmeasured confounders could persist in the observational dataset.

One way to make the propensity score adjustment work and also make better use of observational data is to collect more variables, even when they seem to be inrelevant to the topic we are investigating. For instance, Mustafa *et al.* found some factors that are common to the localized prostate cancer treatment decision-making literature [12] including treatment type, some socioeconomic factors like personal income and education level and personal reasons like urinary function and ability to work. Therefore, before an observational study, it is good to consult experts in various fields and discuss which variable may be needed to collect.

No enough variables Variables in our dataset may be too few. The dataset that Wong [69] used had 13 variables. But our dataset from NCR consists of only 7 variables. In their dataset, some variables that indicate the patient's baseline health or sociological characteristics that seemed less relevant to the tumor, such as information on self-reported health, region, race and income were included. As mentioned above, these "irrelevant" variables may be needed to check how they would cause the potential bias of treatment selection without patients' or doctors' attention. The insufficient number of feature also led to a simple network, which is too hard to find some more complex patterns or information.

In some other researches in Canada and America, some of the sociodemographic variables are associated with the frequencies of PSA testing. In a research project in Canada, Gorday *et al.* found that higher education level and higher household income were associated with higher frequencies of PSA testing. What's more, men in the age group of 60-69 years had a higher testing rate than men aged above 70 years old [34]. A similar conclusion was obtained by Mitchell *et al.* in a study of prostate cancer screening in urban African-American men. In this study, increased education level and higher income were positively correlated to prostate cancer screening [52]. Some references had already found the association between the frequencies of receiving a PSA test and education level or personal income. A higher education level and household income may lead to increased access to health care compared to other population. Especially for localized prostate cancer, which is of low- or intermediate risk with an average 5-year survival rate of 90%, the better quality of health care may play a more important role for survival. In summary, the unmeasured differences in education level or personal income might explain why the survival rate of general male population was lower than cancer patients after adjusted for age and year of diagnosis. However, considering the difference in medical system between America and Europe, this statement remained unclear. Some more studies or experiment may be needed to prove that.

RCT is needed for reference Although it is difficult to conduct an RCT with long-term follow-up, and RCT like this may be needed. In our experiment, using different models would

lead to different conclusions. Analysing the same dataset with other methods would also lead to other conclusions. What's more, the weak association between active treatment and survival obtained by the BN model may only true for the low- to intermediate-risk prostate cancer we were investigating, rather than advanced or metastatic prostate cancer, which has a higher risk of death. Therefore, when analysing observational data, researchers need to view these conclusions critically, and an RCT would help to prove less unbiased results, which could be more reliable for reference.

References

[1] https://www.cbs.nl/.

[2] https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/tests.html.

[3] https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/tests-and-procedures/digital-rectal-exam-dre.

[4] https://www.cancer.gov/about-cancer/diagnosis-staging/staging.

[5] https://en.wikipedia.org/wiki/Censoring_(statistics).

[6] https://en.wikipedia.org/wiki/Bayesian_network.

[7] https://www.iknl.nl/en/ncr/.

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

[9] Jill L Adelson, DB McCoach, HJ Rogers, Jonathan A Adelson, and Timothy M Sauer. Developing and applying the propensity score to make causal inferences: variable selection and stratification. *Frontiers in psychology*, 8:1413, 2017.

[10] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[11] Peter C Albertsen, James A Hanley, David F Penson, George Barrows, and Judith Fine. 13-year outcomes following treatment for clinically localized prostate cancer in a population based cohort. *The Journal of urology*, 177(3):932–936, 2007.

[12] Mustafa Andkhoie, Desneige Meyer, and Michael Szafron. Factors underlying treatment decision-making for localized prostate cancer in the us and canada: A scoping review using principal component analysis. *Canadian Urological Association Journal*, 13(7):E220, 2019.

[13] World Medical Association et al. World medical association declaration of helsinki. ethical principles for medical research involving human subjects. *Bulletin of the World Health Organization*, 79(4):373, 2001.

[14] Gunnar Aus, Svante Bergdahl, Pär Lodding, Hans Lilja, and Jonas Hugosson. Prostate cancer screening decreases the absolute risk of being diagnosed with advanced prostate cancer—results from a prospective, population-based randomized controlled trial. *European urology*, 51(3):659–664, 2007.

[15] Seth K Bechis, Peter R Carroll, and Matthew R Cooperberg. Impact of age at diagnosis on prostate cancer treatment and survival. *Journal of Clinical Oncology*, 29(2):235, 2011.

[16] Henry K Beecher. The powerful placebo. *Journal of the American Medical Association*, 159(17):1602–1606, 1955.

[17] Irad Ben-Gal. Bayesian networks. *Encyclopedia of statistics in quality and reliability*, 1, 2008.

[18] Stefano Beretta, Mauro Castelli, Ivo Gonçalves, Roberto Henriques, and Daniele Ramazzotti. Learning the structure of bayesian networks: A quantitative assessment of the effect of different algorithmic schemes. *Complexity*, 2018, 2018.

[19] Anna Bill-Axelson, Lars Holmberg, Mirja Ruutu, Michael Häggman, Swen-Olof Andersson, Stefan Bratell, Anders Spångberg, Christer Busch, Stig Nordling, Hans Garmo, et al. Radical prostatectomy versus watchful waiting in early prostate cancer. *N Engl J Med*, 352:1977–1984, 2005.

[20] Otis W Brawley. Prostate cancer epidemiology in the united states. *World journal of urology*, 30(2):195–200, 2012.

[21] Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

[22] William G Cochran and Donald B Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446, 1973.

[23] Prostate Conditions Education Council. Gleason score, 2018.

[24] Adnan Darwiche, Keith Casico, David Allen, Hei Chan, Mark Chavira, James Park, Denis Zaloznyy, and Mike Zaloznyy. Samiam: Sensitivity analysis, modeling, inference, and more. *Software available from http://reasoning. cs. ucla. edu/samiam*, 2017.

[25] Cameron Davidson-Pilon, Jonas Kalderstam, Noah Jacobson, sean reed, Ben Kuhn, Paul Zivich, Mike Williamson, AbdealiJK, Deepyaman Datta, Andrew Fiore-Gartland, Alex Parij, Daniel WIlson, Gabriel, Luis Moneda, Kyle Stark, Arturo Moncada-Torres, Harsh Gadgil, Jona, Karthikeyan Singaravelan, Lilian Besson, Miguel Sancho Peña, Steven Anton, Andreas Klintberg, Javad Noorbakhsh, Matthew Begun, Ravin Kumar, Sean Hussey, Dave Golland, jlim13, and Abraham Flaxman. Camdavidsonpilon/lifelines: v0.25.1, August 2020.

[26] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[27] Chuong B Do and Serafim Batzoglou. What is the expectation maximization algorithm? *Nature biotechnology*, 26(8):897–899, 2008.

[28] Valerie J Easton and John H McColl. Statistics glossary v1. 1. 1997.

[29] Stephen B Edge, David R Byrd, Michael A Carducci, Carolyn C Compton, AG Fritz, FL Greene, et al. *AJCC cancer staging manual*, volume 7. Springer New York, 2010.

[30] Julie A Gazmararian, Mark V Williams, Jennifer Peel, and David W Baker. Health literacy and knowledge of chronic disease. *Patient education and counseling*, 51(3):267–275, 2003.

[31] Alan S Gerber, Donald P Green, Edward H Kaplan, Ian Shapiro, Rogers M Smith, and Tarek Massoud. The illusion of learning from observational research. *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*, pages 9–32, 2014.

[32] Sharon H Giordano, Yong-Fang Kuo, Zhigang Duan, Gabriel N Hortobagyi, Jean Freeman, and James S Goodwin. Limits of observational data in determining outcomes from cancer therapy. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 112(11):2456–2466, 2008.

[33] Fred Glover and Manuel Laguna. Tabu search. In *Handbook of combinatorial optimization*, pages 2093–2229. Springer, 1998.

[34] William Gorday, Hossein Sadrzadeh, Lawrence de Koning, and Christopher Naugler. Association of sociodemographic factors and prostate-specific antigen (psa) testing. *Clinical biochemistry*, 47(16-17):164–169, 2014.

[35] Kristin A Greco, Joshua J Meeks, Simon Wu, and Robert B Nadler. Robot-assisted radical prostatectomy in men aged ≥ 70 years. *BJU international*, 104(10):1492–1495, 2009.

[36] Henrik Grönberg. Prostate cancer epidemiology. *The Lancet*, 361(9360):859–864, 2003.

[37] Daniel B Hall, Robert F Woolson, William R Clarke, and Martha F Jones. 16 cochran-mantel-haenszel techniques: Applications involving epidemiologic survey data. *Bioenvironmental and public health statistics*, 18:483–500, 2000.

[38] William Hamilton and Deborah Sharp. Symptomatic diagnosis of prostate cancer in primary care: a structured review. *Br J Gen Pract*, 54(505):617–621, 2004.

[39] Eduardo Hariton and Joseph J Locascio. Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13):1716, 2018.

[40] Dawn L Hershman and Jason D Wright. Comparative effectiveness research in oncology methodology: observational data. *Journal of clinical oncology*, 30(34):4215–4222, 2012.

[41] Julian PT Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 2019.

[42] Jian Huang et al. Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*, 24(2):540–568, 1996.

[43] Jonas Hugosson, Monique J Roobol, Marianne Månsson, Teuvo LJ Tammela, Marco Zappa, Vera Nelen, Maciej Kwiatkowski, Marcos Lujan, Sigrid V Carlsson, Kirsi M Talala, et al. A 16-yr follow-up of the european randomized study of screening for prostate cancer. *European urology*, 76(1):43–51, 2019.

[44] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[45] J Richard Landis, Eugene R Heyman, and Gary G Koch. Average partial association in three-way contingency tables: a review and discussion of alternative tests. *International Statistical Review/Revue Internationale de Statistique*, pages 237–254, 1978.

[46] Antoni Ligeza. Artificial intelligence: A modern approach: By stuart russell and peter norwig. isbn# 0-13-103805-2, pp. 932+ 28. prentice hall, englewood cliffs, new jersey, 1995, 1995.

[47] Hans Lilja, David Ulmert, and Andrew J Vickers. Prostate-specific antigen and prostate cancer: prediction, detection and monitoring. *Nature Reviews Cancer*, 8(4):268–278, 2008.

[48] Nathan Mantel. Chi-square tests with one degree of freedom; extensions of the mantel-haenszel procedure. *Journal of the American Statistical Association*, 58(303):690–700, 1963.

[49] Brian R Matlaga, L Andrew Eskew, and DAVID L McCULLOUGH. Prostate biopsy: indications and technique. *The Journal of urology*, 169(1):12–19, 2003.

[50] John E McNeal, Arnauld A Villers, Elise A Redwine, Fuad S Freiha, and Thomas A Stamey. Histologic differentiation, cancer volume, and pelvic lymph node metastasis in adenocarcinoma of the prostate. *Cancer*, 66(6):1225–1233, 1990.

[51] Rupert G Miller Jr. *Survival analysis*, volume 66. John Wiley & Sons, 2011.

[52] Jamie Mitchell. Examining the influence of social ecological factors on prostate cancer screening in urban african-american men. *Social Work in Health Care*, 50(8):639–655, 2011.

[53] Radhakrishnan Nagarajan, Marco Scutari, and Sophie Lèbre. Bayesian networks in r. *Springer*, 122:125–127, 2013.

[54] Cecilia Nardini. The ethics of clinical trials. *Ecancermedicalscience*, 8, 2014.

[55] Sarang Narkhede. Understanding auc-roc curve. *Towards Data Science*, 26, 2018.

[56] John R Packer and Norman J Maitland. The molecular and cellular origin of human prostate cancer. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1863(6):1238–1260, 2016.

[57] Judea Pearl. *Causality*. Cambridge university press, 2009.

[58] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[59] Pascal Probst, Kathrin Grummich, Patrick Heger, Steffen Zaschke, Phillip Knebel, Alexis Ulrich, Markus W Büchler, and Markus K Diener. Blinding in randomized controlled trials in general and abdominal surgery: protocol for a systematic review and empirical study. *Systematic reviews*, 5(1):48, 2016.

[60] David B Resnik. Randomized controlled trials in environmental health research: ethical issues. *Journal of environmental health*, 70(6):28, 2008.

[61] Alan H Roberts, Donald G Kewman, Lisa Mercier, and Mel Hovell. The power of nonspecific effects in healing: Implications for psychosocial and biological treatments. *Clinical Psychology Review*, 13(5):375–391, 1993.

[62] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[63] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[64] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[65] Marco Scutari. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.

[66] Vianda S Stel, Friedo W Dekker, Giovanni Tripepi, Carmine Zoccali, and Kitty J Jager. Survival analysis ii: Cox regression. *Nephron Clinical Practice*, 119(3):c255–c260, 2011.

[67] Chengwei Su, Angeline Andrew, Margaret R Karagas, and Mark E Borsuk. Using bayesian networks to discover relations between genes, environment, and disease. *BioData mining*, 6(1):6, 2013.

[68] Judith A Turner, Richard A Deyo, John D Loeser, Michael Von Korff, and Wilbert E Fordyce. The importance of placebo effects in pain treatment and research. *Jama*, 271(20):1609–1614, 1994.

[69] Yu-Ning Wong, Nandita Mitra, Gary Hudes, Russell Localio, J Sanford Schwartz, Fei Wan, Chantal Montagnet, and Katrina Armstrong. Survival associated with treatment vs observation of localized prostate cancer in elderly men. *Jama*, 296(22):2683–2693, 2006.

A Strength of all the possible arcs obtained by model averaging

No.	From	To	Strength	Direction
1	incjr	diffgr_map	1	1
2	tumor_stage	group	1	1
3	age	group	1	1
4	survival_less_than_2	survive_GE_5_years	1	1
5	survive_GE_2_years	survival_less_than_2	1	1
6	PSA	age	1	0.5
7	age	PSA	1	0.5
8	PSA	incjr	1	0.5
9	incjr	PSA	1	0.5
10	age	survive_GE_5_years	0.963	1
11	group	survive_GE_2_years	0.896	1
12	diffgr_map	tumor_stage	0.831	1
13	PSA	group	0.113	1
14	PSA	survive_GE_2_years	0.0405	1
15	age	survive_GE_2_years	0.04	1
16	PSA	survive_GE_5_years	0.018	1
17	tumor_stage	PSA	0.011	0.5
18	PSA	tumor_stage	0.011	0.5
19	group	survive_GE_5_years	0.0075	1
20	diffgr_map	survive_GE_2_years	0.0015	1
21	tumor_stage	survive_GE_2_years	0.001	1
22	tumor_stage	survive_GE_5_years	0.0005	1

If the value of direction is 1, the edge is an directed edge. If the value of direction is 0.5, this edge is an undirected edge.