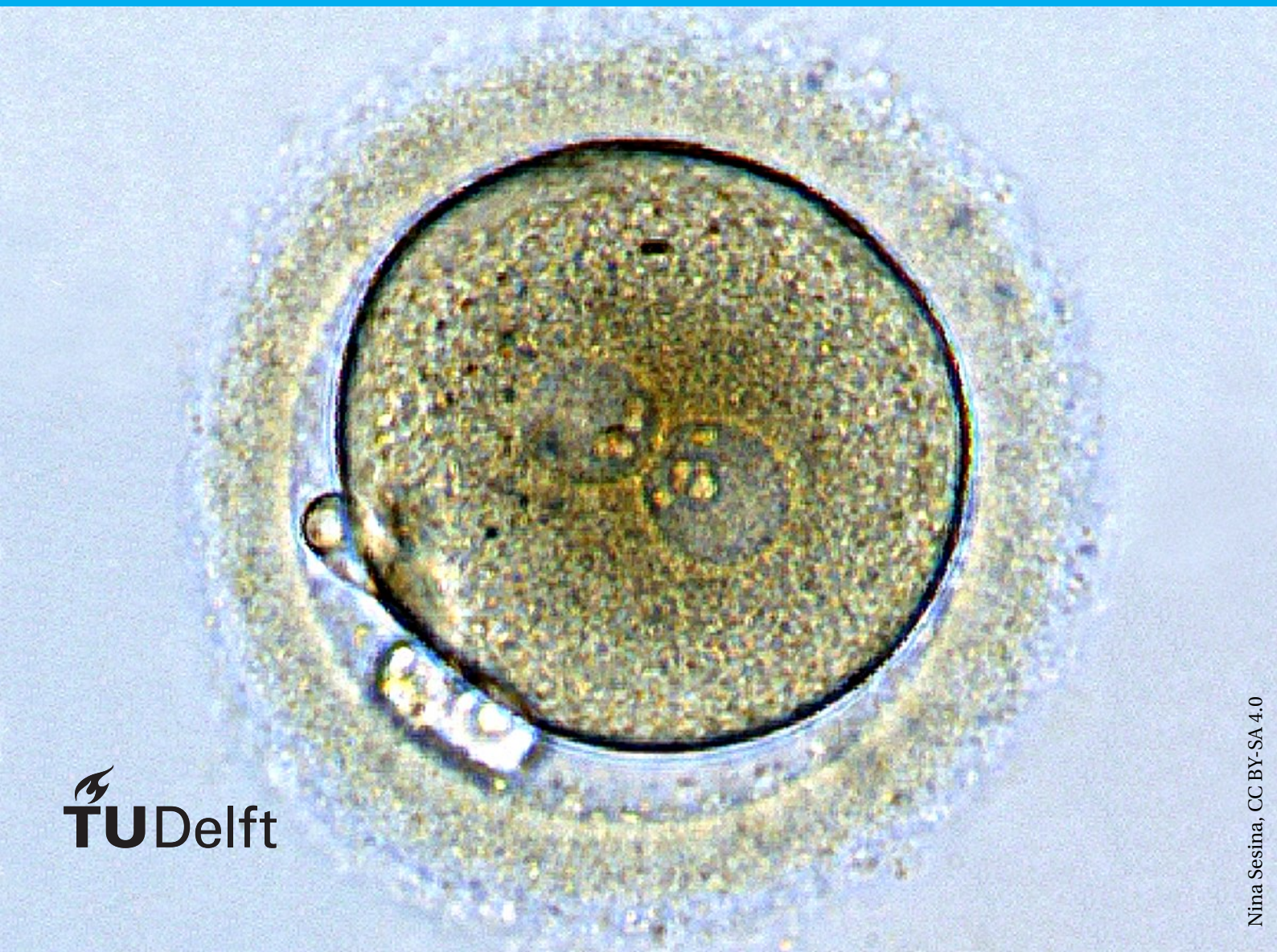
Causal inference with multi-state models, applied on estimating the effect of the IUI treatment timing for couples with unexplained subfertility

Master Thesis

Matea Skypala



TU Delft

Causal inference with multi-state models, applied on estimating the effect of the IUI treatment timing for couples with unexplained subfertility

Master Thesis

by

Matea Skypala

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday March 31, 2021 at 14:00 PM.

Student number:	4915143	
Project duration:	May 1, 2012 – March 31, 2021	
Thesis committee:	Dr. ir. G. F. Nane,	TU Delft, supervisor
	Dr. ir. N. van Geloven,	Leiden University Medical Center, external supervisor
	Prof. dr. S. le Cessie	Leiden University Medical Center, external supervisor
	Prof. dr. ir. G. Jongbloed,	TU Delft

Abstract

This thesis examines statistical methods to find the right timing of intrauterine insemination treatment relative to the start of the follow-up of the couples. Intrauterine insemination is a fertility treatment conducted by injecting refined sperm into a woman's uterus. Lots of research has been done on timing of steps within one IUI cycle. However, not a lot of research has been done on investigating if couples should be advised to start with the IUI immediately after consulting a fertility clinic, or wait a few more months to see if the pregnancy occurs naturally during this time. To analyze this problem, treatment strategies are defined in the following way: a couple stays on expectant management until some predetermined time when the first IUI cycle is started, unless they become pregnant before that time. Strategies analyzed involve starting the treatment at 0, 3, 6 or 9 months, or not at all until the end of the analysis, which is 1.5 years after diagnosis.

Pregnancy probability for each couple is estimated with a multi-state Cox proportional hazards model. Then, the model is connected to the causal inference setting in order to compute the counterfactual pregnancy probability of each couple in the population . This thesis explores how multi-state models can be used to answer causal questions. To do this, a 3-state multi-state model is carefully connected to the causal inference theory and the assumptions this framework relies on are listed and commented upon. Treatment strategies are implemented through making an intervention on the IUI starting time. Then, a causal inference method, G-computation, is used to estimate the expected pregnancy probability of all couples in the population, given that everyone follows the same treatment strategy. Then, individual pregnancy probabilities are averaged in order to obtain the expected pregnancy rates in the population for each strategy. Methodology combining multi-state models and causal inference is new and only one similar study has been found so far in the literature. Multi-state models are currently used mostly for predictive analyses and it is of great interest to use them to answer causal questions as well.

The results of this study show that there is no significant effect of the time the IUI has been started on the pregnancy rate in the population after 1.5 years, but there is significant difference between being and not being treated.

Contents

List of Figures

List of Tables

Listings

1 Introduction

This thesis examines the question of what is the right timing of the intrauterine insemination (IUI) treatment for couples with unexplained subfertility. The study is designed in the way to advise doctors on what to recommend the couples that come to a fertility clinic and are diagnosed with unexplained subfertility. For this reason the reference time for the treatment timing is the time of a couple's start of the follow-up.

In this section, the key aspects of the topic are introduced, the basis for understanding why the problem is interesting and why it is of interest to be researched. The first subsection defines the condition of unexplained subfertility, discusses certain statistics on this topic and argues why this research is important to conduct. Then, next subsection introduces the IUI treatment, describes the procedure within one IUI cycle, potential risks that accompany it, provides a thorough comparison with expectant management (EM) and another similar fertility treatment, discusses the success rates of all three approaches and summarizes the research done on the IUI so far, with the emphasis on the research done on the timing of the IUI treatment relative to the start of the follow-up. Afterward, a subsection introducing the dataset used in this study is included. Second to last, a subsection on the aim of the thesis is provided, including the research questions and methodology chosen on how to answer them. Lastly, a subsection describing the organisation of the rest of thesis report is included, providing a deeper overview of its structure and setting expectations on how the analysis will be conducted.

1.1. Unexplained subfertility

This research analyzes the IUI treatment timing for couples with unexplained subfertility. Subfertility is defined as not being able to get pregnant naturally for at least 12 months. After asking for medical help, couples first undergo a so-called fertility work-up, testing for e.g. progressive sperm count and fallopian tubes blockage. The factors tested cover the common causes why the couple could have difficulties getting pregnant. However, when the factors' values are within normal ranges, the couple's subfertility is considered to be unexplained. This makes treating their subfertility more difficult, since treatment is given only with a hope of help but without any certainty. Therefore, most often the first assisted reproductive technique (ART) is to try the IUI treatment, due to its low invasiveness and lower costs compared to the other fertility treatments.

Furthermore, unexplained subfertility is a common problem among couples wanting to get pregnant. USA national statistics shows that approximately 1 in 8 couples are affected by some kind of subfertility problem. It is estimated that around 10% of those are unexplained subfertilities.[1]. This implies that millions of couples deal with this problem all over the world and thus, research on treatment effects for this type of couples is of strong medical interest.

1.2. Intrauterine insemination

IUI is a fertility treatment, used to help couples with diagnosed subfertility to get pregnant. It is the first in line of fertility treatments, due to its non-invasive approach and lower price compared to other approaches. It is performed by inserting concentrated sperm sample into the uterus with a thin needle, at the time of woman's ovulation. IUI has been used for human reproduction from the end of the 18th century. It is believed that

[1] The data on subfertility statistics is taken from https://www.fertilityanswers.com/13-stats-know-infertility/

Arabs used this techniques for breeding stallions already in the 14th century. However, first human pregnancy resulting from IUI is considered to be the one in 1784 performed by Lazarro Spallanzani, helping a couple where the man had hypospadias[2].

The treatment is aimed to help couples with male and generally unexplained subfertility. Unexplained subfertility means that there is no apparent reason for not succeeding in getting pregnant. On the other hand, male subfertility can include low sperm count, low motility of sperm cells, impotence and others. When possible, the semen is acquired by masturbation and then concentrated, meaning some of the fluid has been removed, and when needed, sperm is treated with certain drugs to increase motility or to select healthy sperm cells for insemination. When semen cannot be acquired in this way, physicians can extract sperm from testicles surgically. Lastly, IUI is also used for women wanting a child without a partner, thus using a sperm from sperm bank. Figure 1.1 illustrates the processing of semen sample and injecting it into woman's uterus[3].

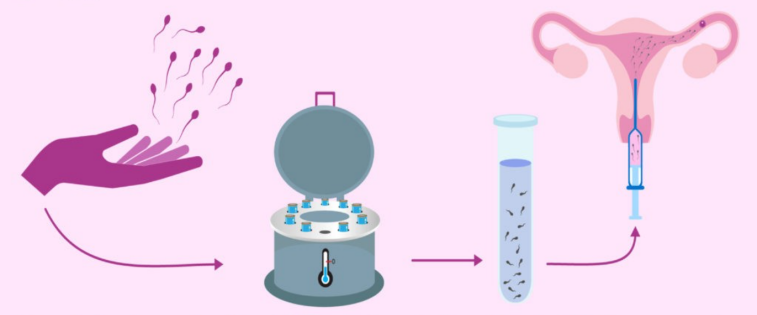


Figure 1.1: Illustration of the IUI steps: first the semen sample is collected, then processed and lastly injected to the woman's uterus

Furthermore, IUI can be accompanied by a hormonal therapy called ovarian stimulation (OS). This include medication stimulating the growth of ovarian follicles, keeping oocytes in follicles and finally rupturing the mature follicles and releasing the eggs. This happens exactly 36 hours after the medication has been given. This approach allows clinicians to control the menstrual cycle before and during ovulation, provides collection of multiple eggs through ovarian hypersimulation, endorsed thickening of the endometrial wall in uterus, ovulation in anovulatory women and more accurate timing of ovulation in general.

However, the OS treatment often induces negative side effects and increases potential risks. All women who underwent OS experience enlarged ovaries, and some develop potentially dangerous ovarian hyperstimulation syndrome (OHSS). This is a state characterized by abdominal pain and nausea, shortness of breath, fluid retention and leakage into the abdominal area, weight gain, blood clots and others. OHSS is categorised in mild (around one in three women), moderate and severe (around 1% of women) category[4]. Women are closely monitored after receiving the hormonal therapy, and if it is observed that ovaries have swollen, nausea medication and drainage of the fluid are available as treatment against the symptoms. Usually the symptoms resolve within a few days naturally, except for when pregnancy has been achieved, symptoms can last up to a 2-3 weeks.

Moreover, multiple pregnancies can often occur as a consequence of multiple follicles being matured. Multiple pregnancy presents a risk for both mother and the babies. Babies are often born underweight and can have body defects due to sharing the womb with their siblings. Moreover, probability of miscarriage is higher than in a single pregnancy, as well as twin-to-twin transfusion syndrome. Furthermore, the mother is in a higher risk of gestational high blood pressure, gestational diabetes, anemia, C-section delivery and more severe bleeding after giving birth.

If OS is given next to IUI, often every second month is taken as a rest from the medication, which allows ovaries and the whole body to recover. However, if the woman does not suffer from any fertility problems mentioned, it is not necessary to conduct the OS with the IUI. Nevertheless, it has been shown that OS increases pregnancy probability in ovulatory women, see for example [15], and thus it is recommended and often taken during IUI when dealing with male and unexplained subfertility.

[2]The information on IUI history taen from Zhu, Tian, "Intrauterine Insemination". *Embryo Project Encyclopedia* (2009-07-22). ISSN: 1940-5030 http://embryo.asu.edu/handle/10776/1992

[3]The image in the figure comes from the source: http://www.keyafertility.com/iui.php

[4]More information and statistics on OHSS can be found on the website https://theduff.co.uk/ovarian-hyperstimulation-syndrome-ohss-ivf/

Next to IUI, other ART approaches include ovulation induction, IVF, intracytoplasmic (morphologically selected) sperm injection, egg and/or sperm donors and surrogacy among the others. A Canadian survey reported that 75% of couples have tried to get pregnant at some point during their relationship, and among those around 15% of couples has sought medical help for this[5], making ART of significant importance to research and develop further. Each ART approach solves certain set of problems in woman's and/or man's subfertility, for example irregular ovulation, impotence, low number of motile sperm cells, unhostile uterus and many more. Closest in procedure to IUI and the most popular fertility treatment in general is IVF. It is based on extracting multiple matured eggs through OS treatment, acquiring semen sample and processing it, then fertilizing the eggs outside of the body and returning the zygotes to the uterus. This procedure also allows to freeze the eggs, sperm cells or zygotes and use them for later if needed. Thus, OS is given once and more attempts to get a lasting pregnancy can be done after that. Furthermore, since hormonal therapy is necessary practice for this procedure, the risk of OHSS and multiple pregnancy is high. Thus, IVF is more a invasive and more expensive procedure when compared to its counterpart IUI. For this reason, couples often start with a few cycles of IUI before giving IVF a try, even though IVF has been confirmed by multiple studies to have high success rates of pregnancies. In studies analyzing effectiveness of IUI this common trend leads to informative censoring.

USA national survey from 2017[6] shows that IVF success rate is about 40% for one IVF cycle for a woman under 35 years old, after which this probability significantly decreases. Moreover, [40] shows that due to the technological improvements, IVF success in the first 5 cycles is above 80%, despite of less zygotes being implanted. Lastly, data from CYN Fertility clinic [7] comparing success rates of one cycle of IVF and IUI based on female age covariate can be seen in Figure 1.2. It shows a decreasing trend of pregnancy probability in the female age covariate, and more importantly the significant difference in success rates of IVF and IUI for younger women, namely 50% versus 12% for age below 35 years, and decreasing to 9% versus 5% for age above 40 years respectively. Conclusion that can be drawn is that for higher age, IVF yields similar results to IUI and therefore IUI is dominant treatment option in that case. However, before being 35, a woman needs to decide if the higher success rates of IVF are worth the financial costs and undertaking a more invasive treatment.

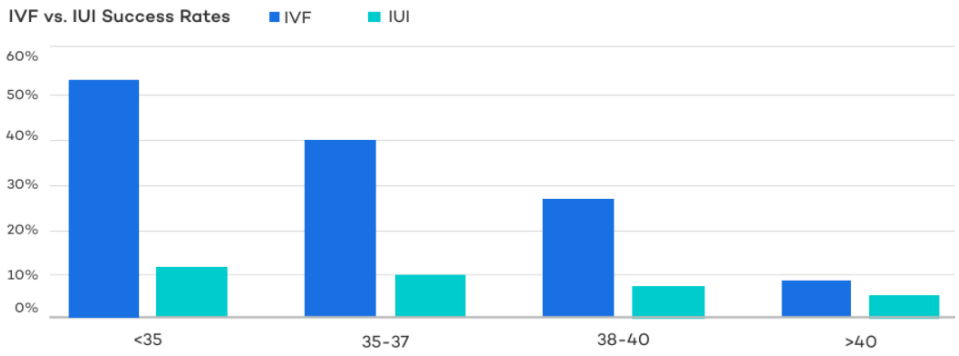


Figure 1.2: Comparison of IVF and IUI success rate for one cycle based on female age covariate

Still, the crucial question is whether IUI significantly increases chances of pregnancy and whether it is worth undertaking. There have been many studies exploring the chances of pregnancy with and without IUI treatment. Certain studies, for example [32], show no improvement in pregnancy probabilities when IUI has been given compared to trying to conceive naturally, which also called expectant management. However, in many more studies it has been shown that the chances of successful pregnancy differ from couple to couple, based on their fertility indicators, such as female age, duration of trying to get pregnant, male sperm count, percentage of progressive sperm and tubal blockage [19]. It is difficult to give an average success rate

[5] Data on the Canadian fertility survey statistics can be found at https://www150.statcan.gc.ca/n1/pub/82-003-x/2012004/article/11719-eng.htm

[6] USA national survey on ART can be found here https://web.archive.org/web/20200204120048/https://www.sartcorsonline.com/rptCSR_PublicMultYear.aspx?ClinicPKID=0

[7] Overview of statistics of IUI success rates based on the mentioned covariates and more can be found at https://www.cnyfertility.com/iui-success-rates/

of IUI procedure, since the probabilities vary significantly based on the couple's covariates, and therefore the estimates are volatile and can significantly differ from study to study. To give an example of variability of pregnancy probability from couple to couple, it is known that the quality of eggs decreases significantly after the age of 35. Furthermore, tubal blockage decreases chances of success, even though OS can improve success rates in these cases significantly, but this comes with the financial and health-related price. Moreover, the medication taken during the treatment influences the outcome results, as well as the number of IUI cycles which differs per study. Some studies report success rates after 3 or 6 cycles, or after a period of 1 year. In the Netherlands IUI is given in a maximum of 6 cycles in a row, after which, if no pregnancy has occurred, a more invasive treatment is started.

Nevertheless, the average success rate of 1 IUI cycle is shown to be just above 10%, as mentioned in [8]. This success rate decreases over time, which represents that the IUI is not a proper solution for the infertility problem a couple has. An extensive study [17] on IUI shows various conclusions on certain problems the couples might have and concludes that the female age is the strongest indicator of IUI success, the effect of which is also seen in Figure 1.2. On the other hand, IUI together with OS was extensively studied in [8]. They observed 39% pregnancy rate after 3 cycles and 57% after 6 cycles based on 1035 couples in the study. However, the extremely high pregnancy rates can be attributed to the young female population considered, with an average age of 28.9 years and standard deviation of 5.4 years. Furthermore, the probability of achieving natural pregnancy in 1 year-time based on 6 covariates (female age, subfertility duration, tubal blockage, doctor referral, type of subfertility and progressive sperm count) can be estimated on the Dutch association website Freya[8]. This website provides accurate estimates for couples who satisfy certain criteria, such as female age below 38 and subfertility duration being at least 1 year. The results come from the Dutch fertility study conducted in 2000-2004.

There is no scientific consensus about which couples should undertake IUI and which should not. [9] performed a randomised intention-to-treat study on 201 couples and concluded a significant difference in pregnancies from the two groups, namely 31% in the treatment group versus 9% in the control group. Intention-to-treat is a study design that randomly assigns couples in treatment and control group, but they are able to choose differently due to their specific situation. On the other hand, IUI can improve not too severe male subfertility and it was seen that for couples with higher chances of natural pregnancy, IUI does not improve chances significantly, while for the couples with lower chances IUI can be beneficial in increasing pregnancy probability [36].

Lots of research has been done on timing of steps within one IUI cycle, such as [41], [19] and [17]. These research topics include understanding the ovulation and learning how to control it, washing the sperm and its influence on the babies, the timing of performing IUI compared to the ovulation time and the number of IUI cycles performed. The latter means conducting a double IUI, i.e. before and after ovulation, which has been shown to increase chances of pregnancy followed by male subfertility, but not unexplained subfertility [19].

However, not a lot of research has been done on investigating if couples should start with the IUI immediately when consulting a fertility clinic, or wait a few more months, to see if the pregnancy occurs naturally during this time. The state of consulting a doctor but still trying to get pregnant naturally and without any fertility treatment is called expectant management. It represents the natural course of a couple's journey to pregnancy, where visiting a doctor does not make any difference and therefore it represents what would happen if no treatment influence is made on their time to pregnancy. [37] compared cost-effectiveness of EM, IUI and IVF and their combinations throughout 3 years, each being followed for one year exactly. The strategies in this study are noted as names of the treatment during each of the three years analyzed. For example, EM-IUI-IVF is a strategy representing being on EM during the first year of a followup, then receiving IUI cycles during the second year, and receiving IVF cycles during the third year of the followup. Through causal inference authors concluded that strategies EM-IUI-IVF and EM-EM-IVF are dominant over the other combinations and can be compared further through assumed monetary value per live birth and the known prices of the treatments. The analysis reported in this thesis employes a similar approach, but with differently defined strategies.

Furthermore, a study from [4] shows that in 3 years time, starting immediately with the IUI or waiting for 6 months on EM yields the same cumulative pregnancy probabilities. A study described in [7] has investigated a similar topic of whether couples should take a few cycles of IUI before starting with the IVF treatment. They

[8] Pregnancy probabilities based on 6 covariates for each couple can be estimated on the website https://www.freya.nl/probability.php

considered the cost-efficiency compared to the pregnancy rates, and concluded that IVF should be used immediately. This all leads us to conclude that there is no simple answer to when should a fertility treatment be started and which one is the best to undertake. More research studies are required, preferably conducted on larger number of couples, due to the variability of the prognoses from couple to couple.

To sum up, the available literature implies that IUI is a useful fertility treatment, yielding a significant increase in success rates compared to the expectant management for certain subfertile couples, and has clear advantages over IVF in situations when fertility problems can be avoided by refining the semen sample and controlling the ovulation cycle.

1.3. Dataset

The dataset that was used in this study comes from a Dutch national project on fertility treatments, in the time period between January 2000 and October 2005. 7 out of 38 hospitals in the study collected data on the IUI treatment. From this data, couples with unexplained subfertility were chosen for this project. This includes 1896 couples in the time span of at most 4 years of follow-up. The variables tested during the couples' work-up that may influence the future pregnancy probability are included in the dataset. These are

- female age
- subfertility duration
- gynecologist referral
- infertility type
- fallopian tubal blockage
- progressive sperm count

Other variables in the dataset, such as exact dates of start and end of a couple's follow-up and the exact hospital the couple was treated in, are excluded from the analysis. More on the description of the 6 variables included in the study is provided in Section 4.

Moreover, a variable representing the time of starting the IUI treatment is included in the dataset. Furthermore, the exact number of IUI cycles after this time is not available in the dataset. However, any pregnancy in a few months after the IUI cycle can still be considered a consequence of the IUI if the hormonal therapy has accompanied IUI. Therefore, the assumption in this research is that all pregnancies after the first IUI cycle has been started are a consequence of the IUI treatment. Lastly, the exact time of first month of pregnancy is recorded, or the time a couple has left the study. Brief statistics about the data are that 406 (21.41%) couples have become pregnant without any fertility treatment started, while 863 (45.52%) couples have started with the IUI treatment during their follow-up, out of which 163 (18.89%) couples have become pregnant during the follow-up. More in-depth statistics on these variables are provided in Section 4.

1.4. Aim of the thesis

There are two important aspects to this study. The first one is to investigate the influence of the IUI timing on the pregnancy probability. The first step in understanding its effect is to compare the pregnancy probability predictions for a given couple, given different treatment timings. These probabilities are estimated and averaged over the whole population, which are in this case all couples from the dataset introduced in Section 1.3. Five different fixed points of treatment timing are analyzed, namely starting at 0, 3, 6, 9 months or not starting at least until the end of the time frame of this study, chosen to be 1.5 years after diagnosis. Discussion on why these time points were chosen is given in Section 7. Since it is expected that the treatment always increases the pregnancy probabilities, even only slightly for some couples, a trade-off needs to be made based on the negative side effects of the IUI for the couple. These drawbacks of taking the treatment include the potential risks of the hormonal therapy accompanying IUI and the financial costs for the couple. Thus, in this thesis we facilitate such a trade-off based on the number of couples that are being treated when all couples start with the treatment at the same time relative to the completion of their fertility work-up phase.

To sum up the the first goal of the thesis, the explicit research question to answer is:

1. How does the time of the start of the IUI relative to the work-up completion influence the probability of pregnancy for couples with unexplained subfertility in the first 1.5 years from their work-up completion?

Then, an overview of the methodology chosen to tackle the above research question leads to the second aspect of this study. The causal inference framework is necessary to make the pregnancy probability predictions for treatment timings that were not necessarily observed in the real life. By quantifying the causal effects, the framework can answer causal questions about what would happen given that some counterfactual event has taken place. Since statistics can directly assess only the correlations among observed variables, causal inference requires specific tools in order to provide unbiased results from a statistical analysis. The theory describing it is quite new and it is crucial to connect it with various statistical settings and models in order to make the causal analyses more accurate and applicable to various research problems.

The other part of the methodology considers the model used to make the predictions of the pregnancy probability. Due to the data being of the time-to-event form and possibility of an intermediate event, namely the start of the treatment, a multi-state model is used to describe the transitions between the states. These states include i) trying to conceive naturally, ii) receiving the IUI treatment and continuing to trying to conceive and iii) the state of being pregnant. Multi-state models offer more flexibility in the modeling than using only one model to predict the probability to achieving pregnancies, by including the time to treatment variable through, for example, a time-varying covariate.

Therefore, the second goal of this study is to make the connection between multi-state models and causal inference. There has not been much research combining the two frameworks. Only [10] has been found so far. The multi-state models are currently often used in prediction research, but their connection with causal inference has not yet been explored. Connecting these two frameworks can open up more precise analysis of the causes of why some treatments work, instead of relying only on the observed probability of their success. Among the causal estimation methods, G-computation is chosen to be applied, and thus the connection of this method with the multi-state models is the second focus of this study.

To sum up the second goal of the thesis, the explicit research question to answer is:

2. Is it feasible to use multi-state models and G-computation to answer causal questions?

1.5. Organisation of the thesis report

The rest of the thesis report is organised as follows. First the survival analysis and all relevant topics related to it are introduced in Section 2. This includes the basic modeling concepts, such as survival probability and hazard functions, then the Cox proportional hazards model and the multi-state models. Then, Section 3 covers the theory on causal inference relevant to this research problem. The estimands of interest in this study are defined and the method used for quantifying the causal effects, the G-computation method, is described. These two sections are designed to present the theory, without any application to this research problem, but to cover all aspects relevant to the application. Next, the dataset used for this study is described in Section 4, together with the descriptive analysis of the covariates used for modeling. After that, in Section 5 the theory on survival analysis and causal inference is applied on the research problem at hand. This includes describing the model used for modeling the pregnancy probability of the couples, addressing all the necessary modeling assumptions related to the survival and causal frameworks and applying the G-computation method in order to estimate the estimands of interest. After this, Section 6 presents the results of the analysis explained in the previous section. The results are briefly commented on while presented. Then, Section 7 includes a discussion on the whole research problem, the chosen methodology and the posed research questions, conclusions drawn from the research study and a few suggestions of future research building on this work. Finally, Appendices A, B and C contain parts of the analysis that would slow down the flow of the report, but are important to be provided. Lastly, Appendix D includes the R code used to conduct the analysis.

2

Survival Analysis

This thesis project deals with data on time until pregnancy of a couple, with possible intermediate event of starting the IUI treatment. Survival analysis is the tool devised to analyze this time-to-event data and models from this theory, namely Cox proportional hazards and multi-state models, are the cornerstone of this thesis methodology. This section covers the needed theory of survival analysis, before connecting it to the problem at hand in Section 5.

In the rest of the section a brief history overview of the survival analysis is presented, as well as definitions of its basic concepts of interest, namely survival and hazard functions. Furthermore, assumptions needed to estimate the mentioned concepts from survival data are laid out. Then, as a popular model to use, Cox proportional hazards model is used in study and its basics are introduced. Lastly, the multi-state models, an extension of the basic survival models, are introduced and commented on.

2.1. Introduction

Survival analysis as a branch of statistics dates back to the 17th century, when the first life table was designed, and has been intensively improved upon from the middle of the 20th century, when first regression models were introduced. With first applications being in biology and epidemiology, data often follows lifetime span of individuals, so the event of interest is usually referred to as 'death'. This analogy accounts for the name 'survival analysis' as analysis of the survival time before individual's death. Often there is an intermediate event of interest, called 'illness'. This is an event that may or may not be observed before 'death', but can have strong influence on the probability of that event.

Epidemiology and thus survival analysis mostly relies on observational data, especially when causes of diseases are analyzed, or the effects of treatments that are difficult to randomly assign. Patients in the study start the trial 'alive', i.e. before the event of interest has occurred. The time of 'death' varies from patient to patient and can be influenced by some of the patient's characteristics, called covariates. These can include patient's age, gender, work status, but also blood pressure or certain hormones' levels at the time of measurement. Usually covariates are measured at the beginning of the study. In addition there may be ones that vary in time during the followup.

Moreover, even with a suitable study design, one shortcoming of the epidemiology trail comes from the length of human life: researchers cannot follow each patient until their death, but need to end the trial after a period of time decided in advance, usually a few years. This leads to incomplete information on the patients who did not experience the event of interest in the period of the trial. However, the information they provide to the stud is still valuable and should not be disregarded. Therefore, these patients are censored at the moment of the end of the trial, or the moment they leave the study prematurely. For example, if they start a different treatment than the one analyzed, they are excluded from the study. This is called right censoring. It is also possible to accommodate for left censoring, which occurs when a patient was followed only from after the possibly unknown time that marks the beginning of the followup period and for the interval censoring, where the exact time of this event happening is unknown. In this thesis, we are concerned only with right censoring.

Survival analysis has achieved the largest improvement by the introduction of Kaplan-Meier survival estimator, proposed in [14]. This allowed to non-parametrically estimate the survival probability with one curve, instead of large life tables. Another non-parametric survival curve estimator is Nelson-Aalen estimator, proposed and developed during 1970s. These allowed survival models to be developed, most popular of which is Cox proportional hazards model, proposed in [2]. It is a semi-parametric model and its most important feature is the proportionality assumption. It states that the covariates' influence on the hazard curve is constant over time. The model is explained in detail in Section 2.4. Even though this is the basic and still the most popular approach to the survival analysis, other models have been proposed. These mostly include accelerated failure time models, which are parametric and do not assume the proportionality such as Cox model does. These rely on a different assumption which says that the covariates influence accelerates or decelerates the prognosis of some event of interest by a constant. They are useful when stages in the disease prognosis are present and their underlying mechanics are known.

2.2. Basic concepts

At first glance, it is hard to visualise a survival dataset. One way of getting an insight into what had been happening with the participants is through an empirical survival probability over the trial period. Let T be a positive random variable representing the time of 'death' of a patient relative to the start of the trial. Then, survival probability until at least time t is the probability that a patient from the population survives until time t. It can be written as

$$S(t) = P(T > t)$$

On the other hand, one can define the hazard at each time point. The hazard function captures the rate at which the patient may die just after time t given they survived until time t. This can be written as

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

The hazard function is a useful tool to illustrate the prognosis over the course of a disease, since it gives conditional estimates at each moment of what will happen next, if nothing happened until that moment. This can be visualised through the cumulative hazard function, sometimes making it easier to see the changes in hazard. It can be written as

$$\Lambda(t) = \int_0^t \lambda(s) ds$$

Survival and hazard functions are closely related to each other and one can be expressed in terms of the other. The hazard function can be expressed through the survival function as

$$\lambda(t) = \frac{1}{S(t)} \lim_{\Delta t \downarrow 0} \frac{S(t + \Delta t) - S(t)}{\Delta t} = -\frac{d \log S(t)}{dt}$$

and the survival function can be found through the cumulative hazard function through the following relationship:

$$S(t) = \exp(-\Lambda(t))$$

The most common way to empirically estimate the survival probability curve is through the Kaplan-Meier estimate. To define it, let us first discretize the time into chronological time points when the events take place, namely $0 < t_1 < t_2 < \ldots$. Let n_j be the size of the risk set at time t_j, where $j = 1, 2, \ldots$, containing all 'alive' subjects in the dataset at time t, and let m_j be the number of patients that experienced the event of interest at time t. The Kaplan-Meier estimate is then defined as

$$\widehat{S}(t) = \prod_{j: t_j \leq t} \left(1 - \frac{m_j}{n_j}\right)$$

In the same fashion, the hazard function can empirically be approximated by

$$\widehat{\lambda}(t_j) = \frac{m_j}{n_j}, \quad \text{for } j = 1, 2, \ldots$$

2.3. Assumptions and biases

Every modeling area relies on certain assumptions. The main assumptions in an area of interest come from the theoretical needs of models used, mostly designed in a way that the most of the usually analyzed problems satisfy them. Of course, exceptions from the assumptions from time to time need to be made, and solutions for those cases have been researched as well. Furthermore, when certain modeling assumptions are violated, biases can occur in the results of the study. Especially when dealing with a complex real-life problem, it is often difficult to account for all influences from the real world into a model. For these reasons it is crucial to be aware of the assumptions the analysis relies on and potential biases to watch out for when assessing the significance of the results.

The following subsections introduce the most important assumptions and biases in survival analysis in general, and list some common ways to deal with them when necessary. The assumptions and biases will be commented on again in Section 5.2, from the perspective of this thesis project.

2.3.1. Selection bias

Selection bias is a wide concept covering many types of biases which all come from the fact that certain selection needs to be made in the process of conducting the research, be it selection of patients participating in a medical study, data to test the results on or trying out multiple models and analysis approaches until researcher obtains results that satisfy his or hers expectations of the study. It is not specific only to survival analysis, but is important to accommodate for in observational studies such as this one.

In survival analysis, one of the concerns is the patient selection bias, which comes from patients often being volunteers for the studies and therefore the sample of population considered is often not a perfect random sample from the population. Many ways have been proposed to correct this bias. The general approach to deal with this is to condition the outcomes of the patients on their covariates. Therefore, the covariate space does not have to be covered equally to obtain estimates on the population level, but it is important that the whole covariate space is covered by enough samples, and that the distribution of the population's covariate values is known. If the space is not covered well enough, the results cannot be generalised for the whole population, but only to the part of the population similar to the participants of the study.

2.3.2. Survivor bias

Survivor bias or the survivorship bias comes up when the focus is put only on the individuals that passed a certain selection criterion, and the ones that did not are disregarded from the analysis. Example of this is making a financial performance analysis including only companies that are still in business, or when traits of successful people are analyzed without taking into account the traits of the individuals who wanted to become successful but failed on the way. In survival analysis this manifests in a higher likelihood of patients of better health to receive the treatment, compared to the ones who die earlier. The latter ones are then falsely modeled as less likely to get the treatment, where in fact treatment might have been randomly assigned to all patients throughout some time period, but they did not manage to live to get theirs. For the same reason, this type of bias can also cause implications of higher success of the treatment.

The survivor bias is a type of time-dependent bias in the survival analysis, and can be decreased or even resolved by a time-dependent analysis. This means allowing covariates and treatment variable to change over time, thus avoiding the confusion of the treatment effect versus the effect of being of better health at the beginning of the study.

2.3.3. Independence assumption

The independence assumption states that the censoring time and the time of the event of interest are independent conditioned on the patient's covariates. This is one of the foundation assumptions that the whole survival analysis relies on and most of the already developed models rely on it. In case the assumption holds, the censoring can be completely ignored. This is done through the assumption that the behaviour of a patient who was censored would in expectation follow the behaviour of the rest of the patients (with similar covariates) still in the study.

This, however, might not be true for all research problems. Often it is violated in observational studies, where patients can leave the study for any personal reasons and so some may be more likely to leave than others. A common example of the informative censoring is when multiple treatment are available for certain condition, but only one is being studied. Then, patients with worse prognosis often opt for a more invasive treatment than the one being researched and thus leave the current study.

The way to approach this problem is to incorporate all the knowledge one has about the censoring: the characteristics and the time of the patients being censored. For example, in an analysis with competing risks, a framework that allows multiple final events of interest, this means that if a patients experiences one event, he or she will not experience any other event of interest. This information is included in the analysis by leaving this patient in the risk sets for the events of interest until the study end, therefore simulating the fact that he or she did not experience those events of interest in this time. Some methods to deal with informative censoring come from other large areas such as dealing with missing data, sensitivity analysis of the estimands, in order to be aware of the best and worst cases for the study results, or paying lots of attention to the possibility of informative censoring when designing a study, as discussed in [30].

2.4. Cox proportional hazards model

Most survival models are based on distinguishing a so-called baseline hazard and the effect of the covariates on the hazard throughout the time. Proportional hazards models are based on the proportionality assumption stating that the change in a covariate implies a constant change in hazard in each time point. This leads to proportional hazard curves for any two different patients. [2] introduced a model now known as Cox proportional hazards model, which has become the most commonly used in survival analysis. Next to the proportionality assumption, in its basic form it assumes linearity of the covariates' effects to the log hazard function. If λ_0 stands for the baseline hazard, $\mathbf{X}$ is the covariate vector and β is a coefficient vector, then the hazard function is modeled as

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{X}) \tag{2.1}$$

The linearity can be avoided by using some function of the covariates, but this is equivalent to performing some transformation of the data and using the linear form of the transformed data.

Since the main assumption of the model is that the hazards are proportional, the analysis is often interpreted through the hazard ratios, which are the ratios between the baseline hazard and the hazard after a unit change in a certain covariate. The hazard ratio for a covariate j equals $\exp(\beta_j)$ and covariates with hazard ratios larger than 1 are interpreted as having a positive effect on the hazard, therefore implying shorter life-span for patients with higher values of that covariate.

2.4.1. Estimation

The baseline hazard and the beta coefficients can be easily estimated from the data. The beta vector can be found through the partial likelihood maximisation. The reason for using the partial and not the full likelihood is that the baseline hazard function has no restrictions on its form and therefore one would need to perform maximization over the infinite dimensional vector space, which is not possible. Therefore, Cox proposed to use the partial likelihood instead. Here we provide the definition of the partial likelihood in case of no tied events, meaning that m_j equals 1 for every j. Then the partial likelihood consists of a product over event times of hazards of the patient who experienced the event at t_j over the sum of hazards of all patients in the risk set at that time, labeled as R_j. A more detailed text on the partial likelihood can be found in [13] or [12]. If $\mathbf{X}_j$ is a vector of covariates of a patient that experienced the event of interest at time t_j and N is the number of patients in the dataset, then the partial likelihood can be written as

$$L_p(\beta) = \prod_{j=1}^{N} \frac{\exp(\boldsymbol{\beta}^T\mathbf{X}_j)}{\sum_{i\in R_j}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)}$$

The estimated $\hat{\boldsymbol{\beta}}$ which maximizes the partial likelihood is then used for estimating the baseline cumulative hazard. A well-known and most often used such estimator is the Breslow's estimator. In [13] the estimate is derived from the MLE in more detail, and here we write it in its final form

$$\widehat{\Lambda}_0(t) = \sum_{j:t_j\le t} \frac{1}{\sum_{i\in R_j}\exp(\hat{\boldsymbol{\beta}}^T\mathbf{X}_i)}$$

2.4.2. Time-varying extensions

Certain relaxations on the model can be made. One can allow for time-varying covariate values, which is useful in e.g. disease progress modeling, thus allowing for treatment covariate and other medical tests to be included over time. This does not influence the beta vector and the baseline hazard estimation, and it is easy to include the extention in the dataset through the long data format, explained in Section 4.1.

Furthermore, one can allow that the importance of the covariates differs in time, therefore time-varying beta coefficients can be used. In this case the proportionality assumption no longer holds. Thus, this extension is a way of dealing with the hazards non-proportionality in the model. It complicates the computation of the beta vector and certain assumptions on the form of the coefficients need to be made, after which the estimation procedure stays the same.

2.5. Multi-state models

Multi-state models extend the described framework by allowing multiple intermediate events and events of interest, here called 'transient' and 'final'/'absorbing' states, respectively. Once a patient is in the absorbing state, he or she can no longer transition to another state, thus marking the end of his or hers followup. The analysis is based on inspecting the possible transitions between the states, probabilities of each transition, often based on information on each patient, and prediction of the overall population distribution over the states at each time point.

These models are suitable to be represented by a graph, having the states on the nodes and possible transitions as arrows. Therefore, we often inspect the model by inspecting its graph and graph theory is often used for analysis of more complicated graphs. Moreover, if transitions between two states are allowed to go in both ways between states, the model is called 'bi-directional'. This is opposed to a 'uni-directional' model that allows only one direction of transition between each two states. Our focus lies in the latter ones, without any recurrent events. This means that no loops or cycles are allowed on the graph. The simplest such multi-state model is a three-state model, often called 'illness-death' model, and an illustration of it can be seen in Figure 2.1.

The theory of multi-state models is a direct extension of the competing risks theory in the survival analysis. How to handle risk sets in the competing risk analysis was briefly covered in Section 2.3.3. In the multi-state setting, transitions starting in the same states are modeled as competing risks, with events of interest being entering one of the other possible states. Moreover, simultaneous entering and leaving the states implies that patients also leave and enter the risk sets. This provides knowledge about what happens with a patient after he or she has been censored from some transition model, allowing for a more accurate analysis.

2.5.1. Time scales

It is important to distinguish two possible ways of measuring the time in the multi-state models: the 'clock-forward' and the 'clock-reset' approach. The clock-forward time scale consists of tracking the time passed between the start of the patient's followup and an event of interest, which in a multi-state model is a state transition. This means that every followup starts at time 0, and time is continuously tracked through intermediate events until the final state has been reached or an individual has been censored. This is illustrated in Figure 2.2 on the top. The figure shows time intervals capturing the period a patient is in some state, and an arrow pointing to the moment in relation to which the time in that state is being measured, which is the start of the trial in this time scale.

On the other hand, clock-reset approach resets the time after each event occurs, i.e. when a new state has been entered. Then the time for each patient is measured in relation to the time the current state has been entered. An illustration of this can be seen in Figure 2.2 on the bottom. Again, the time intervals show periods of time a patient is in one state, and an arrow shows from which moment the time in that state is measured.

Both time scales contain the same amount of information and are easily transformed from one to another, but provide different approaches to the analysis. One possible difference is in which time point, start of the trial or entering a certain state, is more important for modeling the transitions in the model. By considering

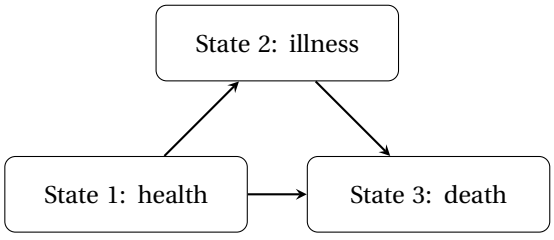


Figure 2.1: The simplest multi-state model: an illness-death model

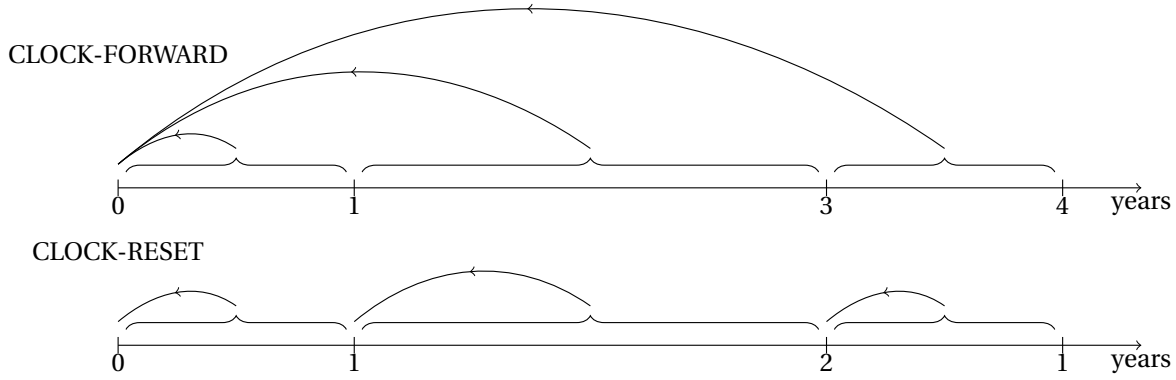


Figure 2.2: Clock-forward and clock-reset time tracking for a patient that experienced events of interest at times 1, 3 and 4 years after the start of the trial, i.e. he was in the first state for 1 year, in the second state for 2 years and in the third state for 1 year; the arrows represent the time point from which the time to event is tracked in each time interval

this reason, the time-reset approach will be considered in this thesis and more details are provided in Section 5.1.

2.5.2. Transition probabilities

The multi-state analysis is based on the transition probabilities. For each transition, one can define a separate hazard function, also called transition intensity. These are the hazard functions in a basic survival analysis in the case only this transition is analyzed. If one models a transition from state e to state f, noted by $e \to f$, and T stands for a time-to-event variable with event being transition to state f, then the transition hazard in both clock-forward and clock-reset time scales is defined as

$$\lambda_{ef}(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}$$

Similarly, the cumulative hazard for the same transition $e \to f$ is defined as

$$\Lambda_{ef}(t) = \int_0^t \lambda_{ef}(s)\, \mathrm{d}s$$

Furthermore, the survival function and its relationship to the hazard and cumulative hazard function are somewhat different than formulas in Section 2.1. Using the example of an illness-death multi-state model illustrated in Figure 2.1, probability of survival in state 1 depends on both hazards of transition 1 and 2, while the survival probability for state 2 resembles the formula we have already seen, due to hazard of only one possible event happening. For survival in state 2, one extra covariate can be added in order to have the extended Markov model, representing the time of arrival to the state 2. It is usually denoted with letter r, which is used as a subscript for hazard and cumulative hazard for transitions from state 2, and for survival function in state 2. Thus, the survival probability functions for this 3-state model can be written as

$$S_1(t) = exp(-(\Lambda_{12}(t) + \Lambda_{13}(t)))$$
$$S_{2,r}(t) = exp(-\Lambda_{23,r}(t))$$

With the use of the survival probability functions, functions representing transition probabilities can be defined. Modeling transition from state e to f, in a time interval (s, t), is done through modeling the probability of being in the state f in time t given being in the state e at time s. This means that the direct transition from state e to f can happen, or some different path can be taken. The transition probabilities for an illness-

death multi-state model can be written as in the following

$$\begin{aligned} P_{12}(s,t) &= \int_s^t \lambda_{12}(u) S_1(s,u) S_{2,u}(u,t) \mathrm{d}u \\ P_{23,r}(s,t) &= \int_s^t \lambda_{23,r}(u) S_{2,r}(s,u) \mathrm{d}u \\ P_{13}^2(s,t) &= \int_s^t \lambda_{12}(u) S_1(s,u) P_{23,u}(u,t) \mathrm{d}u \\ P_{13}^1(s,t) &= \int_s^t \lambda_{13}(u) S_1(s,u) \mathrm{d}u \\ P_{13}(s,t) &= P_{13}^2(s,t) + P_{13}^1(s,t) \end{aligned} \tag{2.2}$$

where $S_e(s,t) = S_e(t)/S_e(s)$. Moreover, P_{13}^2 stands for probability of path $1 \to 2 \to 3$, whereas P_{13}^1 for the probability of direct transition $1 \to 3$. These two summed up make for the total probability of reaching state 3 from state 1, written as P_{13}.

To estimate these transition probability functions, it is enough to plug the estimates of the transition hazard probabilities in the above formulas. For the clock-forward Markov multi-state models, one can estimate transition probabilities through the Aalen-Johansen estimator. It is based on the Markov property which can be used only in the clock-forward time scale, and since that is not in the focus of this thesis, it will not be introduced, but can be read about in [1]. For clock-reset time scale, there is no different estimator available than the transition probabilities from equation (2.2). For this study, estimation of transition probabilities using the transition equations will be described in Section 5.

3

Causal inference

This thesis poses a causal question of what would the pregnancy rates in the population be if all couples followed the same treatment strategy. In order to answer this question, causal inference needs to be used. This section provides the theoretical background of the general theory for answering causal questions using the model fitted on observational data, namely the structural theory. This section explains the importance of developing this theory precisely, the pitfalls when not using the theoretical setting properly, and how to conduct the causal analysis and safely draw conclusions on counterfactual scenarios.

The section is organised as follows. First, causal inference is introduced and its development over the past two decades is briefly covered. Then, the structural theory is described, with its notation and basic concepts. Next, three causal models, seen as restrictions of the structural theory, are introduced. Furthermore, the main estimands are defined and the assumptions of the causal framework are laid out and commented upon. Finally, G-computation, the method of interest used to estimate the estimands, is described in detail.

3.1. Introduction

Two separate frameworks can be distinguished in statistical analysis in general: prediction and causal inference. The framework is chosen based on the nature of a problem at hand and poses different end goals of the analysis. In the prediction framework one can use all available information about the variables in the model to predict the outcome as precise as possible. There it is reasonable to ignore the causal relationships, since variables not directly influencing the outcome can be informative for the probability of the outcome value. As opposed to this, the causal framework is used to describe the causal structure among variables and quantify the causal effects. The difference between the causal and prediction framework is nicely described in [31].

Causality is a broad term and is crucial for human understanding of the world. Some causal relationships are exact and can be expressed through equations, for example physical equations. Other causal relationships have certain noise present in them, or there is a lack of information to infer the exact relationships. These are assessed through statistical analysis. As Paul Holland states in [11], statistics is used to examine the "effects of causes", as opposed to the "causes of effects". He argues that this is the right approach to take, since one can describe the cause of an effect through various scales of the problem, going from physical objects into chemical reactions, thus the right view of the cause does not exist, or is forever disputable.

[23] points out that before the 2000s, causal analysis was rarely discussed, only sometimes mentioned when discussing the results, commented on in free interpretation and without any notation whatsoever, and often wrongly mistaken for the association relationships.

Moreover, the causal effects cannot be directly computed from the data nor from the distributions the data comes from. This is the demarcation line between correlation and causal relationships and concepts, as is emphasized in an causal inference overview [23]. For this reason, statistical analysis as it is does not provide broad enough framework for the proper causal analysis to take place in, and new notation needs to be introduced. This was started in [28] and [21], with their frameworks described in detail in the following subsections and later used in the application of the methodology to the research study in Section 5.

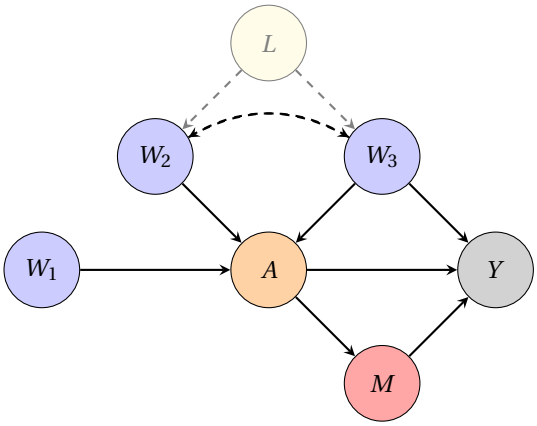


Figure 3.1: Representation of a causal relationships and correlations among variables in a graph; correlation between (W_2,W_3), represented by a two-sided dashed arrow, and their relation to A and Y, with a presence of a mediator variable; A=independent variable, Y=dependent variable, W_1=non-confounding variable, W_2,W_3=confounding variables, M=mediator variable, L=latent variable; colors chosen to emphasize different types of variables

3.2. Structural theory

In order to talk about the causal questions, assumptions, methods and algorithms, a mathematical theory providing unambiguous language is required. Following [23], the structural theory, sometimes called structural causal model (SCM), is presented. It is a general theory for causal inference, meaning that it accommodates all other theories developed in this area in a way that they can be viewed as its restrictions. It provides a framework subsuming three most commonly used approaches to the causal inference, namely structural equation modeling, potential outcomes framework and representation of causal structure through graphs. Moreover, other approaches are also covered by the theory, such as interventional [42], decision analytical [5], sufficient component [27] and probabilistic [33] approaches to causation. Lastly, the structural theory accommodates systematic axioms, definitions and algorithms to conduct the causal inference, and also focuses on "demystifying enigmatic notions" which include mediation, confounding, ignorability, exchangeability, comparability, etc.

The structural theory is presented through its graphical, structural and counterfactual aspects. The following subsections describe these aspects separately, but it should kept in mind that they fit together and make a more complete approach to the causal inference. Furthermore, throughout the following subsections, main concepts and assumptions will be pointed out as paragraph titles, which will help explaining and summarizing the SCM better. Lastly, the interest of this thesis lies in the Rubin's potential-outcomes framework, which will be later used in this more general SCM framework. For this reason, it is first presented as in [28] and then connected to the structural theory framework.

3.2.1. Graphical causal representation

Causation is easily visually represented through directed graphs. The approach was first mentioned by Wright in [43] who wanted to incorporate in his model the assumption, or a fact in this case, that symptoms don't influence diseases but diseases influence symptoms. To achieve this distinction he used graphical representation of causal relationships, and called them path diagrams.

When setting up a graph of the model, variables are set as nodes, and arrows represent their pairwise relationships. In order to analyze causation, it is important that a graph is non-cyclic. An example of a causal graph can be seen in figure 3.1. One-sided arrows represent the causal relations between the two variables, more specifically that one variable influences the other, and the second variable does not have an effect on the first one. When correlations are present, they can be represented through double-sided arrows between the variables. These correlation arrows represent that there is a latent common cause of the variables that is not present in the set of variables in the model. A double-sided arrow in a causal graph is equivalent to adding a new variable that has a causal influence on the correlated variables, which is also shown in figure 3.1 with a faded latent variable L.

Moreover, the absence of certain arrows in a causal graph is a consequence of the so-called independence

assumption, and is a part of the assumptions on the causal structure of the variables. A model without any assumptions on the causal structure would imply that all variables can be correlated, and so the model would be susceptible to pick up patterns from the data which are generally not there.

Usually, a causal question refers to the influence of one variable to another. Common notation here is to call the variable that causes some effect as the independent variable, and the variable on which the effect is being made as the dependent variable. The independent variable can have a direct or indirect influence on the dependent variable. Namely, if there exists a variable on the path between dependent and independent variables, it is called a mediator, and it requires a slightly more complicated approach than the analysis of a direct influence does. The mediator can be seen in Figure 3.1 as variable M. Methods to infer the causal effect when mediator variables exist and can be read about in De Stavola et al. [6], but are not in the focus of this thesis. From now on, the models and methods concern the direct causal relationships exclusively.

Another division of the variables in causal models is based on the origin of their values. Variables whose values are determined from outside of the model are called exogenous variables, and the ones whose values are determined through other variables in the model are called endogenous variables. So, in the graph from Figure 3.1, variables W_1, W_2 and W_3 are exogenous, and variables A, M and Y are endogenous.

Confounding

Important concept in causal inference is confounding. This is the situation when certain variables influence both the dependent and independent variables. The confounding variables might distort the relation between exposure and outcome variable. For this reason, the method to avoid this biased causal effect is to condition on the all confounding variables and analyze the causal effect in each strata of the confounding variables. However, it is often difficult to even consider all the possible confounding variables, and lots of them might be difficult to quantify, for example in social sciences. A set of variables sufficient to condition on needs to be found.

The back-door method is easy and most commonly used criterion to find this sufficient set of variables. It is based on finding a set of variables that block all backdoor paths from the independent to the dependent variable which contain the arrow into the independent variable, i.e. they are not descendants of the independent variable. Then the change in the independent variable can be directly compared to the change in the outcome variable, due to only paths from the independent to the dependent variable being susceptible to change.

To illustrate confounding effect and how the back-door criterion actually works, the model from Figure 3.2 is used. We assume that all variables influencing the dependent and independent variable are included in the analysis. Back-door paths from A to Y are

$$
\begin{aligned}
&A \leftarrow W_4 \rightarrow Y,\\
&A \leftarrow W_3 \leftarrow W_1 \rightarrow W_4 \rightarrow Y,\\
&A \leftarrow W_3 \leftarrow W_2 \rightarrow W_5 \rightarrow Y \text{ and}\\
&A \leftarrow W_3 \leftarrow W_1 \rightarrow W_4 \leftarrow W_2 \rightarrow W_5 \rightarrow Y.
\end{aligned}
$$

It should be noted that W_4 does not block the path last path listed above, since all arrows are entering this variable. Therefore, set $\{W_4\}$ is not a sufficient set to condition on in order to correct the confounding bias. Furthermore, it is important to emphasise that it is not advisable to condition on all variables apart from the dependent and independent variable. This has long been missed in the causal studies, but [29] and many others show that including unnecessary variables may increase the bias of the causal estimates. The reason for this are examples when two variables are independent, but when conditioned on certain strata of another variable, they show traces of dependence among them. Therefore, the back-door criterion gives us the following sets as sufficient sets to condition on:

$$\{W_1, W_2, W_4\},\ \{W_1, W_4\},\ \{W_2, W_4\}, \{W_3, W_4\}, \{W_5, W_4\},\ \{W_3, W_2\}, \text{ and } \{W_1, W_5\}.$$

Moreover, since e.g. $\{W_1, W_4\} \subset \{W_1, W_2, W_4\}$, the latter set is reducible, meaning that a subset of it is a sufficient set to satisfy the back-door criterion as well.

The backdoor adjustment theorem, proved in [20], says that if there exists a set of variables F such that they blocks all the back-door paths between the independent variable A and the dependent variable Y ar-

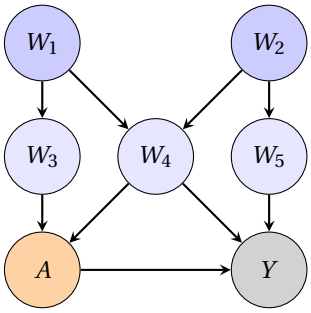


Figure 3.2: Path diagram used for illustrating confounding and the back-door criterion for a sufficient conditioning set of variables: A=independent variable, Y=dependent variable, W_1, W_2=exogeneous variable, W_3, W_4, W_5=endogeneous variables

rowing at A, then the causal effect of A on Y is identifiable and can be computed through formula

$$P(y|a) = \sum_{W \in F} P(y|a,w)P(w) \tag{3.1}$$

The variables are denoted by capital letters and their values by small letters. This notation for random variables and their values is used throughout the thesis report.

3.2.2. Structural equation modeling

The structural equation modeling (SEM) on its own is a popular model to use for causal inference. It is based on modeling endogenous variables through deterministic or probabilistic equations of other variables in the model. This equation modeling means that the exogenous variables are the input of the system and the model then predicts the values of the endogenous variables. The model is easy to represent through a path diagram, which is the reason SEM and path diagrams go so well together when analyzing causal research questions.

For example, SEM for the model in Figure 3.2 can be written as

$$\begin{aligned} W_3 &= h_{W_3}(W_1) \\ W_4 &= h_{W_4}(W_1, W_2) \\ W_5 &= h_{W_5}(W_2) \\ A &= h_A(W_3, W_4) \\ Y &= h_Y(A, W_4, W_5) \end{aligned}$$

This way of describing the model is flexible in the form of the causal relationship among the variables and covers the independency assumption in a way that all parents of each variable are listed as the input of that variable's equation. However, there is no way of specifying the correlation between nodes W_2 and W_3 purely through these equation, since the latent variable is not included in the model.

Furthermore, the model equations need to comply with the so-called structural property, which states that a change in the functional form of one equation does not change functional forms of any other equation. This property is crucial for making interventions, the basis of causal inference in SEM framework, described in the following.

Interventions

Causal effects are analyzed through interventions, i.e. changes in some endogenous variables despite the values predicted by the model. This is due to these variables being influenced by other variables and thus they are suitable to be influenced. Contrary to this, exogeneous variables, such as someones age, are not influenced by anything else and thus are not suitable to make interventions on. Interventions can be seen as assigning some exposure variable certain exposure level to all individuals in a population. Technically they can be interpreted as either effects of some latent causes, or a small miracle that changes certain values in the model. This is done through substituting the equation of that endogenous variable with a simple expression of the variables equals a value of choice. The earlier mentioned structural property for the modeling equations allows the interventions to be well defined in a way that this out-of-nowhere value of the exposure variable is autonomous, i.e. one does not need to know its origin in order for the rest of the model predictions to be plausible. Furthermore, the intervention makes the exposure variable an exogenous one, breaking the

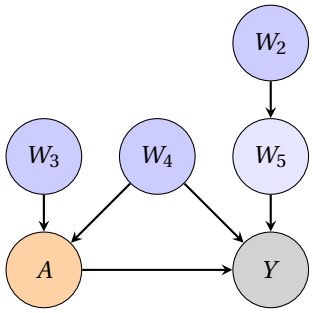


Figure 3.3: Path diagram of the model from Figure 3.2 after an intervention $W_3 = w_3$: A=independent variable, Y=dependent variable, W_2, W_3, W_4=exogeneous variable, W_5=endogeneous variable

arrows entering this exposure variable and thus changing the causal structure of the model. This procedure yields a different distribution over the modified model, used later to infer answers to causal questions.

As a short example, making interventions on variables $W_3 = w_3$ and $W_4 = w_4$ on the model from Figure 3.2 would yield a new causal model, seen in Figure 3.3, and new structural equations following new causal structure:

$$\begin{aligned} W_3 &= w_3 \\ W_4 &= w_4 \\ W_5 &= h_{W_5}(W_2) \\ A &= h_A(W_3, W_4) \\ Y &= h_Y(A, W_4, W_5) \end{aligned}$$

The notation used to describe the interventions and used for later causal analysis follows the one proposed in [21]. Concepts will be described on the example graph in figure 3.2, noted as C. Interventions are mostly used to set the exposure level of the independent variable. This is written as $set(A = a)$ or $do(A = a)$, for exposure variable A and exposure level a. This allows to inspect the distribution of an outcome variable Y, which is a descendent of A in the model C, whose new distribution under this intervention can be written as

$$P(Y|do(A = a))$$

It is not immediately clear if this probability distribution equals the conditional distribution $P(Y|A = a)$, since conditional probability may capture only one part of the population, contrary to the outcome after the intervention, which includes the whole population. Thus, the new outcome after an intervention is denoted as

$$Y_{C_a}(u)$$

where C_a represents the model C altered by the intervention $A = a$, and u representing a context of the model, i.e. exogenous variables noted as vector U. In the example of the model from Figure 3.2, $U = (W_1, W_2)$.

Identifiability

For this new outcome variable $Y_{C_i}(u)$, there is a crucial question of identification of its probability distribution. In order to make any conclusions from the observed data, the counterfactual distributions after employing interventions need to be connected to the distributions that can be identified from the data. [34] proves that the possibility of identification depends on the set of assumptions in the model, and not the model itself. [22] provides the assumption sufficient for the identification property to hold, which will not be introduced in this report, but is equivalent to the back-door criterion for finding the sufficient conditioning set. Thus, if the sufficient conditioning set exists, the sufficient assumption for identifiability holds as well. Then, [22] shows that for every analyzed exposure level a it holds that

$$\begin{aligned} P(Y|do(A = a)) &= P(Y|A = a) \text{ and} \\ E(Y|do(A = a)) &= E(Y|A = a) \end{aligned} \tag{3.2}$$

From this equivalence of counterfactual and observable distributions, the causal questions that are answerable by interventions are translated to the probability language and are therefore identifiable from the data by statistical analysis. It was crucial to have proper assumptions satisfied, in order for the statistical estimates to be unbiased for a given problem.

3.2.3. Rubin's potential outcomes model

The framework of potential outcomes is based on the intuitive interpretation of the causality as the difference in outcomes if only the parameter of interest is changed, and everything else stays the same. This means comparing and analysing the so-called potential outcomes, introduced in [28]. The first mention of the potential outcomes was by Neyman in his master thesis [18], but only in a setting of randomised experiments. Rubin generalized the notion for observational and experimental studies, and so the potential outcome model setting is often called the Rubin, or Neyman-Rubin, causal model.

The outcome variable Y is modeled through its possible outcomes Y_a and $Y_{a'}$ in cases when the dependent variable A takes some values a and a' respectively. In its original formulation, nothing much more is said about these counterfactual outcome variables, which are only assumed to exist, and the approach for estimating their marginal and joint distributions is left to the researchers themselves. However, Rubin developed theory on how to use these counterfactual outcomes to infer the causal estimands. This ambiguity is, in opinion of [23], the main reason why the potential-outcome framework is still not widely accepted and used among the statisticians and social scientists. Therefore, in order to properly set axioms of the counterfactual language, these counterfactual outcomes need to be properly defined.

One of the main assumptions of the potential outcomes framework is the consistency assumption, which states that

$$A = a \implies Y_a = Y$$

where Y is the observed outcome of the dependent variable. In other words, the assumption says that in the case that we have observed the outcome with exposure level a, this is exactly the outcome that would realize "had A been set to a". Consistency assumptions concerns only the observed outcomes, thus making this a weaker assumption than identifiability, which concerns all possible exposure levels. Moreover, to improve the readability of the report, further axiomatization can be found in Appendix A, where other underlying assumptions and definitions of this framework are described and connected to the structural theory.

The fundamental problem of causal inference

The main setback of the potential outcome analysis is that at least half of the outcomes are missing, since only one of the potential outcomes can be observed. This is known as the fundamental problem of causal inference. However, this does not stop the setting from being useful to answer causal questions. This setback is controlled through the treatment assignment. In the case of the randomized treatment assignment and a large enough sample size, the difference in the means of the treatment group outcomes is an unbiased estimate of the population causal effect. In the case of non-randomized treatment assignment, various methods have been proposed to be used in order to correct the bias stemming from the treatment assignment. The most popular methods to correct this treatment assignment bias are described in detail in section 3.5 and Appendix B.

The following subsection introduces the main estimands of interest, i.e. values that can be learnt from the causal framework proposed so far and are as close to the intuitive answers to our causal questions as possible.

3.3. Estimand of interest

Let us define the central value of interest, the expected intervention outcome (EIO). It is derived from a common causal inference treatment effect estimand, namely the average treatment effect (ATE). It should be emphasized that the term EIO estimand was not found in other literature but is here defined for the convenience in this research study.

First ATE is defined. Let Y be an outcome random variable, e.g. 1 if the patient experienced final event of interest during the study and 0 if he/she has not. ATE captures the expected difference in a patient's outcome if he/she was exposed to some event, e.g. treatment, in contrast to the case of non-exposure. This captures the causal effect of the treatment to the outcome, since everything else, i.e. patient's covariates, are taken the same, and only the exposure variable is changed. Let A be the random variable of exposure, binomial or multinomial. Next, Y_a notes the outcome of a treatment strategy when $A = a$ and Y_{a^*}, its counterfactual outcome, when the treatment was not observed. Then, ATE is defined as

$$\psi_a = \mathbb{E}(Y_a - Y_a^*) \tag{3.3}$$

When dealing with multiple strategies which need to be compared, and especially when there is no one obvious way to define a reference strategy for the strategies researched, one may compute only part of ATE: the expectated outcome after an intervention has been done, namey EIO. This is the case in this study, and therefore, for each strategy $A = a$ we define the EIO estimand of interest as

$$\xi_a = \mathbb{E}(Y_a) \tag{3.4}$$

EIO is a more general and easier to use estimand than ATE, and can be easily linked to ATE. For each strategy a and its counterfactual strategy a^*, EIO estimand is easily connected to ATE through the following expression

$$\psi_a = \xi_a - \xi_{a^*} \tag{3.5}$$

3.4. Assumptions and biases

Following the subsection on the assumptions and biases in the survival analysis modeling, the same should be discussed in the causal inference framework. Certain assumptions and biases, for example the consistency assumption in the Rubin's potential outcomes framework or confounding bias in general, have already been introduced in Section 3.2. Others common in causal inference are listed and commented upon below. A brief comment about how to resolve the bias if the assumption does not hold is mentioned. Each assumption is commented if it holds in this thesis and testing these assumptions will be done in Section 5.2.

3.4.1. No unmeasured confounding assumption

The no unmeasured confounding assumption says that a sufficient subset of all variables involved in confounding of the exposure and the outcome variable are included into the analysis. In most research problems this is not testable. However, it is crucial the hypothesis is satisfied when analysing a causal effects of variables, since if the assumption does not hold, it is impossible to differentiate between the confounding variable's influence and the real causal effect of the independent variable of interest.

3.4.2. Stable unit treatment value assumption

The stable unit treatment value assumption (SUTVA) in causal inference concerns the consistency of the treatment exposure across all the patients in the study. The assumption states that the exposure of one patient does not effect outcome of any other patient in the study, and the level of the exposure is the same for each individual across that exposure level.

In case the assumption does not hold, the analysis is made more difficult, and one way of correcting the treatment definition is to define the treatment as a combination of treatments for all individuals influenced by a certain patient.

3.4.3. Positivity assumption

Positivity assumption makes sure that conclusions from a study on certain dataset are not extrapolated to the parts of population not included in the dataset. This is especially important when counterfactual predictions are being made and thus it is a foundation for any causal inference method. The assumption is formulated as that every patient in a study has a positive probability of receiving any treatment that is analyzed. If X marks the covariates of a patient, x possible values of X, A is the exposure variable, and if values it can take are labeled with a, then the positivity assumption can be written as

$$P(A = a|X = x) > 0, \ \forall a, x$$

It is interesting that this assumption can be also framed as the 'no causation without manipulation' statement. This means that if a participant in the study cannot obtain both treatment and no-treatment states, then his personal causal effect does not exist, since he cannot obtain the counterfactual exposure. The assumption is testable and should always be checked by looking into the data. If it does not hold, either the scope of the problems needs to be decreased, the treatment needs to be differently defined, or the dataset cannot be used to answer the causal question of interest.

3.5. G-computation

Rubin's potential outcomes model have been introduced earlier in section 3.2.3. The model can be used to estimate causal estimand EIO, introduced in Section 3.3. The most popular methods to do this are G-

computation, inverse probability of treatment weighting (IPTW), and their combination called targeted maximum likelihood method (TMLE). A study using IPTW has already been done on this same research problem in [13]. Therefore, in this study the G-computation method is used. The other two methods, IPTW and TMLE, can be read about in Appendix B and are presented to allow comparison of the available methods.

G-computation was introduced by Robins in [25] as a generalised method for the treatment effect estimation. Originally, the method estimates average treatment effect (ATE), but here we use it to estimate expected intervention outcome (EIO), both introduced in Section 3.3. The estimation is done through the G-formula. Namely, for Y being an outcome variable, A an exposure variable and W the vector of confounding variables, the G-formula is written as the following

$$P(Y) = \sum_{w} \sum_{a^*} P(Y|A = a^*, W) P(A = a^*|W) P(W) \tag{3.6}$$

Then, G-formula is conditioned on the exposure levels of interest. Namely, G-formula conditioned on an intervention $A = a$ is written as

$$P(Y|A = a) = \sum_{w} P(Y|A = a, W) P(W)$$

Then, using the notation of the counterfactual outcomes Y_a after intervention $A = a$, this implies the estimation of the EIO through

$$\mathbb{E}(Y_a) = \mathbb{E}_W[\mathbb{E}(Y_a|W)]$$

The algorithm to conduct the G-computation is described as follows:

1. Fit a model for the outcome Y with an exposure variable A and the confounding covariates W on the dataset available for the study. The way this is done in this thesis will be explained in Section 5.3.
2. Predict the potential outcome $\widehat{Y}$ at exposure level a for each patient in the dataset using the model from the previous step.
3. The EIO estimate is then calculated as $\hat{\psi}_a = \frac{1}{N} \sum_{j=1}^{N} \widehat{Y}_a$, where N equals the size of the dataset from the first step.
4. Quantify the uncertainty of the obtained estimate from step 3 by the bootstrap procedure. This implies taking a number of datasets M, e.g. 500, of the same size N, sampled with replacement from the original dataset used for the study. Then, the model is fitted on a bootstrap sample dataset (step 1), and steps 2 and 3 are conducted on it. This is done for each bootstrap sample, which provides M estimates of $\hat{\xi}$. The 95% confidence intervals for the $\hat{\xi}$ are computed as the 2.5th and the 97.5th percentile of the set $\{\hat{\xi}_{am}\}_{m=1}^{M}$.

Next to the assumptions mentioned in this section the causal inference is relying on, the crucial assumption for G-computation to provide unbiased estimates is the no model misspecification assumption. This relates to the model used to make counterfactual predictions, in this thesis being the multi-state model with Cox proportional hazards modeling the transitions between the states. The assumption will be checked in Section 5.2.

4

Data Description

This section describes the dataset at hand, which is used to tackle the problem of evaluating the effect of the IUI treatment starting time. Firstly, the format of the dataset is commented on. Then, the quality of the dataset is assessed, through description of the criteria used for the patients to be a part of this medical study. Then, variables used for the analysis are listed and their descriptive statistics are presented. Lastly, for a better understanding of the data, basic survival analysis on pregnancy when ignoring the treatment assignment is conducted.

4.1. Data format

Survival data is usually longitudinal data, meaning that the patients are followed through time and each observed event is recorded in the dataset. This can be written in the wide or long format. Wide format contains one row for each patient, and has multiple columns containing all the covariates, possible changes in the covariates and the time of change, the events of interests and the possible intermediate events, with their times of occurrence.The long format contains multiple rows for each patient, therefore reducing the number of necessary columns int he dataset. Each row follows the patient in some time interval, mostly between the events, or in time period specified in advance. Both formats can be equivalently used and are chosen based on the difficulty of the further analysis of the data.

In this study, both wide and long format are provided. Long format is more suitable for the analysis with multi-state models, so mostly that one will be used. Furthermore, since only one intermediate step is tracked, the treatment indicator, it is easy to transform one format to the other and both are equally suitable.

4.2. Patient selection

The methodology is applied and tested on a dataset consisting of couples with unexplained subfertility, some of which received the Intrauterine Insemination (IUI) treatment and some stayed on expectant management until their pregnancy or the end of the followup. The couples were a part of a large study on fertility carried out in The Netherlands, in the period between January 2000 and October 2005. 7 out of 38 hospitals included in the study collected data on the IUI treatment.

A couple is said to be subfertile if their duration of subfertility lasted for at least 12 months and all the other parameters indicating infertility were inside the normal ranges: the menstrual cycle was between 23 and 35 days, the total motile sperm count at least 10^6 and the couple was not diagnosed with a blockage of both fallopian tubes. The last criterium means that they did not take tests like hysterosalpingogram (HSG) or that the results of such tests were that one or none fallopian tube is blocked. These values were determined on a fertility workup, which every couple had to undergo before entering the study and after a period of trying to conceive naturally. More on the workup can be read about in [3].

After completion of the workup, a couple decides, with advice of their doctor, which treatment strategy to follow. A treatment strategy is based on the initial time that the couple stays on the expectant management before starting with the IUI. Strategies are explained in more detail in Section 5.3.

Table 4.1: Variables in the dataset on patients with unexplained subfertility who recieved IUI or stayed on expectant management before getting pregnant or being cencored

Type of Variables	Variable Name	Description
descriptive	sleuteli	Couple's ID
	t0_newadj	Date of finishing the workup
	datecons	Date of registration at a fertility center
	lok	Location where the couple received their workup
time-fixed covariate	vrouleef_newadj	Female age at workup completion (in years)
	duursubf_newadj	Subfertility duration at workup completion (in years)
	infvrouw1	Subfertility occurance (1=primary, 2=secundary)
	verwijz3	Referral from gynecologist (1=yes, 0=no)
	sa1prog	Percentage of progressive motile sperm
	savcmgem	Total motile sperm count
	tubapat1_t0	VTD for a blocked fallopian tube (yes/no/no vtd)
time-dependent covariate	treatedIUI	Started IUI (0 until the first treatment, 1 thereafter)
	oiiui	Ovarian simulation with IUI (1=yes, 0=no)
outcome	t_fusp_t_newadj	Time to the first event/censoring (in years)
	Z	Indicator of an event (1=pregnancy, 0=otherwise)
	start	Time of entering the state (in years)
	stop	Time of leaving the state (in years)

4.3. Variables

The dataset describes 1896 couples through 17 variables, if viewed in wide format, or 20 if viewed in the long format. They are listed in Table 4.1.

First, there are 4 descriptive variables: couples' identification numbers, dates of the start of the workup and the location of where they received their workup. These variables are assumed to be independent of the outcome variables and therefore are not included in the models. This assumption allows for the conclusions of the study to be drawn for the whole population of subfertile patients and is the base of the sruvival analysis.

Next, 7 time-fixed covariates are the parameters that can influence the outcome and among which, the most influential ones will be used for the further analysis. Subfertility duration is the time the couple actively tried but failed to naturally conceive. The subfertility occurance indicates if a woman has succeeded to conceive before the clinical trial, with the primary subfertility meaning that she did not have any pregnancies before and the secundary subfertility meaning that she had a lasting pregnancy before. VTD is abbreviation for a Visual Tubal Diagnostics test. However, this test was not mandatory for all couples, and a lot of them did different fallopian tube tests that are not recorded in this dataset. For this reason, this parameter is excluded from this study.

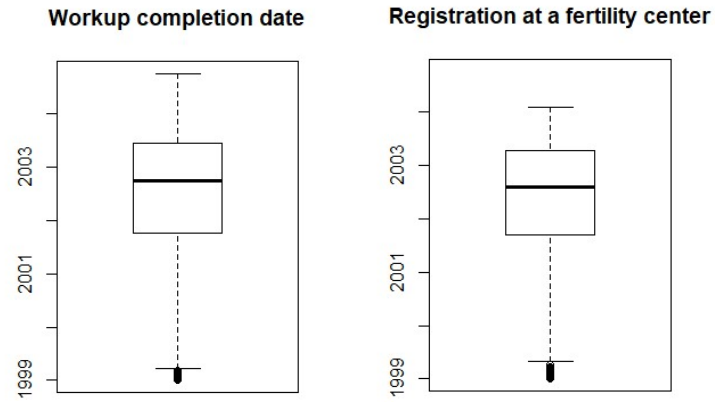
Moreover, 2 time-dependent covariates consist of indicator of the IUI treatment and if ovarian simulation (OS) was given with it or not. They are equal to 0 until the first treatment, when they change to 1 or stay 0.

Lastly, there are 4 outcome variables in the long format and 7 in the wide format. Time to the first event, which can be read from the outcome variables. Variable Z indicates if a couple got pregnant at the time of that event. The same holds for Z1 and Z2 in the wide format. Variables start and stop contain the time passed from the end of the workup to the event that happened. $start$ represent entering a state and $stop$ leaving a state. At the start of the followups, all couple have start=0. In the wide format, if treatment is the first event, then $start2$=$stop1$. Z2 and $stop2$ can be NA in case there was no treatment, so the first event is an absorbing event.

4.3.1. Descriptive analysis

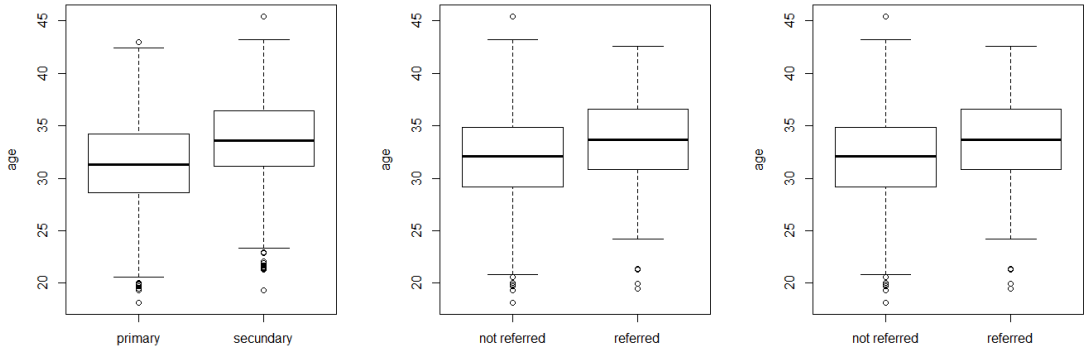
In this subsection, variables are analysed and the results are summarised and graphically presented, to get a better insight into the data.

The descriptive variables are analysed first. Time distribution of received the workup and registration in a fertility center are presented in the Figure 4.1 on the left and in the center respectively. Couples mostly



Location of the fertility center	Number of couples on the location
AMC, Amsterdam	683
MMC, Veldhoven	293
TSZ, Tilburg	300
SMG, Venlo	254
ZGN, Gooi Noord	118
HZMC, Zaandam	125
SZE, Emmen	123

Figure 4.1: Summary of descriptive voariables: boxplots of time distribution of couples (left) finishing their workup and (center) registering at a fertility center and (right) a table of fertility center locations and the number of couples registered there



(a) Degree of subfertility against female age

(b) Referral against female age

(c) VDT test against female age

Figure 4.2: Boxplots illustrating relationship of the categorical variables versus female age

finished their workup in 2002, even though some workup data can be a couple of years old. Couples mostly registered just before doing their workup, which is reasonable to expect. Locations where the couples received their workup and the distribution of couples in these 7 center and be seen in Figure 4.1 on the right.

Next, 7 time-fixed covariates are considered. In Table 4.2 on the left, mean and standard deviation of the 4 numerical covariates is listed. Table 4.2 on the right shows their correlations. Of all correlations, the progressive sperm and the total sperm count stand out with their high corerlation of 0.48. Furthermore, total sperm count is much more volatile than the progressive sperm, so it is reasonable to chose only progressive sperm from the two for the further analysis. Other correlations are 0.1 or smaller. On the other hand, the factor time-fixed covariates are summarised in Table 4.3. Since all 3 covariates have different responses, they are labeled as 1, 2 and 3, and described inside the table in the brackets. 10% of the couples were referred from a gynecologist, 35% had had a pregnancy before entering the study and 14% of the couples tested by VDT were diagnosed with a blockage of a fallopian tube. Moreover, they are plotted against a numeric variable, for example female age, to see their relationship to the numeric covariates. The boxplots can be seen in Figure 4.2. The boxplots for the other numeric covariates look similar as well. However, due to their low representation of the different outcomes, referral and VDT diagnosis will be excluded from the further analysis.

Furthermore, the two time-dependent covariates, indicating if the IUI treatment has been started and if OS has been given next to it are analysed. Out of 1896 couples, 863 (45%) has started with the treatment, and

Table 4.2: Mean and standard deviation of the numeric time-fixed covariates in the dataset

Covariate	Mean	Standard deviation
Female age	32.12	4.35
Subfertility duration	2.04	1.41
Progressive sperm	33.34	18.98
Total sperm count	67.30	98.10

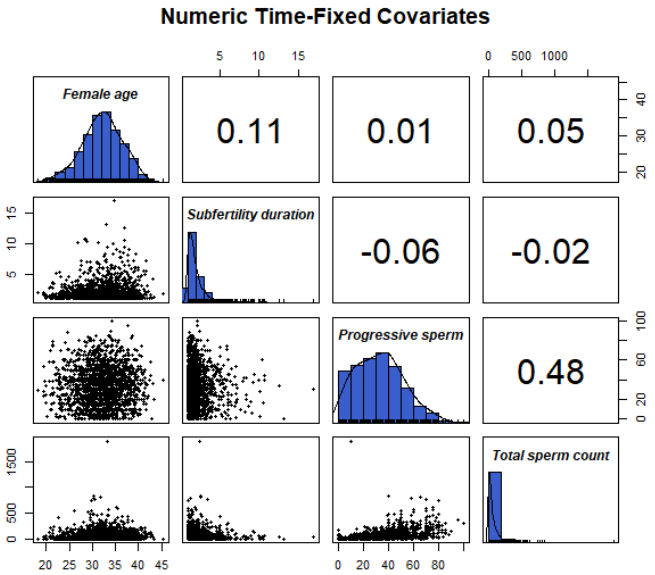


Table 4.3: Output summary for factor time-fixed covariates in the dataset

Referral from gynecologist	Subfertility occurance	VTD test
192 (referred)	1217 (primary)	127 (one-sided blockage)
1704 (not referred)	679 (secundary)	739 (no blockage)
		1030 (no test)

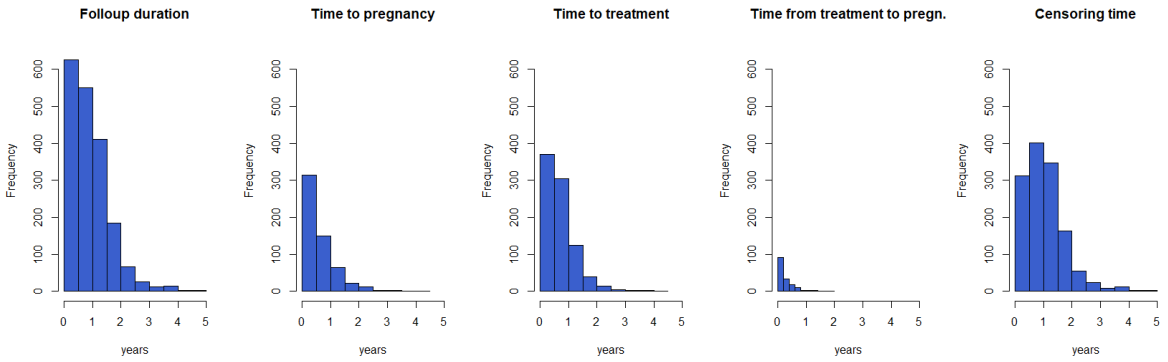


Figure 4.3: Frequencies of the fertility study outcomes: follow-up time, time to pregnancy, time to treatment, time from treatment to pregnancy and censoring time respectively from left to right

712 (82% of the treated couples) have received OS during the treatment. For this project, OS is not in the focus of interest, and due to the lack of representation of the treatment without OS, this variable is excluded from the further analysis.

Lastly, outcome variables are summarised and presented. Variables Z, $start$ and $stop$ are a convenient way of writing down information like when the couple received the treatment, when did the get pregnant and how long were they in the followup. The latter can be computed for all couples, and distributions of the times to treatment, pregnancy, censoring or time between treatment and pregnancy can be approximated. The frequencies of the mentioned random variables over time are presented in Figure 4.3. Out of 1896 couples, 30% got pregnant in total and 21% got pregnant naturally. 45% of the couples received the treatment, out of which 18% got pregnant. Follow-up time goes up to 5 years, but most couples left the trial within a year. There is a sharp decrease in couples present in the study after 1.5 years into the follow-up. For this reason, 1.5 years has been chosen to provide a time frame of this analysis. Similar is true for receiving the treatment, while most pregnancies occurred within first 6 months of the followup. Furthermore, peak of censoring time is between 6 months and 18 months. Lastly, most of the couples who received the treatment got pregnant within 6 months and only a few between first and second year after the first treatment. This can be attributed to the way the treatment is provided. Most couples receive the IUI in cycles of 6 or 9 months. In case of not getting pregnant in this period, they often seek a different treatment and therefore get censored in this dataset.

5

Application

This section connects the research problem of the effect IUI treatment timing on pregnancy for couples with unexplained subfertility, described in Section 1, with the theory on survival analysis and causal inference from Sections 2 and 3. The model used for the IUI treatment timing is a multi-state model, put in a causal setting in order to answer a causal question of IUI treatment timing effect on the pregnancy, is introduced in Section 5.1. The model is introduced from both the survival view and the causal setting, and the combination of the two is thoroughly described. Next, assumptions and biases encountered in survival analysis and causal inference presented in Sections 2.3 and 3.4 are discussed in the context of this research problem in Section 5.2. Lastly, steps of the G-computation method and the way the treatment effect can be compared to the expected number of treated couples are described in Section 5.3.

5.1. Model

To start the analysis of the IUI treatment timing effect, one needs to model the problem first. Therefore, this subsection described the design of the multi-state Cox model used for causal predictions of the pregnancy. First the multi-state model is connected to the causal inference setting. Then, their graphical connection is introduced, and interventions made on a multi-state model are commented on. Lastly, the implementation of the multi-state model in R is described.

Multi-state model

First, there are 6 time-fixed covariates that describe each couple. These are female age, subfertility duration, type of infertility, doctor's referral, test for ovarian tubes blockage and progressive sperm count. In the model they are represented as variables $X_1, ..., X_6$. They represent the information available about each couple and will be used to make pregnancy probability prediction for each individual. A slight correlation among the variables exists, but it is not crucial to explicitly model the structure of the common causes of the variables in order to conduct the causal analysis and therefore this is not discussed in this section. Correlations and other descriptive analysis of the covariates are computed and described in Section 4.

Then, the time from the work-up completion to the IUI and the pregnancy is represented by positive random variables. It will also be useful to introduce a time variable representing the time from starting the IUI until the pregnancy. Notation for these variables comes from the transitions in a multi-state model that they represent. Namely the time of starting the IUI treatment relative to the work-up completion is denoted as T_{12}, the time of natural pregnancy relative to the work-up completion as T_{13} and the time of pregnancy after the treatment has been started and relative to the start of the treatment as T_{23}. Therefore, the observed time of pregnancy equals

$$T = \begin{cases} T_{12} + T_{23}, & \text{if } T_{12} < T_{13} \\ T_{13}, & \text{if } T_{12} > T_{13} \end{cases}$$

Furthermore, indicator functions are used to describe in which state in a multi-state model a couple is. The multi-state models are defined in Section 2.5. In this model, the states of interest are State 1: expectant management (EM), State 2: receiving the IUI treatment and State 3: being pregnant. The IUI treatment is

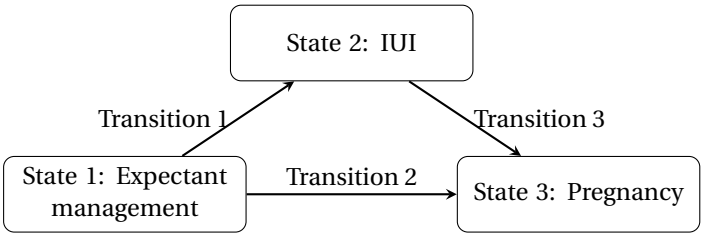


Figure 5.1: 3-state multi-state model applied on the research problem

represented by a function $A(t)$, indicating the status over time of the IUI treatment received by a certain couple and defined on the time interval $[0, T]$. $A(t)$ is defined as

$$A(t) = \begin{cases} 0, & t < T_{12} \\ 1, & t \geq T_{12} \end{cases}$$

Similarly, pregnancy is represented by a function $Y(t)$ indicating the status over time of pregnancy for a certain couple, also defined on the time interval $[0, T]$. $Y(t)$ is defined as

$$Y(t) = \begin{cases} 0, & t < T \\ 1, & t \geq T \end{cases}$$

The states of the multi-state model are based on the values of the indicator functions $A(t)$ and $Y(t)$. Namely, state 1 is defined by having both $A(t) = 0$ and $Y(t) = 0$, which holds in the time interval $t \in [0, \min\{T_{12}, T\})$. Then, if $T_{12} < T$, a couple has started the treatment before getting pregnant naturally and has transitioned to state 2. This state is defined as the time when $A(t) = 1$ and $Y(t) = 0$. Then, at time $t = T$ this couple transitions to state 3, defined as time when $Y(t) = 1$, i.e. during the time $t \in [T, \infty)$. Lastly, a transition directly from state 1 into state 3 is possible, in case of $T < T_{12}$. The described multi-state model can be visualized as in Figure 5.1.

The transitions in the multi-state model are modeled by Cox proportional hazards models, which were defined in Section 2.4. Each transition models the hazard of the event of interest happening. Definition of hazard functions can be seen in Section 2.2. Specifically, Transition 1 (EM → IUI) models the hazard of IUI treatment being started at time t, namely the case of $T_{12} = t$, for some time t. The same goes for Transitions 2 (EM → pregnancy) and 3 (IUI → pregnancy) for modeling the hazard of $T_{23} = s$ or $T_{13} = t$ respectively. Moreover, the clock-reset time scale is used in this model, due to the assumption that after the treatment has been started, the start of the followup time becomes less important for the pregnancy occurrence than the time that has passed from the first treatment cycle.

Let $\boldsymbol{X}^6 = (X_1, ..., X_6)$ be a vector of covariates and $\boldsymbol{\beta}_{ij} = (\beta^1_{ij}, ..., \beta^6_{ij})$ a vector of covariates' coefficients, with $\boldsymbol{\beta}'_{ij}$ representing the transpose of vector $\boldsymbol{\beta}_{ij}$, for $i, j = 1, 2, 3$, where i is the state the transitions starts at and j the state the transitions ends at $(i < j)$. Moreover, in order to emphasize from which time point the time is being measured, t is used for time from the work-up completion, and s is used for time from the start of the IUI treatment. Then, the transition hazards are modeled as

$$\begin{aligned} \text{Transition 1 (EM} \rightarrow \text{IUI)} &: \lambda_{12}(t|\boldsymbol{X}^6) = \lambda_{12_0}(t)\exp(\boldsymbol{\beta}'_{12}\boldsymbol{X}^6), \quad \text{for } t \geq 0 \\ \text{Transition 2 (EM} \rightarrow \text{preg)} &: \lambda_{13}(t|\boldsymbol{X}^6) = \lambda_{13_0}(t)\exp(\boldsymbol{\beta}'_{13}\boldsymbol{X}^6), \quad \text{for } t \geq 0 \\ \text{Transition 3 (IUI} \rightarrow \text{preg)} &: \lambda_{23}(s|\boldsymbol{X}^6) = \lambda_{23_0}(s)\exp(\boldsymbol{\beta}'_{23}\boldsymbol{X}^6 + \beta^7_{23}T_{12}), \quad \text{for } s \geq 0 \end{aligned} \tag{5.1}$$

Causal aspect

The multi-state model is now connected to the notation in the structural theory, which is described in Section 3.2. Firstly, a model needs to be represented by a causal model. This means having a graphical representation through a path diagram and set of structural equations modeling all endogeneous variables, i.e. variables that are influenced and thus modeled by other variables in the model. Each transition in a multi-state model can be represented by a causal model. In each transition's causal model, covariates on which the hazard of that transition are conditioned on are exogeneous variables, and the time until the transition is the endogeneous variable, namely $T_{12}, T_{13}, T23$. The expressions in equation (5.1) represent the structural equations for the three causal models.

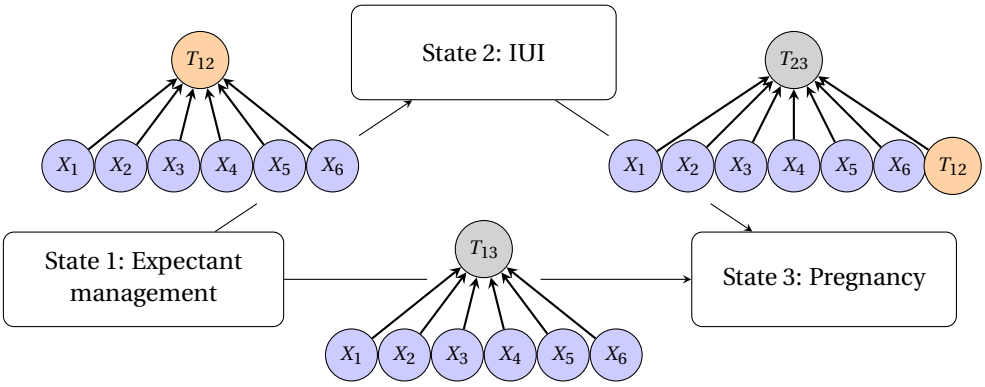


Figure 5.2: Multi-state causal graph for modeling the IUI problem: three transitions between three possible states a couple can be in, modeled through six covariates in blue (X_1 to X_6 stand for female age, subfertility duration, doctor referral, infertility type, tubal blockage and progressive sperm count), time-to-treatment variable T_{12} in orange and time-to-pregnancy variables T_{13}, T_{23} in grey, colors chosen for easier differentiating between the important variables

Then, one can combine the multi-state graph with the causal graphs modeling the transitions in one large graph. In this thesis, this looks as graph in Figure 5.2. This visual representation of multi-state model in causal setting provides convenient setting to analyze making interventions in the model. Therefore, to improve readability, the graph shown in Figure 5.2 will be referred to as a multi-state causal graph. It should be emphasized that this notation was not found in other research studies.

The most important detail of combining the multi-state models with the causal inference setting is the way the interventions are made on a multi-state model. Interventions can be made on endogeneous variables, as described in Section 3.2. In the multi-state causal graph in Figure 5.2, it can be seen that the endogeneous variables are T_{12}, T_{13} and T_{23}. Therefore, the possible interventions in this model are interventions on the times of events of interest happening. In this thesis, it is of interest to intervene on the time until the IUI has been started. From the medical point of view, among the three possible transitions, this is also the only transition that is reasonable to condition on.

So let t_A be a level of exposure determined by the chosen intervention, later known as treatment strategy, for the time of the IUI treatment variable T_{12}. Subscript A comes from common notation in causal inference of A being the exposure variable, notation which is also used in Sections 3 and 5.3. Then, based on the structural theory described in Section 3.2, the structural equation modeling this variable becomes deterministic at the specified time, in this case noted as $T_{12} = t_A$. This transition will occur only if a couple has not become pregnant before T_{12}, i.e. if the couple is in the risk set of Transition 1 just before time t_A. Therefore, the hazard for the Transition 1 (EM → IUI) is no longer applicable way to model T_{12}, and so new set of structural equations are

$$\begin{aligned}
&\text{Transition 1 (EM} \to \text{IUI)} : \; T_{12} = t_A \\
&\text{Transition 2 (EM} \to \text{preg)} : \; \lambda_{13}(t|\boldsymbol{X}^6) = \lambda_{13_0}(t)\exp(\boldsymbol{\beta}^{2\prime}\boldsymbol{X}^6), \quad \text{for } t \geq 0 \\
&\text{Transition 3 (IUI} \to \text{preg)} : \; \lambda_{23}(s|\boldsymbol{X}^6) = \lambda_{23_0}(s)\exp(\boldsymbol{\beta}^{3\prime}\boldsymbol{X}^6 + \beta_7^3 T_{12}), \quad \text{for } s \geq 0
\end{aligned} \tag{5.2}$$

Section 3.2 described that the arrows going into the variables intervened on are removed from the causal graph. This is true for the causal graph representing the Transition 1 in the multi-state model. There, couple's covariates no longer influence the time of starting the treatment. However, in a multi-state causal graph the transition from EM to IUI is still feasible, and happens at time t_A after the work-up completion, is the couple is still on the EM just before t_A. In vocabulary of multi-state models, the transition happens if the couple is in the risk set for Transition 1 just before time t_A. Therefore, this transition is now represented differently in the multi-state causal graph, for example by a dashed arrow and the fixed time of the transition written next to the arrow. This can be seen in Figure 5.3.

Implementation in R

Lastly, we describe the procedure of fitting the multi-state model on the data, which results in transition hazards, which can be used further to predict the causal estimands, as will be later described in Section 5.3.

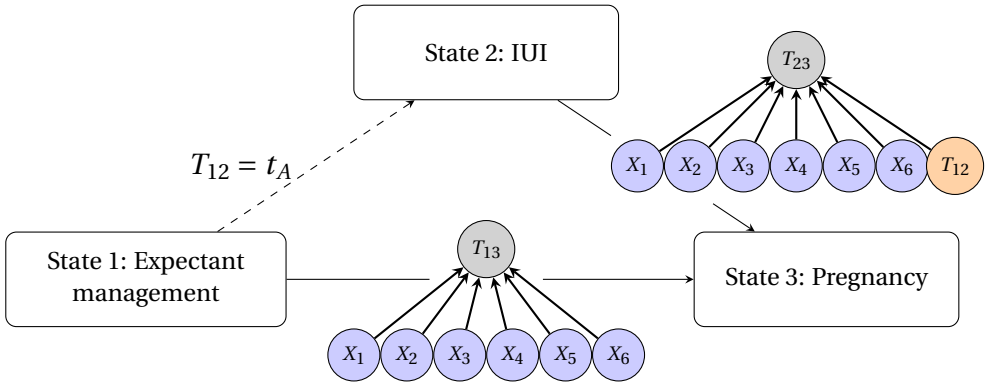


Figure 5.3: Multi-state causal graph for modeling the IUI problem after an intervention has been done on the time to IUI treatment T_{12}: a deterministic transition between expectant management (EM) and IUI at time t_A, and two stochastic transitions from EM and IUI into the pregnancy; transitions modeled through six covariates in blue (X_1 to X_6 stand for female age, subfertility duration, doctor referral, infertility type, tubal blockage and progressive sperm count), time-to-treatment variable T_{12} in orange and time-to-pregnancy variables T_{13}, T_{23} in grey, colors chosen for easier differentiating between the variables

First, data needs to be prepared for the already implemented function estimating hazards to be applied on. The data preparation follows [24] and the *expand.covs()* function transforms data from wide to long format, described in Section 2.4.1. A thorough description of the required dataset form is described in Appendix 4.1.

Then, the model is fitted on the data through the function *coxph()*, implementing the estimation procedure described in Section 2.4.1. This yields estimates of the baseline hazard and the covariate coefficients for each transition. The estimated baseline hazards and covariate coefficients are then used to predict hazards for any couple from the population of the couples with the unexplained subfertility. Predicting the hazards for an arbitrary couple can be done with the *msfit()* function from the *mstate* package, or computing them directly from the Cox hazard equations (2.1).

Later in Section 5.3 the estimation of the transition probabilities from the predicted hazards is described. The whole code of fitting the model on the data can be found in Appendix D. In the following subsection, modeling assumption and possible biases are commented on in the context of the model described.

5.2. Addressing assumptions and biases

In this subsection, the modeling assumptions and potential biases from survival and causal analysis specific for this data application are commented on. The most important sources of bias are the selection and attrition bias introduced in Section 2.3, the latter one stemming from the informative censoring, and confounding bias introduced in Section 3.4. The assumptions being made are no unmeasured confounding, stable unit treatment value, positivity, consistency and correct model specification assumption, introduced in Section 3.4. The assumptions and biases listed are discussed in more detail in Section 5.3 in the context of this research.

The assumptions and potential biases in this research problem are the following:

- To start with, the selection bias is present in almost any observational study, this one included, and comes from participants being volunteers for the study. This study was an all-comers study, meaning that all couples seeking medical help with getting pregnant were asked to join the study. In the dataset it is lowered even more by conditioning the outcomes on couples' covariates and therefore having a better overview of the type of couple participating in the study, as well as the parts of population for which the study results are generalizable.

- Next, the attrition bias in this context comes from the couples with lower probability of natural pregnancy often starting with the IVF treatment and thus being censored from the study. This leads to informative censoring. However, their decision to start with IVF is largely influenced by the covariates in the dataset, and thus this bias is also lowered by the conditioning the predictions on the covariates. Finally, [38] showed that using inverse probability weighting to correct for informative censoring in context of

fertility treatments does not influence the survival curves significantly.

- Moreover, the indication bias, a kind of confounding bias, arises when couples with lower probability of natural pregnancy often start with the IUI treatment earlier than the others. This bias is lowered by conditioning on the couples' covariates, which provides more accurate predictions for couples' probability of getting pregnant. This is a crucial bias to combat in this research study and thus the reason why conditioning on variables decreases the bias is described in more detail in Section 5.3.
  In order for conditioning on covariates to be enough to eliminate the indication bias, the no unmeasured confounding assumption is crucial to be satisfied. It is not directly testable, and it is considered to hold in this research problem. The 6 time-fixed covariates in this dataset are the most important factors influencing the probability of getting pregnant, used in studies such as [16]. Lastly, variables that might be subjective characteristics of a couple, e.g. their impatience to get pregnant and therefore starting the treatment early or their personal belief in a more invasive treatment than their doctor may be advising them, are not likely to influence the probability of pregnancy, with or without the treatment, and are therefore not confounding variables.

- Furthermore, the stable unit treatment value assumption (SUTVA) should be mentioned. The treatment is given equally to all couples receiving it. 90% of the couples who received the IUI treatment also received the ovarian simulation, which is introduced in Section 1. Moreover, the amount of hormones is accustomed to each woman for the maximal effect. Therefore, couples are assumed to be treated equally. On the other hand, treatment of one couple has no influence on outcome of any other couple.

- In addition, the positivity assumption needs to be tested. This is done empirically by looking into the data and seeing if certain types of couples always received the treatment, or, on the other hand, never received the treatment around 0, 3, 6 or 9 months into the followup. Due to difficulties of inspection of space coverage in more than 2 dimensional spaces, the space coverage is analysed through pair-wise space coverage of covariates, and due to their low mutual correlations, the results can be generalized to conclude if the total covariate space is covered well. Moreover, since treatment time is continuous, a time interval of 1-month length around the specified times is created and scatterplots on all pairs of covariates are plotted. The plots are compared to the pair-wise covariate scatterplots for all couples in the dataset, which was provided in Table 4.2. The assumption holds, and more on this testing procedure can be read in Appendix C.2.

- Then, structural property of the structural equations in the causal models needs to hold. This property was introduced in Section 3.2 and refers to the independence of each structural equation of the form the other equations are written in. In this study this is satisfied, since modeling one transition in the multi-state model does not influence the way other transitions are modeled. For example, changing the way female age influences the probability of starting the IUI treatment does not influence the way any covariate influences the pregnancy probability before or after the IUI has been started.

- The consistency assumption is not testable but is assumed to hold in this research study. It should be noted that the observed outcome in the dataset is an indicator function of pregnancy, and our analysis of the treatment effect is based on hazard functions and therefore it is not directly comparable to the observed outcomes. However, the next assumption, the correct model specification assumption, ensures that the consistency assumption can be safely assumed.

- Lastly, the assumption of correct model specification is important to test. The assumption was introduced in Section 3.5 as the main assumption G-computation relies on. The model specification relies on the fit of the model with the research problem and data at hand. The main assumptions made for the multi-state model using Cox proportional hazards model for modeling transitions between the states are do covariates act linearly in log hazards and the proportionality of hazards. Lastly, for the robustness of the model, presence of influential outliers is checked.
  First, the linearity of the log hazards and the covariates is assessed through the martingale residuals. The analysis provided in Appendix C.2 implies that all but subfertility duration in Transition 1 satisfy the linearity assumption. The logarithm of the subfertility duration variable is then taken and this transformed variable satisfies the linearity assumption, so this transformation is implemented in the model for Transition 1.
  Then, the most common test of hazards proportionality is the inspection of the Schoenfeld residuals through the Chi-square test, implemented in R function *cox.zph()*. The analysis of this problem is

provided in Appendix C.2. Results imply that most of the covariates support the proportionality assumption. However, female age, doctor referral and 1-sided tubal blockage do not comply with this assumption in the transition 1 model. This is resolved by separating the female age variable into two variables conditioned on the age being below or above some theshold, in this case 32 years. Moreover, two baseline hazard functions are allowed for both no and 1-sided tubal blockage. With this implemented, the hazards proportionality assumptions holds in all transition models.
Finally, the existence of influential outliers is analyzed. This is done through examination of deviance and so-called dfbeta residuals, described in more detail in Appendix C.2. The results of this analysis show that there are no influential outliers in this study.

5.3. Causal estimation

In order to answer the causal research question of quantifying the effect of treatment timing on the pregnancy rates for couples with unexplained subfertility, we need to make sure the causal estimands can be identified from the modeling setting in an unbiased way. For this reason, formal steps to do this, introduced in [23], are followed. The steps are numbered and immediately commented upon in the text.

1. "Define: Express the target quantity Q as a function $Q(C)$ that can be computed from any model C."
The target estimand is the expected intervention outcome (EIO). In order to define it in the context of this research problem, the treatment strategies need to be defined. The treatment in this dataset study is given continuously over time. However, due to the dataset being relatively small, a few treatment strategies are defined in order to draw conclusions on the IUI treatment timing effect more easily. A treatment strategy is defined as being on the expectant management until time T_{12}, a time when a couple starts their first IUI cycle, in case the couple has not become pregnant before time T_{12}. After receiving the treatment, they are still trying and waiting to get pregnant. We are interested in 5 separate timings for T_{12}, which are 0, 3, 6 and 9 months after the workup completion, or not giving the treatment at least before the end of the analysis being at 1.5 years from the beginning of the followup, which is from no on referred as no treatment at all.

One way of writing down the strategies is through T_{12}, by setting the random variable to a specific value, namely

$$T_{12} = t_A, \text{ for } t_A \in \{0, 0.25, 0.5, 0.75, \infty\}$$

Moreover, since treatment strategies influence the outcome function $Y(t)$ through the T_{12}, which then equals t_A, outcome function when following certain strategy is a counterfactual outcome function, written as $Y_{t_A}(t)$.

Treatment strategies are used to define EIO, the target quantity to estimate. EIO estimator is introduced in equation (3.4) for a general outcome variable Y, and in this context it is written as $\xi_{t_A}(t) = \mathbb{E}(Y_{t_a}(t))$, for some $t \geq 0$ and strategy $T_{12} = t_A$. Due to computational expensiveness, $\xi_a(t)$ is computed at 1.5 years after the start of the followup, therefore providing the final estimand of this study

$$\xi_{t_A}(1.5)$$

for $t_A \in \{0, 0.25, 0.5, 0.75, \infty\}$. Lastly, for more convenient notation in the analysis and generalizability of the computation, the final time is noted as $t^* = 1.5$.

2. "Assume: Formulate causal assumptions using ordinary scientific language and represent their structural part in graphical form."
The graphical form of the model is represented by the path diagram, seen in Figure 5.2, and the structural equations, seen in equations (5.1). both described in Section 5.1. Moreover, the causal assumptions important for this step of conducting causal inference, introduced in Section 3, are assumptions regarding the absence of certain causal relationships. Namely, in this model the assumptions are that the time of the pregnancy does not influence the time of the IUI, and that pregnancy and IUI time do not influence any of the covariates.

3. "Identify: Determine if the target quantity is identifiable."
Based on a sufficient condition for identifiability, described in Section 3.2.2, the target quantity $\xi_a(1.5)$ can be identified from this model since there exists a set of covariates that satisfy the back-door criterion. Since the no unmeasured confounding assumption is being made, from the path diagram in Figure 5.2 is it easy to see that all 6 time-fixed covariates need to be included in the conditioning set. This is due to each covariate

having a direct influence on both IUI and pregnancy variables. Therefore, the path IUI $\leftarrow X \rightarrow$ pregnancy is conditioned on only if the covariate X is included in the conditioning set. This holds for every covariate X in all three transition models. Therefore, given the assumptions, $\xi_a(1.5)$ is identifiable with the sufficient conditioning sets of vectors $\boldsymbol{X} = (X_1, ..., X_6)$ for Transitions 1 (EM$\rightarrow$ IUI) and 2 (EM$\rightarrow$ preg), and $(\boldsymbol{X}, t_A)$ for Transition 3 (IUI$\rightarrow$ preg).

4. "Estimate: Estimate the target quantity if it is identifiable, or approximate it, if it is not."
Lastly, the previous step implies that the target quantity is identifiable, and it is estimated through the method of G-computation, introduced in Section 3.5. A detailed procedure of estimation of $\xi_{t_A}(t^*)$, for $t^* = 1.5$ and $t_A \in \{0, 0.25, 0.5, 0.75, \infty\}$ is done with the model designed and implemented as described in Section 5.1 is used, and by following the 4 steps of conducting the G-computation method described in Section 3.5.

The outcome variable estimated for each couple is the probability of becoming pregnant in the first 1.5 years of the followup, namely $P(Y(t^*) = 1)$. Therefore, when conditioning on the G-formula from equation (3.6) on the wanted level of exposure $T_{12} = t_A$ and following the notation introduced in Section 5.1, the probability of pregnancy while following the treatment strategy can be written as

$$P(Y_{t_A}(t^*) = 1) = P(Y(t^*) = 1 | T_{12} = t_A) = \sum_{\boldsymbol{X}} P(Y(t^*) = 1 | T_{12} = t_A, \boldsymbol{X}^6) P(T_{12} = t_A | \boldsymbol{X}) P(\boldsymbol{X})$$

Then, the counterfactual outcome Y_{t_A} after intervention $T_{12} = t_A$ can be estimated as

$$\xi_{t_A}(t) = \mathbb{E}(Y_{t_A}(t^*)) = P(Y_{t_A}(t^*) = 1) = \mathbb{E}_{\boldsymbol{X}}[P(Y_{t_A}(t^*) = 1 | \boldsymbol{X})]$$

Since the couples are assumed to be randomly sampled from the population of subfertile couples, the distribution of the couples' covariates is approximated through the already sampled 6-tuples of the covariate values. Therefore, the expectation over the couples' covariates is taken to be the average of the pregnancy probability estimates of all couples in the dataset.

The pregnancy probabilities $P(Y_{t_A}(t^*) = 1 | \boldsymbol{X})$ are computed with the help of transition probabilities in the multi-state model. General transition probabilities are listed in the equations (2.2). However, due to the long computational time of the clock-reset approach, only the formulas computing the values needed are used in this thesis and are presented here. This simplification comes from the fact that all couples start in state 1 at time 0, and that the time of interest when the pregnancy probability is evaluated for each couple equals $t^* = 1.5$ years. The equations used in the computation are

$$\begin{aligned}
&\text{hazards learned from the model:} && \lambda_{12}(t|\boldsymbol{X}),\ \lambda_{13}(t|\boldsymbol{X}),\ \lambda_{23,r}(t|\boldsymbol{X}),\\
&\text{cumulative hazards learned from the model:} && \Lambda_{12}(t|\boldsymbol{X}),\ \Lambda_{13}(t|\boldsymbol{X}),\ \Lambda_{23,r}(t|\boldsymbol{X})\\
&\text{survival in EM (state 1) before fixed} && \\
&\text{transition time } T_{12} = t_A \text{ until } t < t_A\text{:} && S_1^*(t|\boldsymbol{X}) = \exp(-\Lambda_{13}(t|\boldsymbol{X}))\\
&\text{pregnancy without IUI (transition 2) before time } t\text{:} && P_{13}^*(t|\boldsymbol{X}) = \int_0^t \lambda_{13}(u|\boldsymbol{X}) S_1^*(u|\boldsymbol{X}) \mathrm{d}u \qquad (5.3)\\
&\text{survival in IUI (state 2) started at time } r\text{, until } s\text{:} && S_{2,r}(s|\boldsymbol{X}) = \exp(-\Lambda_{23,r}(s|\boldsymbol{X}))\\
&\text{pregnancy with no treatment } (T_{12} = \infty) \text{ until } t^*\text{:} && P_{13}^*(t^*|\boldsymbol{X}) = \int_0^{t^*} \lambda_{13}(u|\boldsymbol{X}) S_1^*(u|\boldsymbol{X}) \mathrm{d}u\\
&\text{pregnancy with treatment strategy } T_{12} = t_A,\ t_A < \infty, \text{ at } t^*\text{:} && P_{13}^{t_A}(t^*|\boldsymbol{X}) = P_{13}^*(t_A|\boldsymbol{X}) + S_1^*(t_A|\boldsymbol{X}) \cdot\\
& && \cdot (1 - S_{2,t_A}(t^* - t_A|\boldsymbol{X}))
\end{aligned}$$

Formulas shown here are a bit different than the ones in equation (2.2) in Section 2.5.2. The difference in the survival and transition probabilities is marked with an asterisk, and this is not related to the notation of the time of interest t^*. Firstly, since the time of the Transition 1 (EM $\rightarrow$ IUI) is known and fixed throughout the computation, model 1 is not used in the formulas. Survival in State 1 depends only on the transition

to pregnancy, i.e. Transition 3. This influences the pregnancy probability when treatment is not given, P^*_{13}, which then relies on this new survival probability. The last formula in equation (5.3) computing the pregnancy probability when the treatment is started at $T_{12} = t_A,\ t_A < \infty$ is a sum of two probabilities. First one is the probability of getting pregnant before this time, namely $P^*_{13}(t_A|\boldsymbol{X})$, and the second one the probability of getting pregnant after this time. For probability of getting pregnant after time t_A, the probability of not getting pregnant before time t_A when knowing the fixed transition time $T_{12} = t_A$, namely $S^*_1(t_A|\boldsymbol{X})$, is multiplied by the probability of Transition 2 (IUI → preg), when IUI has been started at t_A. This second transition needs to happen in the time interval (t_A, t^*) and when this is put into time-reset clock scale, this becomes $(0, t^* - t_A)$. Lastly, the probability of making this transition equals the probability of not surviving in State 2 until time $t^* - t_A$, due to having only one possible transition to make from that state. Also, it should be emphasized that time index t represents the time from the beginning of the followup, and time index s the time from the first IUI cycle. Therefore, the analysis follows the clock-reset time scale.

Then,

$$P(Y_{t_A}(t^*) = 1|\boldsymbol{X}) = \begin{cases} P_1^{t_A}{}_3(t^*|\boldsymbol{X}), & t_A \in \{0, 0.25, 0.5, 0.75\} \\ P^*_{13}(t^*|\boldsymbol{X}), & t_A = \infty \end{cases}$$

Let us now summarize the G-computation procedure. First, the Cox proportional hazards model is fitted on the dataset used in this study. This provides estimates of the hazard functions for each couple in the dataset. Then, the hazard estimates are plugged in the equations (5.3) in order to get estimates for each couple's pregnancy probability until 1.5 years in the followup. These estimates are then averaged to get the EIO estimate, namely

$$\hat{\xi}_{t_A}(t^*) = \frac{1}{N} \sum_{i \in I} \hat{P}(Y_{t_A}(t^*) = 1|\boldsymbol{X}_i)$$

where N is the size of the dataset.

The last part of the G-computation procedure is to quantify the uncertainty of this estimate. This is done by bootstrapping samples with replacement from the original dataset, and conducting the same procedure on each bootstrap sample. This means refitting the multi-state model on the bootstrap sample and through the hazard estimates computing the pregnancy probability estimates. These probability estimates are then averaged and a bootstrap estimate of $\hat{\xi}_{t_A}(t^*)$ is obtained. This is done 200 times with the sample size N. The number of 200 samples was chosen due to the computational time constraints. From the 200 bootstrap samples, 95% confidence intervals (CI) are obtained by sorting the 200 bootstrap estimates in ascending order and using the 5^{th} value as the lower confidence bound and the 195^{th} value as the upper confidence bound.

5.4. Percentage of pregnancies vs. percentage of treated couples

Finally, the comparison of the 5 strategies introduced in the previous subsection is done through comparison of the expected percentage of pregnancies in the population versus the expected percentage of couples undergoing the IUI treatment, evaluated at 1.5 years in the followup, when all couples are following the same strategy. The expected percentage of the pregnancies in the population after 1.5 years equals the expected intervention outcome (EIO) estimate, introduced in Section 3.3 and discussed in Section 5.3. Following the notation from the previous section, the EIO estimate is written as $\hat{\xi}_{t_A}(t^*)$, where $t^* = 1.5$.

On the other hand, the expected percentage of couples undergoing the treatment is computed through computing the average probability of couples becoming pregnant naturally before the time of the first treatment cycle, namely T_{12}. For strategy $T_{12} = a_{t_A}$, the average probability of getting pregnant before the time of receiving the treatment can be written as $\xi_{t_A}(t_A)$. Then, the expected percentage of couples undergoing the treatment equals $1 - \xi_{t_A}(t_A)$.

The value of $\xi_{t_A}(t_A)$ is estimated in the same way as the main estimate of $\xi_{t_A}(t^*)$, and so the procedure described in Section 5.3 is followed. The only difference is that now $t^* = t_A$ instead of 1.5 years. This implies the estimand being computed as follows

$$\hat{\xi}_{t_A}(t_A) = \frac{1}{N} \sum_{i=1}^{N} \hat{P}(Y_{t_A}(t^*) = 1|\boldsymbol{X})$$

This estimate can be bootstrapped in the same way as the EIO estimate, as explained in Section 5.3, and is computed from the same 200 bootstrap samples used for the bootstrap of the EIO estimate. Therefore, the 200 bootstrap estimates of $\hat{\xi}_{t_A}(t_A)$ are used to compute the 95% confidence intervals of the main EIO estimate. This is done in the same way as before, namely sorting the 200 values in ascending order, and using the 5^{th} value as the LCB and the 195^{th} value as the UCB.

From these estimates of the probability of natural pregnancy before receiving the treatment, the expected percentage of the couples undergoing the treatment is computed. Namely, the main estimate of it is $1 - \hat{\xi}_{t_A}(t_A)$. The corresponding 95% CI is computed in the following way. Let $[a, b]$ be the 95% CI of the $\hat{\xi}_{t_A}(t_A)$, then the 95% CI for $1 - \hat{\xi}_{t_A}(t_A)$ equals $(1 - b, 1 - a)$.

It can be noted that when $t_A = 0$, the whole population will be treated, and therefore $1 - \xi_{t_A}(t_A) = 1$ as its true value. Moreover, when $t_A = \infty$, none of the couples will be treated and therefore $1 - \xi_{t_A}(t_A) = 0$.

Further analysis can be done on this project by defining a utility function over how many extra pregnancies in the population are worth how many extra treatments. The next section provides the results of applying the newly designed methodology introduced in this section to the real data and provides insight into this trade-off.

6

Results

This section presents the results of the analysis described in Section 5. First, the model used is commented through the hazard ratio estimates. Then, percentages of the couples who achieve a pregnancy in the population and the number of treated couples, when everyone would follow the same treatment strategy, are presented and commented upon. Furthermore, the probability curves over time of the treatment strategies are depicted, applied on a certain example couple.

More about the model used for the analysis, introduced in Section 5.1, can be read about in Appendix C. This includes the comparison of the model's survival predictions to their non-parametric Kaplan-Meier estimates from the data.

The first level of inspection of a fitted model is through its coefficients, or in the case of Cox model, its hazard ratios. They are computed for each transition, with the clock-reset time scale, and are presented in Table 6.1. Hazards ratios larger than 1 mean that an increase in the covariate's value increases hazard of reaching the state of interest, and hazards ratios smaller than 1 have the opposite effect. Furthermore, hazard ratios represent only the effect of the unit change of the covariate value on the hazard curve, and therefore the range of a covariate's values needs to be taken into account when determining the covariates' influences on the hazards. For this reason, the covariates' units are displayed in the table as well. Lastly, the 95% confidence intervals for the hazard ratio estimates are displayed in brackets.

Table 6.1: Hazard ratios for the covariates in the three transition models, with their asymptotic 95% confidence intervals in brackets and units the covariates are modeled in, using the clock-reset time scale

	Hazard ratio			
Covariate	Transition 1 EM → IUI	Transition 2 EM → preg	Transition 3 IUI → preg	Unit
Female age	1.018 (1.002, 1.035)	0.952 (0.930, 0.974)	0.991 (0.949, 1.034)	years
Subfertility duration	0.928 (0.884, 0.975)	0.773 (0.690, 0.867)	0.979 (0.857, 1.119)	years
Infertility type	1.279 (1.100, 1.487)	0.740 (0.601, 0.910)	0.873 (0.613, 1.243)	binary
Doctor referral	0.561 (0.435, 0.722)	0.673 (0.444, 0.018)	1.312 (0.787, 2.186)	binary
Tubal blockage "none"	1.718 (1.309, 2.254)	0.662 (0.399, 1.101)	0.620 (0.282, 1.364)	binary
Tubal blockage "1-sided"	2.179 (1.888, 2.515)	1.056 (0.848, 1.314)	1.156 (0.818, 1.634))	binary
Progressive sperm count	0.997 (0.994, 1.001)	1.001 (0.996, 1.006)	1.001 (0.993, 1.010)	percentage
Treatment timing			0.983 (0.715, 1.351)	years

Reference category for infertility type is 'primary infertility', for doctor referral is 'no doctor referral'.
The tubal blockage represents the outcome of a diagnostics test of fallopian tubes blockage, with the available options 'no blockage', '1-sided blockage' and 'no test'.

Table 6.1 suggests that a few covariate in all three models are non-influential, which is seen by their confidence intervals containing number 1. All covariates are influential in model 1, representing the transition from expectant management (EM) into the IUI treatment. On the other hand, four out of seven covariates are

influential in model 2, representing transition from EM into the pregnancy. These are female age, subfertility duration, infertility type and doctor referral. Lastly, none of the covariates are influential in model 3, representing the transition from the IUI into the pregnancy. Moreover, it can be noticed that covariate for the start of the IUI is, next to being not significantly influential to the pregnancy probability, also valued almost the same as the subfertility duration variable, and their confidence intervals are also similar, with the time of the treatment being slightly more volatile estimate. This supports the assumption that the time after the start of the follow-up is equally important and influential to the pregnancy probability as the subfertility duration is, and they they can be added together and the results would be similar.

Then, the estimates on the population pregnancy probability at 1.5 years after the start of the follow-up, when whole population follows the same treatment strategy, are presented in Table 6.2 in the second column. Their bootstrap 95% CI are provided in brackets and all percentages are proportions in the population. Furthermore, the procedure of how to obtain these estimates was described in Section 5.3. The table shows that not treating anyone yields approximately 32.6% of pregnancies among the couples. However, if the treatment is given to the couples not already pregnant after 9, 6 or 3 months, the pregnancy probabilities are 41.7%, 41.8% and 41.0%, respectively. Lastly, if everyone is treated immediately, the pregnancy probability is around 42.4%. It can be seen that the timing of starting the IUI treatment yields almost the same pregnancy probability in the first 1.5 years of the follow-up, but it increases the pregnancy probability for around 1.28 times compared to the no treatment strategy. Moreover, the strategy to start the treatment immediately is the most volatile, with the length of the 95% CI of 26%, and this decreases to 14%, 10% and 7% for delaying the treatment, and 6% for not giving the treatment at all.

Furthermore, the percentages of the treated couples in the population when following certain strategy are presented in Table 6.2 in the third column, with the bootstrap 95% CI written in the brackets. The procedure of obtaining these estimates is described in Section 5.4. When no treatment is given, all pregnancies are natural pregnancies, therefore the treated couples percentage is 0%. It should be noted that 1-0.326=0.674% of couples are neither pregnant not have received the treatment at the end of 1.5 years of follow-up when following no treatment strategy. Furthermore, if the treatment is given to the couples not already pregnant after 9, 6 or 3 months, the percentage of couples that receive the treatment are 78.8%, 83.7% and 90.1%, respectively. Lastly, if everyone is treated immediately, all couples receive the treatment and therefore the percentage of couples treated is 100%. The first and last estimates are sharp, and the rest are volatile, with the 95% CI range of 7%, 4% and 3% respectively. These CI are more narrow than the CI for the pregnancy percentages due to the time of the estimation being smaller than 1.5 years.

From this analysis it can be concluded that delaying the IUI treatment on average yields the same pregnancy probability at 1.5 years after the workup as starting the treatment earlier, and moreover less IUI treatments need to be conducted among the couples seeking help. However, it is important to start the treatment at some point. This analysis can conclude that delaying the treatment for 9 months, among the choices proposed in this research, is the most suitable treatment strategy to recommend the couples when one strategy is followed in the whole population and when the goal is to have as few possible treatments with as high as possible pregnancy percentages after 1.5 years after the workup.

Table 6.2: Estimated population pregnancy percentage and estimated percentage of treated couples 1.5 years in the followup, when starting the treatment at a fixed time after the start of the followup, with their 95% CI in the brackets

Treatment strategy	Proportion of couples with pregnancy at 1.5 years	Proportion of couples receiving treatment within 1.5 years
Treat immediately	0.424 (0.302, 0.565)	1.000 (1.000, 1.000)
Treat at 3 months	0.410 (0.343, 0.485)	0.901 (0.888, 0.914)
Treat at 6 months	0.418 (0.374, 0.476)	0.837 (0.819, 0.857)
Treat at 9 months	0.417 (0.383, 0.454)	0.788 (0.766, 0.812)
No treatment	0.326 (0.292, 0.358)	0.000 (0.000, 0.000)

The estimated average pregnancy probabilities in the population when following all 5 treatment strategies are visualized over time in Figure 6.1. It can be seen from the figure that the treatment provides immediate increase in the hazard of getting pregnant. In the long run, however, the pregnancy probabilities including the treatment are close to each other, but significantly different compared to the strategy when treatment was

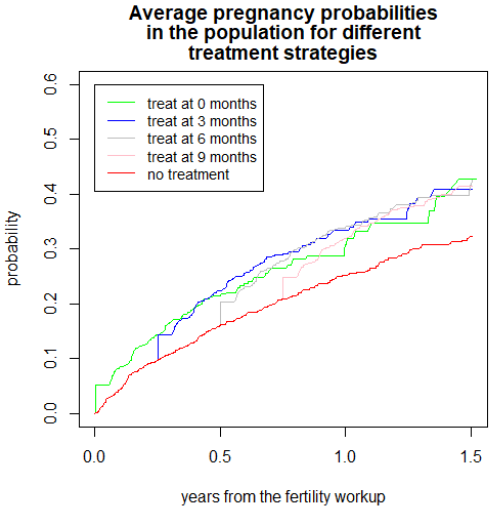


Figure 6.1: Estimated average cumulative pregnancy probability for the population of subfertile couples when following different treatment strategies

not given. This is supported by the 95% confidence intervals of the estimates shown in Table 6.2. This implies that undertaking the treatment significantly improves the probability of getting pregnant, but the timing of the treatment in the first 9 months has no influence on long term (1.5 years) prognosis.

Then, two examples of couples with higher and lower probabilities of naturally get pregnant are displayed. The higher pregnancy probability couple is taken to have female age 25 years, subfertility duration 1 year, progressive sperm count 80 percent, no referral from gynecologist, secondary infertility type and no tubal blockage test done. Their probability curves can be seen in Figure 6.2 on the left. On the other hand, the lower pregnancy probability couples is taken to have female age 38 years, subfertility duration 4 year, progressive sperm count 15 percent, referral from gynecologist, primary infertility type and 1-sided tubal blockage. Their probability curves can be seen in Figure 6.2 on the right. The two couples presented are designed to get a sense of how do the counterfactual pregnancy probabilities change from couple to couple. From the two examples it can also be seen that the IUI treatment increases the pregnancy probability significantly and is a few times higher than the natural pregnancy probability. Moreover, the effect of the treatment timing is larger for a couple with lower natural pregnancy probability, whereas couples with higher pregnancy probability can expect much lower effect of the treatment, even negative one, meaning the pregnancy probability with the treatment being lower than the natural pregnancy probability. This violation of the assumption that the IUI treatment has positive or neutral effect on the pregnancy probability may imply that the model at hand is not suited to make accurate counterfactual predictions for couples farther away from the average pregnancy probabilities in the population. However, the curve of natural pregnancy prediction shown in Figure 6.1 coincides with studies reporting the natural pregnancy probability for couples with unexplained subfertility on the same dataset as this study such as [39] and [35].

The following section provides a discussion on the methodology and the significance of the results presented in this section, and answers the research questions from Section 1.4. Then, conclusions are drawn from the research done and a few recommendations are given for potential future work building on this research.

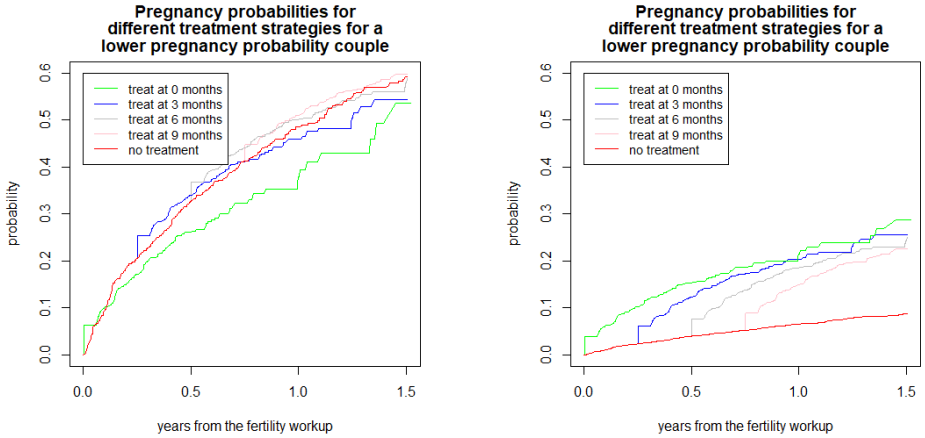


Figure 6.2: Cumulative pregnancy probability for a couple having higher (left) and lower (right) probability of natural pregnancy, when following 5 different treatment strategies. Covariates for the couple with higher probability of natural pregnancies are: female age 25 years, subfertility duration 1 year, progressive sperm count 80 percent, no doctor referral, secondary infertility type and no tubal blockage test done; and covariates for the couples with lower probability of natural pregnancy are: female age 38 years, subfertility duration 4 year, progressive sperm count 15 percent, doctor referral, primary infertility type and 1-sided tubal blockage.

7

Discussion and Conclusion

This section provides a discussion on the research conducted and the two research questions posed in Section 1.4, and draws conclusions from this research study. Firstly, the application part of the research is discussed, namely the importance of the IUI treatment timing. After this, methodology is discussed, namely if the G-computation method works well with counterfactual predictions done with multi-state models. Lastly, a few suggestions for building further on this work are provided.

The first part of the research goals refers to the problem of quantifying the influence the IUI starting time has on the pregnancy probability. This was done by using 3-state multi-state model designed in Section 5.1 connected to the causal inference setting. The model was used to make counterfactual predictions of pregnancy probability in the first 1.5 years in the follow-up for each couple in the dataset, given that they started with the IUI treatment at a predetermined time in case they are not yet pregnant then. The probability predictions are then averaged over all couples in the dataset.

The procedure to get to these estimates, namely designing the multi-state model and connecting it to the causal inference setting, was done in Section 5. Moreover, certain assumptions that need to hold in order for the results to be bias-free were listed and commented upon in Section 5.2. In this study, assumptions that can be checked hold, and those that are not testable are reasoned why they are most likely satisfied or that the potential bias resolving from broken assumptions is lowered as much as possible. Specifically, selection bias and attrition bias are assumed to be low but are not tested, stable unit treatment value assumption, structural property and no unmeasured confounding assumption are reasonable to hold, the latter one implying sufficient assumption for identifiability to hold, then correct model specification assumption is satisfied due to linearity, hazard proportionality and outlier detection tests, and from which the consistency assumption follows, and positivity assumption is checked to hold. Another assumption that was made considers the use of the clock-reset time scale in the multi-state model. The likelihood of the fit of the models with the two time scales is very similar, but clock-reset approach is chosen due to the expectation of time to pregnancy to be influenced dominantly by the time from the treatment as opposed to the time of start of the follow-up.

What also influences the analysis, are the times of starting the IUI treatment and the time frame of the analysis. The time frame was chosen as described in Section 4.3.1, where it was seen that there is a sharp decrease in couples in the study after 1.5 years. Then, treatment starting times were chosen to be evenly spread throughout this time interval, while respecting the positivity assumption, i.e. that there is enough couples around those times who were observed to start the treatment then. The distance between the times analyzed was chosen as 3 months, so not too small nor too wide delay in the treatment, and the number of couples around each timing can be seen in Appendix C.2.

Results of this analysis are shown in Table 6.2, Section 6, where it can be seen that there is a significant difference between pregnancy probabilities if the IUI was started or not. However, the timing of the IUI in the first 9 months seems to be completely non-influential on the final pregnancy probability in the population of couples with unexplained subfertility after 1.5 years in the follow-up. Due to the similarity of the effects of strategies when treatment was employed being similar, the proposed trade-off based on the number of couples treated in the population clearly implicates that delaying the start of the IUI until 9 months in the follow-up is recommendable. However, this is the result based on evaluating the effect of the treatment strategies only at 1.5 years into the follow-up, and not during this time. If the goal is to have the largest prob-

ability of getting pregnant as soon as possible, IUI should be started immediately, based on results in Figure 6.1.

The main drawback of the study is the small dataset used, the size of which is 1896 couples and the fact that the IUI could have been started at any time throughout the 4 years of follow-up. It presents a risk of not being able to pick up important patterns due to lack of data, an dalso the risk of overfitting on the couples in this study. Moreover, another drawback is that there is no details about the cycles of the IUI, for example how often are they given. In this research it is assumed that the effect of the IUI from its start until 1.5 years into the follow-up is the same for all couples and all IUI starting times. For these reasons, it is of strong medical interest to conduct a similar study quantifying the IUI treatment timing effect for couples with unexplained subfertility from the causal inference prespective on a larger dataset, possibly including the data on conducting the IUI cycles.

The second part of the goals of this thesis is discussed. This concerns the methodology used in this study, namely can multi-state models and G-computation be used to answer causal questions. This research shows that it can be done, with the methodology using an illness-death multi-state model and the treatment strategies designed in this project has been thoroughly described in Section 5. Therefore, discussion on this research question is based on the contribution of this methodology to the current state of science. The main point of reference of the work previously done on this topic is [10], the one study found that quantifies the effect of interventions on a multi-state model. The paper use a larger multi-state model and applied more methods and researched two different kinds of interventions, however the description of the work is restricted. This thesis conducts one part of this paper, namely the G-computation method, to see the difference in probabilities of being in one state conditioned on starting from another when conditioning on the time of the transition, in much greater detail. Here, the current causal inference theory is laid down and multi-state model is carefully connected to it. In general, multi-state models are currently mostly used for predictive analyses, and using them for answering causal questions open doors to many yet unanswered research questions. Moreover, the definition of treatment strategies perfectly aligns with the way interventions can be done on a multi-state model. In this work, there are only two possible path from the initial state (EM) into the final one (pregnancy) and so the probability curves of reaching the final state are easier to interpret than in [10]. Therefore, it is easier to track what is happening with one patient transitioning among the states. Additionally, the curves of reaching pregnancy state as shown in Figure 6.1, Section 6, are not possible to acquired by using one survival model with time-varying treatment time covariate. This is another contribution of this work to the field of survival analysis. Finally, G-computation method is an intuitive method to estimate causal estimands, when multiple treatment strategies are being compared.

Lastly, a few suggestions are given for possible future work building on this research. First suggestion is to use the model designed in this study and provide different treatment strategies for different types of couples. This would mean finding a way to make cohorts of similar couples based on their covariates or pregnancy probability predictions. This approach would provide better suited recommendations for couples and it would be interesting to see which types of couples would benefit the most from which IUI treatment timing. This work would come down to finding the criteria based on which to appropriately cluster the couples in population with unexplained subfertility.

Another suggestion for building on this work is to challenge the connection of the multi-state models and causal inference setting in a different context, namely making interventions on a time of a transition being inside a chosen time interval. This can be done by using the hazard function during the period of time of interest. Specifically, one can define a probability of a couple making this transition at time t, where t is some moment inside the chosen time interval, as hazard of the transition at time t divided by the cumulative hazard from time t until the end of the chosen time interval. This would yield the probability of transition being 1 at the end of the chosen time interval, thus confirming that the transition will definitely happen during this time interval. This suggestion can be used to allow some flexibility for couples to choose the exact cycle in which to conduct IUI treatment, thus the model would be better suited to reflect variations in treatment schedules among the couples.

The last suggestion for building on this work is to include the data from the same Dutch study on fertility that includes data on IVF as well, and include it in the multi-state model. Then, interventions could be made on IUI and IVF treatment timing simultaneously, allowing to see which combination of timings provides the highest pregnancy probability in the population of couples with unexplained subfertility. Also, if data on the exact time of the IUI or IVF cycles is available, a more precise trade-off of pregnancy probability and the

number of cycles couples undertake can be done.

A

Axiomatization of the Rubin's causal model

This section covers assumptions of Rubin's potential outcomes model and relates these assumptions to the ones the structural equations modeling is based on.

Firstly, a way of describing the relationships among the variables needs to be included. Without the visual help of path diagrams, this is achieved through independence and exclusion restrictions. To illustrate the definitions, Figure A.1 is used. Therefore, to state that variable W in this model does not have a direct influence on variable Y, but only through variable A, the following can be written

$$\begin{aligned} \textit{independence}&: W \perp\!\!\!\perp \{Y_{w_1}, ..., Y_{w_k}\} \quad \text{for all realisations } w_k \text{ of } W \\ \textit{exclusion}&: Y_{aw} = Y_a, \quad \text{for all realisations } w \text{ of } W \text{ and } a \text{ of } A \end{aligned}$$

Furthermore, to connect the potential outcomes and SEM framework, firstly the counterfactual outcomes are seen as outcomes after the appropriate intervention, namely $Y_a(u) = Y_{C_a}(u),\ \forall a, u$. The consistency assumption is satisfied in SEM in case the model is describes the data well, so this is not related to the research problem, but the statistical modeling approach. Moreover, the two ways to state the variable relationships can be translated to the structural theory language through conditioning on a variable's parents values. To set up the notation, let's call the set of parent variables $PA(W)$, for any variable W, and their possible values written in small letters. Then, the two restrictions can be rewritten as

$$\begin{aligned} \textit{independence}&: Y_{pa(Y)} \perp\!\!\!\perp \{W_{1\,pa(W_1)}, ..., W_{k\,pa(W_k)}\} \text{ for any set of variables } \{W_j\}_j \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{not having a mutual cause with } Y \\ \textit{exclusion}&: Y_{pa(Y)} = Y_{pa(Y),b}, \text{ where } B \text{ is any set of endogenous variables disjoint from } PA(Y) \end{aligned}$$

Last crucial aspect not yet mentioned is finding a set of assumptions sufficient to translate the counterfactual distributions to the ones observable from the data. [26] proposed the concept of conditional ignorability in order to find the set of variables to condition on which will lead to unbiased causal estimates. The property of conditional ignorability implies that the potential outcome Y_a after an intervention $do(A = a)$ is independent of the observed exposure level A, conditioned on some set of variables W and can be written as

$$Y_a \perp\!\!\!\perp A \mid W$$

Furthermore, they prove that if there is a set of variables W that satisfies the conditional ignorability assumption, then the counterfactual probability of Y_a can be computed through conditioning on W. Then, if $P^*(Y_a)$

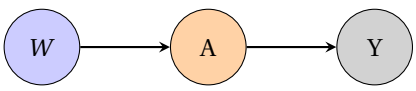


Figure A.1: A simple causal graph for illustration of axioms of potential outcomes model; W=endogenous variable, A=independent variable, Y=dependent variable

is the counterfactual probability distribution, $P^*(Y_a = y) = P(y|do(A = a))$ and equations 3.1 and 3.2 follow. Therefore, the conditional identifiabilty condition, extremely difficult to use agilely, is equivalent to the back-door criterion.

With these key assumptions, definitions and the connection of counterfactual to statistical distributions it can be concluded that SEM is subsumed in Rubin's causal framework, and all that holds in SEM also holds in potential outcomes setting. Therefore, even though the model used for the analysis in this thesis will be defined through SEM and a method of estimation set in Rubin's causal model will be used, the method will yield unbiased causal estimators.

B

Other causal inference methods

The two methods populat for estimating causal estimands that were not used in this study but are briefly refered to are described in this appendix. These are inverse probability of treatment weighting (IPTW) and targeted maximum likelihood estimation (TMLE). At the end, a flowchart of the steps of the two methods and G-computation can be seen in Figure B.1.

B.1. Inverse probability of treatment weighting

This method provides a simpler yet differently approached EIO estimate than G-computation. It is based on weighting the individuals based on the exposure level and obtaining the weighted mean estimate of the treatment effect. The weights equal the inverse of a patient's probability of receiving the treatment, conditioned on his/hers covariate values. If w_j notes a weight for a patient j with exposure a_j and covariate vector w_j, weights can be written as

$$\alpha_j = \frac{1}{P(A = a_j | W = w_j)} \tag{B.1}$$

and, if I is the set of all patients in the dataset, the ATE estimator is given by

$$\hat{\psi}^{IPTW} = \frac{1}{|I|} \sum_{j=1}^{|I|} \alpha_j Y_j$$

This way, groups of patients similar in their covariate values are made approximately equally important i.e. influential on the final estimate of the ATE. This way an observational study behaves as an experimental one, without a loss of available information.

The probability is estimated from the dataset, usually by fitting a model on the exposure variable, depending on the covariates. This can be done with Cox proportional hazards or any other model suitable as a model fit. This step relies on the no model misspecification assumption, as well as the G-computation, only it is applied on the exposure model. If the assumption holds, the IPTW ATE estimate is unbiased.

Dealing with inverse number of possibly small probabilities, the weights might vary and escalate too much to trust the estimator in this case. To resolve this problem, the stabilised IPTW estimator has been proposed. It is based on dividing the covariate vector into a vector C that are the confounders and a vector D that influence only the exposure, but not the outcome variable. Now, the stabilised inverse probability-to-treatment weights are given by

$$\alpha_j = \frac{P(A = a_j | D = d_j)}{P(A = a_j | C = c_k, D = d_j)} \tag{B.2}$$

and the estimator is written in the same way, only with these new weights.

B.2. Targeted maximum likelihood estimation

Targeted maximum likelihood, or TMLE, is an approach that combines G-computation and IPTW. It uses the concept of G-computation algorithm, but has a medium step of weighting individuals just as in the IPTW method. The algorithm then, in short, comes to:

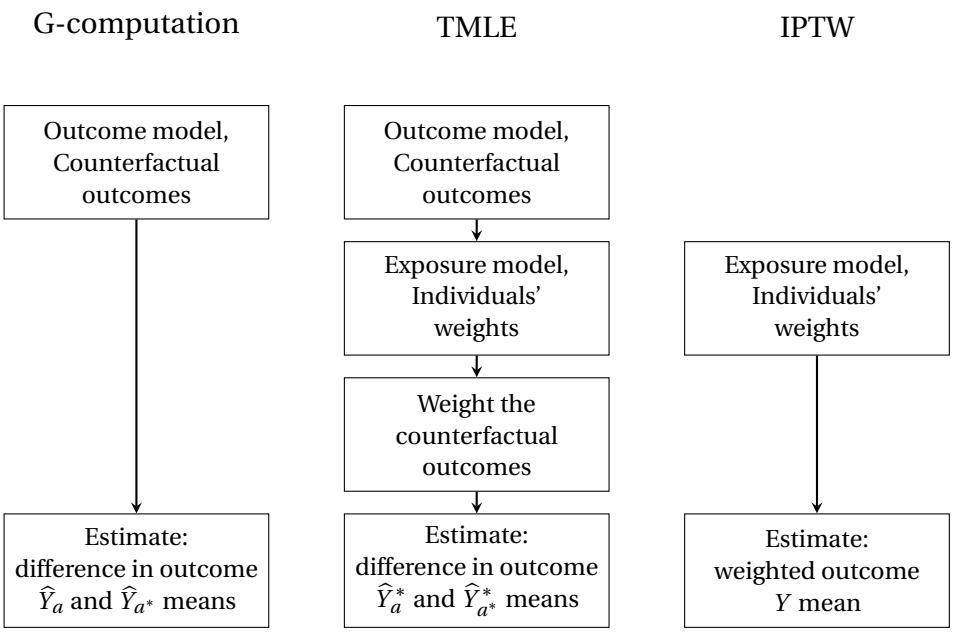


Figure B.1: Visualised steps in the G-computation, IPTW and TMLE algorithms

1. Fitting a model on the outcome variable,
2. Computing counterfactual outcomes on the data subset,
3. Fitting a model on the exposure variable, computing the weights α_j^a and $\alpha_j^{a^*}$ as shown later in equations (B.3) and (B.4),
4. Estimating ATE through the weighted expression $\hat{\psi}_I = \frac{1}{|I|}\sum_{w\in I}[\alpha_j^a \widehat{Y}_a - \alpha_j^{a^*}\widehat{Y}_{a^*}]$,
5. Conducting the steps 1-4 on a bootstrap sample and computing 95% CI for the estimate.

The steps repeat the procedure of the G-computation and IPTW, with the step 4 being the one that combines the two methods. The weights are computed as in equation (B.1) or (B.2), only that both counterfactual outcomes are weighted. By this we get

$$\alpha_j^a = \frac{1}{P(A=a|W=w_j)}, \quad \alpha_j^{a^*} = \frac{1}{P(A=a^*|W=w_j)} \tag{B.3}$$

or similarly, for the stabilised weights,

$$\alpha_j^a = \frac{P(A=a|D=d_j)}{P(A=a|C=c_j,D=d_j)}, \quad \alpha_j^{a^*} = \frac{P(A=a^*|D=d_j)}{P(A=a^*|C=c_k,D=d_j)} \tag{B.4}$$

Combination of the two approaches results in a doubly-robust estimator of ATE. This means that only one f the two models, outcome and exposure model, needs to be correctly specified in order to get an unbiased estimator. The flowchart of the steps of methods G-computation, IPTW and TMLE and their similarities can be found in Figure B.1.

C

Model analysis

This section provides more information about the model that is used in this study, introduced in Section 5.1, and represents an addition to Sections 5 and 6. Firstly, the data format used to implement the model is described in more detail. Then, positivity and proportional hazards assumptions are checked and the procedure of testing them is provided. Lastly, the empirical comparison of the Cox multi-state model's predictions with the non-parametric Kaplan-Meier estimates is conducted.

C.1. Dataset format

Firstly, the format of the dataset used for the analysis in R is described. The format is motivated by the fact that all three transitions are estimated from the data at once, which is feasible to do through stratification. An example of a few rows from the dataset formatted as required to conduct the stratified Cox modeling is presented in Figure C.1. After identification number of a couple, column representing the transition that each row represents is included. This column is used to stratify the dataset and provide the Cox model baseline hazards and hazard ratios for each transition separately.

Therefore, in order to allow different covariate coefficients in each transition model, each covariate column X in the dataset is replaced by three columns named $X.1$, $X.2$, $X.3$. These satisfy that $X.k = X$ and $X.j = 0$ for $\forall\, j \neq k$, where $k = 1,2,3$. Furthermore, this feature allows for including $ttime$ into the model for transition 3 only. An example of the decomposition of one couple who has the female age 32, subfertility duration 2 years and has started the treatment after 6 months in the followup is shown in Table C.1.

Table C.1: Representation of the covariate decomposition on the three transitions from the multi-state model, shown on an example couple

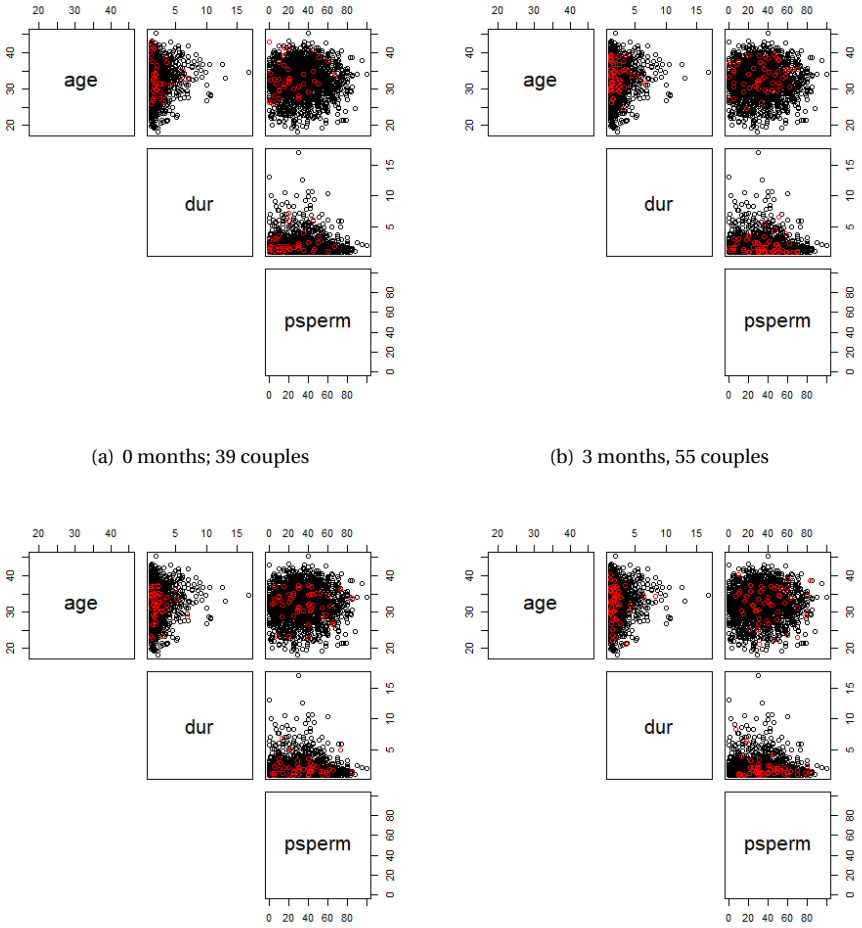
id	trans	age.1	age.2	age.3	dur.1	dur.2	dur.3	...	ttime.3
1	1	32	0	0	2	0	0	...	0
1	2	0	32	0	0	2	0	...	0
1	3	0	0	32	0	0	2	...	0.5

C.2. Testing the assumptions

Positivity assumption

Positivity assumption is introduced in Section 3.4 and it is checked empirically looking into data. As mentioned in Section 5.2, pair-wise covariate scatterplots for couples who received the treatment at 0, 3, 6 and 9 months are compared to the ones of all couples in the dataset. The time interval around these 4 treatment timings is defined as 2 weeks before and 2 weeks after. For example, for starting the treatment at time 0, time interval (0, 0.039) is taken, representing the first two weeks measured in years. On the other hand, for starting the treatment at 3 months, time interval (0.208, 0.292) is taken. The number of couples who received the treatment during this time are 39, 55, 48 and 53 respectively. The space coverage can be seen in Figure C.1. From visual inspection of the reults in figure it is concluded that the pair-wise space is well covered, and so

the conclusion is generalized on the whole covariate space. Moreover, 32 and 19 couples in the dataset have started the IUI treatment in 1-month interval around 12 and 15 months in the follow-up, respectively. This is somewhat lower than other four timings, and so these two will not be considered in the analysis.



(a) 0 months; 39 couples

(b) 3 months, 55 couples

(c) 6 months, 48 couples

(d) 9 months, 53 couples

Figure C.1: Groups of couples who received the treatment around the times indicated in subcaptions colored in red, contrasted in pairwise covariates (female age, subfertility duration and progressive sperm count) to all couples in the dataset colored in black; the number of couples in each group indicated in the subcaption

Moreover, the 5^{th} treatment strategy is to never get the treatment, and therefore the scatterplots for couples who got the treatment at any time before being pregnant or censored are plotted. This scatterplot can be seen in Figure C.2. Almost half of the couples received the treatment and it can safely be assumed that positivity assumption holds in this case as well.

Linearity assumption

Testing for linearity between log hazards and the covariates is the first diagnostic test which needs to be done in order to analyze the fit of the Cox proportional hazards model on the dataset at hand. This is done through inspecting the martingales residuals of the numeric covariates. These residuals capture the difference of predicted and observed outcomes for each couple in the dataset and plotting them compared to the numeric covariates. This allows to see the functional form the coefficients in the Cox model should have in order to capture true effect of the covariates on the log hazard function. The residuals are shown in Figure C.3 for the three numeric covariate and for all three transition models. The fitted lines are close to being linear, except for the subfertility duration covariate in model 1. However, if a logarithmic transformation is done on the

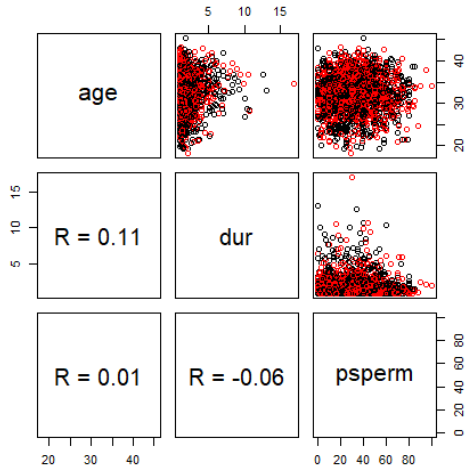


Figure C.2: Couples who got the treatment at any time before being pregnant or censored are indicated in red, contrasting the couples who got pregnant or censored without initiating the treatment colored in black

covariate values, the fitted lines becomes close to being linear, as is shown in Figure C.4. Therefore, after including this transformation in the analysis, the linearity assumption holds.

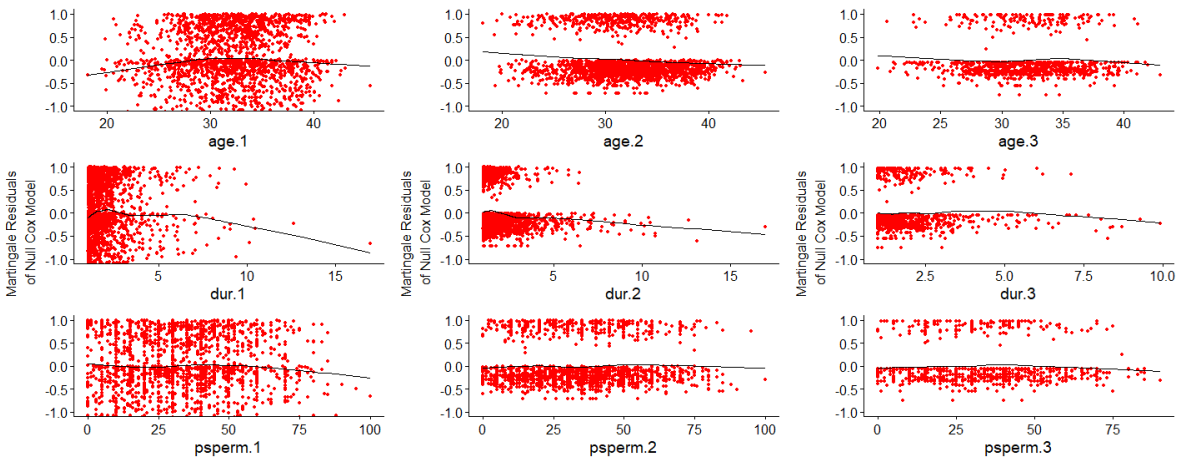


Figure C.3: Martingale residuals for testing the linearity between the log hazards and covariates in the multi-state model: model 1, 2 and 3 from left to right; the numeric covariates female age (age), subfertility duration (dur) and progressive sperm count (psperm)

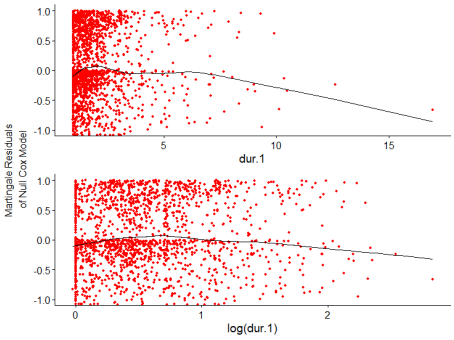


Figure C.4: Martingale residuals for testing the linearity between the log hazards and the subfertility duration covariate in the model 1; original data on the top and logarithmic transformation on the bottom

Hazards proportionality assumption

In order to see if the model satisfies the main assumption of the Cox model, namely the proportional hazards assumption, the Schoenfeld residuals are used as the most common way to test this. Essentially, they represent the difference between the model outcome prediction and the observed outcome in the data. Chi-squared test is preformed on them, and the results are written in Table C.2. The results show insignificant difference in predictions and observed outcomes for all covariates except for female age and 1-sided tubal blockage in model 1.

Table C.2: Chi-square statistics and p-value for Schoenfeld residuals over time for the three transition Cox models

	Transition 1 EM → IUI		Transition 2 EM → preg		Transition 3 IUI → preg	
Covariate	Statistic	p-value	Statistic	p-value	Statistic	p-value
Female age	11.871	0.001	0.118	0.731	0.319	0.572
Subfertility duration	5.622	0.018	0.004	0.948	0.310	0.577
Infertility type	0.825	0.364	0.355	0.551	2.161	0.142
Doctor referral	9.886	0.002	0.998	0.318	0.233	0.629
Progressive sperm count	4.746	0.029	1.156	0.282	0.029	0.865
Tubal blockage "none"	4.116	0.042	0.827	0.363	1.550	0.213
Tubal blockage "1-sided"	21.922	<0.001	0.009	0.923	0.431	0.512
Treatment timing					3.161	0.075
Global test			67.260	<0.001		

The tubal blockage represents the outcome of a diagnostics test of fallopian tubes blockage, with the available options 'no blockage', '1-sided blockage' and 'no test'.

Numeric covariates can be separated into two covariates based on a certain treshold value and binary variables can be used to stratify the model on, thus allowing different baseline hazard function for the two possible values of the covariate. Since 1-sided tubal has the highest Chi-squared test statistic, first the model for Transition 1 is stratifies based on that covariate. The Chi-square test results are now reported only for the Transition 1 model and after this separation can be seen in Table C.3 in the second column. Then, covariate representing if the tubal blockage test was done or not shows high non-proportionality in the hazard, so we also stratify on that covariate. Chi-squared statistic after this can be seen in Table C.3 in the third column. And lastly, only female age covariate is left with high non-proportionality. When certain treshold numbers were tried to minimize the non-proportionality, it turned out that age of 32 was the best to use. Therefore, results after female age covariate was separated into two covariates with treshold value of 32, test statistics can be seen in Table C.3 in the fourth column. The global test statistic for the whole model now equal 35.464 with p-value of 0.018. With this showing high enough hazards proportionality, this version of the model is used further in the analysis.

Table C.3: Chi-square statistics and p-value for Schoenfeld residuals over time for the three transition Cox models

	Stratify on tubal blockage "1-sided"		Stratify on tubal blockage "none"		Separate female age with threshold of 32 years	
Covariate	Statistic	p-value	Statistic	p-value	Statistic	p-value
Female age	9.853	0.002	9.077	0.0026		
Female age < 32					0.682	0.409
Female age ≥ 32					1.985	0.159
Subfertility duration	3.270	0.071	2.763	0.097	2.641	0.104
Infertility type	0.229	0.632	0.390	0.534	0.396	0.529
Doctor referral	7.083	0.008	5.081	0.024	5.006	0.025
Progressive sperm count	6.071	0.014	5.597	0.018	5.576	0.018
Tubal blockage "none"	11.355	0.001				

The tubal blockage represents the outcome of a diagnostics test of fallopian tubes blockage, with the available options 'no blockage', '1-sided blockage' and 'no test'.

Detecting influential outliers

Detecting influential outliers in the data is the last test of goodness of fit for the Cox proportional hazards model. The test is done through so-called dfbeta residuals, which represent each covariate coefficient in the model if one observation has been deleted from the dataset. The test is implemented through the function *ggcoxdiagnostics()*, and the plots of the coefficients for each covariate are shown in Figure C.5. Here we are looking for the maximal influence on the cofficients one observation may have, and conclude that all dfbeta residuals are in reasonably small intervals compared to their influence on the model. For example, the largest difference in coefficient occurs for the subfertility duration covariate in the third model, where the average coefficient value is 0 and the maximal value is 0.02. Because this is a numerical variable obtaining value in range (0,4), this makes for some difference in the model but is not considered as significant.

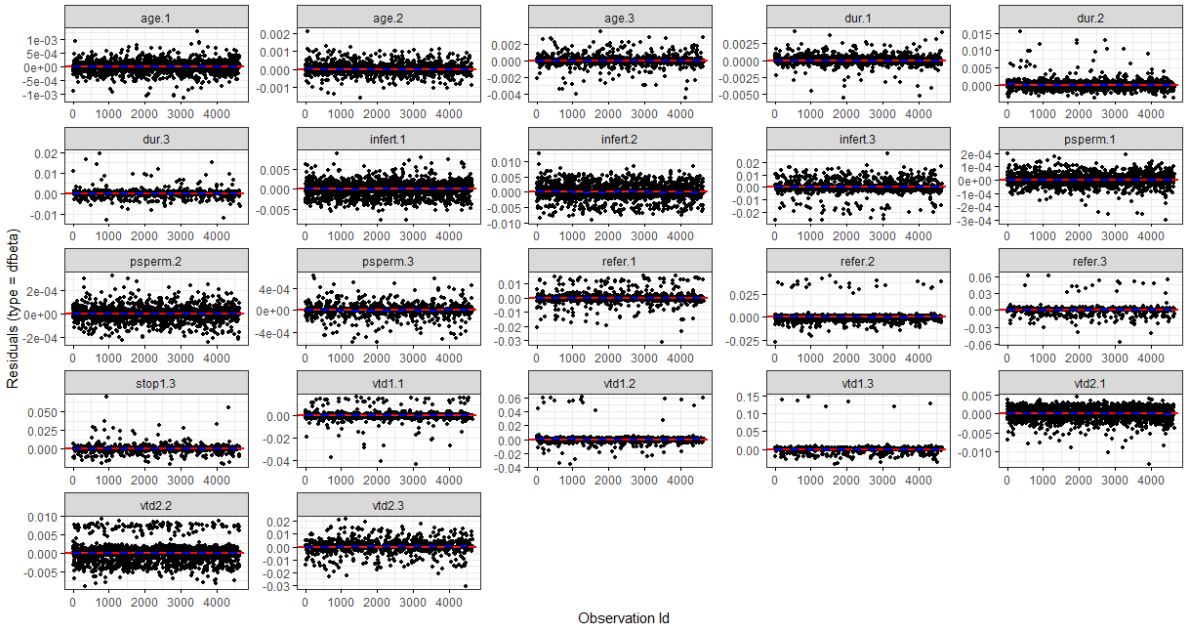


Figure C.5: Dfbeta residuals of the multi-state model on the dataset used in this study

Another way of examining influential outliers is through deviance residuals. These are normalized martingale residuals. Positive residual values mean that this couple got pregnant later than the model predicts, and the negative values that the couple got pregnant earlier than the model predicts. Therefore, outliers should have extreme residual values, since they are the most difficult to predict. The test is implemented through the same function *ggcoxdiagnostics()*, and the plot of residuals for all couples is shown in Figure

C.6. The residuals should have standard normal distribution, and in this case there are more slightly negative residuals and somewhat less positive and larger in absolute value residuals. There are no small set of influential observations, but larger groups which are then not considered outliers.

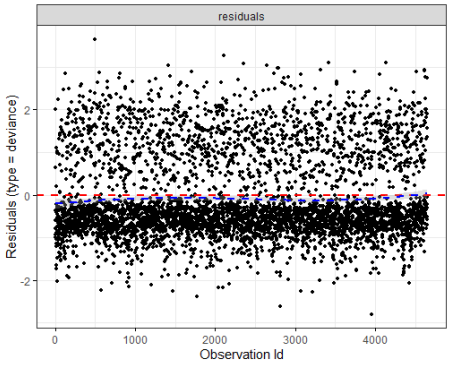


Figure C.6: Deviance residuals of the multi-state model on the dataset used in this study

C.3. Comparison of model predictions and Kaplan-Meier estimates

Then, the empirical analysis of the model is conducted. There are many ways one can compare the model estimates with non-parametric ones from the population. Here the survival probability for each transition is used. The model estimates of survival probability are computed for each couple in the dataset and then averaged over all couples. The survival curves can be seen in Figure C.7 on the left. The non-parametric estimate of the survival probabilities is done with a Kaplan-Meier estimate, and can be seen in Figure C.7 on the right. Survival curve for Transition 2 (EM→ pregnancy) almost completely overlaps in parametric and non-parametric estimate. This may be attributed to having lots of data to estimate this curve. However, curves for survival probabilities for Transitions 1 (EM→ IUI) and 3 (IUI→ pregnancy) are completely different in the two plots. The model largely overestimates the probability of these two transitions happens and thus survival curves soon become really low, much lower than in the non-parametric estimate of the same indicate. This can be attributed to having less data available to model these two transitions and thus lots of couples having unrealistically low estimated survival probabilities for these two transitions.

The model estimate is computed through *survfit()* function applied on a *coxph* object, representing the fitted multi-state model. On the other hand, the non-parametric estimate is computed through the *survfit()* applied on a *Surv()* object, therefore not taking the couples' covariates into account and not specifying any model to underlie the data.

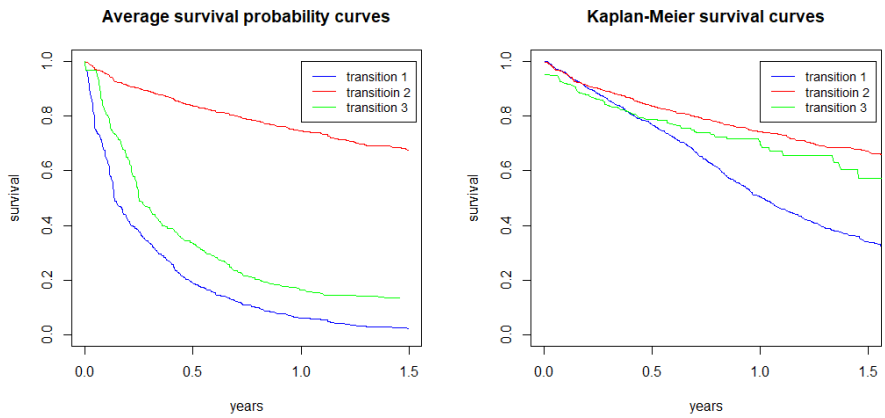


Figure C.7: Multi-state transitions survival probabilities on the left and Kaplan-Meier estimator of the survival probabilities, fitted on the typical couple and using the clock-reset time scale

D

Code

```
library(survival)
library(mstate)
library(plyr)
library(pracma)

#load the data and prepare the dataset (datasets 'long' and 'wide' confidential)
MS <- wide
MS <- rename(MS,c("vrouleef_newadj"="age","duursubf_newadj"="dur","infvrouw1"="infert",
                  "sa1prog"="psperm","verwijz3"="refer","tubapat1_t0"="vtd"))

MS$timePreg <- MS$stop2 # time of pregnancy
MS$statPreg <- MS$Z1 + ifelse(is.na(MS$Z2), 0, MS$Z2) # status of pregnancy
for(i in 1:length(MS$stop1)) if( is.na(MS$timePreg[i]) ) MS$timePreg[i] <- MS$stop1[i]

treat <- rep(0,1896)
for(i in 1:length(treat)) if(!is.na(MS$start2[i])) treat[i] <- 1
treat
table(treat)
MS <- cbind(MS, treat)

mstime = cbind( rep(NA,length(MS[,1])), MS$stop1, MS$timePreg)
msstatus = cbind( rep(NA,length(MS[,1])), MS$treat, MS$statPreg)

tmat <- trans.illdeath( names=c("exp.m.", "IUI", "preg")) #transition matrix
tmat
paths(tmat)


covariates <- c("age","dur","infert","refer","psperm","vtd","stop1") #covariates in the model

MSprep <- msprep(time=mstime, status=msstatus, data=MS, trans=tmat, keep=covariates)
head(MSprep)
events(MSprep)
MSdata <- expand.covs(MSprep, covariates, append = TRUE, longnames = FALSE)
head(MSdata)

#take logarithm of subfertility duration in order to satisfy linearity assumption
MSdata$dur.1 <- ifelse(MSdata$dur.1>0, log(MSdata$dur.1)+1,0)
head(MSdata)


#clock-reset model fitted on the dataset
c1 <- coxph(Surv(time, status) ~ age.1 + dur.1 + infert.1 +  refer.1 + psperm.1 + vtd1.1 + vtd2.1 +
              age.2 + dur.2 + infert.2 + refer.2 + psperm.2 + vtd1.2 + vtd2.2 +
              age.3 + dur.3 + infert.3 + refer.3 + psperm.3 + vtd1.3 + vtd2.3 +
              strata(trans) + stop1.3, data=MSdata, method="breslow")

baseline <- basehaz(c1, centered=FALSE) # baseline cumulative hazard
strata12 <- sum(baseline[,3]=="trans=1") # number of events for transitions 1 and 2
strata3 <- sum(baseline[,3]=="trans=3") # number of events for transition 3
```

```
51
52 baseline1 <- baseline[1:strata12,1]
53 baseline2 <- baseline[(strata12+1):(2*strata12),1]
54 baseline3 <- baseline[(2*strata12+1):(2*strata12+strata3),1]
55
56 #preparing the 'couple' dataframe
57 couple <- data.frame(age=rep(0,3), dur=rep(0,3), infert=rep(0,3),
58                      refer=rep(0,3), psperm=rep(0,3), vtd=rep(0,3),
59                      stop1=rep(0,3), trans=1:3)
60 couple$infert <- factor(couple$infert, levels=0:1, labels=levels(MS$infert))
61 couple$refer <- factor(couple$refer, levels=0:1, labels=levels(MS$refer))
62 couple$vtd <- factor(couple$vtd, levels=0:2, labels=levels(MS$vtd))
63
64 #number of couples analysed
65 n<-length(MS$sleuteli)
66 n
67
68 #time vectors
69 l1 <- baseline[1:strata12,2] # event times for transitions 1 and 2
70 l1 <- c(0,l1) # add 0 in the beginning
71 for(i in 1:length(l1)) if(l1[i]<1.5) k1<-i # k1 the last event time before 1.5 years
72 l2 <- baseline[(2*strata12+1):(2*strata12+strata3),2] # event times for transition 3
73 l2 <- c(0,l2) # add 0 in the beginning
74 for(i in 1:length(l2)) if(l2[i]<1.5) k2<-i # k2 the last event time before 1.5 years
75
76 #initialize
77 notreatment <- 0
78 treatment <- list(0, 0, 0, 0)
79 EMpreg <- list(0, 0, 0)
80
81 for(j in 1:n){
82
83   #couple
84   couple <- subset(couple,select=1:8)
85   couple$age <- MS$age[j]
86   couple$dur <- MS$dur[j]
87   couple$psperm <- MS$psperm[j]
88   couple$stop1[3] <- 0
89   couple$infert <- MS$infert[j]
90   couple$refer <- MS$refer[j]
91   couple$vtd <- MS$vtd[j]
92   attr(couple, "trans") <- tmat
93   class(couple) <- c("msdata", "data.frame")
94   couple <- expand.covs(couple, covariates, longnames = FALSE)
95   couple$strata = 1:3
96
97   ## no treatment
98
99   #cumulative hazards
100  Lam13 <- baseline2*exp(sum(c1$coefficients[8:14]*couple[2,c(10,13,16,19,22,25,28)]))
101  Lam13 <- Lam13[1:k1]
102
103  #survival probabilities
104  S1 <- rep(1,k1)
105  for(i in 1:k1) S1[i] <- exp(- Lam13[i])
106
107  #trasition probabilities
108  P113 <- rep(0,k1)
109  P113[1] <- Lam13[1]
110  for(i in 2:k1) P113[i] <- P113[i-1] + (Lam13[i]-Lam13[i-1])*S1[i-1]
111
112  #pregnancy probability after 1.5 years
113  notreatment[j] <- P113[k1]
114
115
116  ## treatment strategies
117
118  for(b in 1:4){
119
120    a <- (b-1)/4 # treatment at time (b-1)/4 years
121    couple$stop1.3[3] <- a
```

```
    for(i in 1:k1) if(l1[i]<=a) r1 <- i # r1 index of the last time before treatment starts
    for(i in 1:k2) if(l2[i]<=1.5-a) r2 <- i # r2 index of the last time before 1.5 years in clock-reset

    Lam23 <- baseline3*exp(sum(c1$coefficients[15:22]*couple[3,c(11,14,17,20,23,26,29,32)]))
    Lam23 <- Lam23[1:k2]

    S2 <- rep(1,k2)
    for(i in 1:k2) S2[i] <- exp(- Lam23[i])

    P113a <- ifelse(r1>1, P113[r1], 0) # probability of pregnancy before the time of the treatment
    P123 <- 0
    P123[1] <- Lam23[1]
    for(i in 2:r2) P123 <- P123 + (Lam23[i]-Lam23[i-1])*S2[i-1] # probability of pregnancy after the
        treatment
    P123 <- P123*S1[r1] + P113a # cumulative probability of pregnancy

    treatment[[b]][j] <- P123
    if(b>1) EMpreg[[b-1]][j] <- P113a

  }
}

#results: average of all couples' outcomes
mean(notreatment)
for(i in 1:4) print(mean(treatment[[i]]))
for(i in 1:3) print(1-mean(EMpreg[[i]]))
```

Listing D.1: Code of estimating the pregnancy probabilities after 1.5 years in the followup and the probabilities of natural pregnancies before start of the treatment in the population as described in Section 5.3

```
#bootstrap procedure
#continuing on the workspace from the previous code for the main estimates

#binary search function to speed up the computation
bin.search <- function(x,l,u){
  y <- floor((u+l)/2)
  z <- MSdata$id[y]
  if(x==z) return(y)
  if(x>z) return(bin.search(x,y+1,u))
  return(bin.search(x,l,y-1))
}

#initialize
p <- 200 # number of bootstrap samples
set.seed(100) # set seed from which the other p seed are sampled
randseed <- sample(1:n, p, replace=FALSE) # random seeds to sample the bootstrap samples from

m <- length(MSdata$id)

bootnotreat <- rep(0,p)
boottreat <- list(rep(0,p), rep(0,p), rep(0,p), rep(0,p))
bootEMpreg <- list(rep(0,p), rep(0,p), rep(0,p))


for(o in 1:p){

  set.seed(randseed[o]) # set seed for the o-th sample
  randvec <- sample(1:n, n, replace=TRUE) # sample n couples' id numbers with replacement
  randvec <- sort(randvec)
  table <- as.data.frame(table(randvec))

  #create the dataset with the couples with id in randvec
  MSdata$unique <- rep(0,m)
  MSample <- MSdata[FALSE,]
  for(i in 1:length(unique(randvec))){
    ind <- as.integer(as.vector(table[i,1]))
    rep <- as.vector(table[i,2])
    q <- bin.search(ind,1,m) #find couples faster
    first <- 0 #marks the first mention of a couple
    #include all rows (max. 3) with this id
    for(j in 1:5){
```

```
      qq <- q-3+j
      if(qq>= 1 & qq <= m) if(MSdata$id[qq]==ind) for(k in 1:rep){
        MSample <- rbind(MSample,MSdata[qq,])
        if(first==0){
          MSample$unique[length(MSample$id)] <- rep
          first <- 1
        }
      }
    }
  }
  len <- length(MSample$id)
  MSdata$unique <- NULL


  #fit the clock-reset model
  c1 <- coxph(Surv(time, status) ~ age.1 + dur.1 + infert.1 +  refer.1 + psperm.1 + vtd1.1 + vtd2.1 +
                age.2 + dur.2 + infert.2 + refer.2 + psperm.2 + vtd1.2 + vtd2.2 +
                age.3 + dur.3 + infert.3 + refer.3 + psperm.3 + vtd1.3 + vtd2.3 +
                strata(trans) + stop1.3, data=MSample, method="breslow")

  #baseline cumulative hazards
  baseline <- basehaz(c1, centered=FALSE)
  strata12 <- sum(baseline[,3]=="trans=1")
  strata3 <- sum(baseline[,3]=="trans=3")

  baseline1 <- baseline[1:strata12,1]
  baseline2 <- baseline[(strata12+1):(2*strata12),1]
  baseline3 <- baseline[(2*strata12+1):(2*strata12+strata3),1]

  #time vectors
  l1 <- baseline[1:strata12,2]
  l1 <- c(0,l1)
  for(i in 1:length(l1)) if(l1[i]<=1.5) k1<-i
  l2 <- baseline[(2*strata12+1):(2*strata12+strata3),2]
  l2 <- c(0,l2)
  for(i in 1:length(l2)) if(l2[i]<=1.5) k2<-i

  #remove all but one row for each couple
  MSample <- MSample[MSample$unique>0,]
  len <- length(MSample$id)

  #initialize
  notreatment <- rep(0,len)
  treatment <- list(rep(0,len), rep(0,len), rep(0,len), rep(0,len))
  EMpreg <- list(rep(0,len), rep(0,len), rep(0,len))

  for(j in 1:len){

    #couple
    couple <- subset(couple,select=1:8)
    couple$age <- MSample$age[j]
    couple$dur <- MSample$dur[j]
    couple$psperm <- MSample$psperm[j]
    couple$stop1 <- 0
    couple$infert <- MSample$infert[j]
    couple$refer <- MSample$refer[j]
    couple$vtd <- MSample$vtd[j]
    attr(couple, "trans") <- tmat
    class(couple) <- c("msdata", "data.frame")
    couple <- expand.covs(couple, covariates, longnames = FALSE)
    couple$strata = 1:3

    ## no treatment

    #cumulative hazards
    Laml3 <- baseline2*exp(sum(c1$coefficients[8:14]*couple[2,c(10,13,16,19,22,25,28)]))
    Laml3 <- Laml3[1:k1]

    #survival probabilities
    S1 <- rep(0,k1)
    for(i in 1:k1) S1[i] <- exp(- Laml3[i])
```

```
113
114     #trasition probabilities
115     P113 <- rep(0,k1)
116     P113[1] <- Lam13[1]
117     for(i in 2:k1) P113[i] <- P113[i-1] + (Lam13[i]-Lam13[i-1])*S1[i-1]
118
119     #update the average pregnancy probability after 1.5 years
120     notreatment[j] <- P113[k1]
121
122     ## treatment strategies
123
124     for(b in 1:4){
125
126       a <- (b-1)/4 # treatment at time (b-1)/4 years
127       couple$stop1.3[3] <- a
128       for(i in 1:k1) if(l1[i]<=a) r1 <- i # r1 index of the last time before treatment starts
129       for(i in 1:k2) if(l2[i]<=1.5-a) r2 <- i # r2 index of the last time before 1.5 years in clock-reset
130
131       Lam23 <- baseline3*exp(sum(c1$coefficients[15:22]*couple[3,c(11,14,17,20,23,26,29,32)]))
132       Lam23 <- Lam23[1:k2]
133
134       S2 <- rep(0,k2)
135       for(i in 1:k2) S2[i] <- exp(- Lam23[i])
136
137       P113a <- ifelse(r1>1, P113[r1], 0) # probability of pregnancy before the time of the treatment
138       P123 <- rep(0,r2)
139       P123[1] <- Lam23[1]
140       for(i in 2:r2) P123[i] <- P123[i-1] + (Lam23[i]-Lam23[i-1])*S2[i-1] # probability of pregnancy after
              the treatment
141       P123 <- P123*S1[r1] + P113a # cumulative probability of pregnancy
142
143       treatment[[b]][j] <- P123[r2]
144       if(b>1) EMpreg[[b-1]][j] <- P113a
145     }
146   }
147
148   bootnotreat[o] <- weighted.mean(notreatment,MSample$unique)
149   for(b in 1:4) boottreat[[b]][o] <- weighted.mean(treatment[[b]],MSample$unique)
150   for(b in 1:3) bootEMpreg[[b]][o] <- weighted.mean(EMpreg[[b]],MSample$unique)
151
152   if(o%%20==0) print(o)
153
154 }
155
156 # sort the bootstrap estimates
157 bootnotreat_sort <- sort(bootnotreat)
158 boottreat_sort <- boottreat
159 for(b in 1:4) boottreat_sort[[b]] <- sort(boottreat[[b]])
160 bootEMpreg_sort <- bootEMpreg
161 for(b in 1:3) bootEMpreg_sort[[b]] <- sort(bootEMpreg[[b]])
162
163 # lower and upper 95% confidence bounds
164 print(c(bootnotreat_sort[floor(0.025*p)], bootnotreat_sort[ceiling(0.975*p)])) # no treatment
165 for(b in 1:4) print(c(boottreat_sort[[b]][floor(0.025*p)],
166                       boottreat_sort[[b]][ceiling(0.975*p)])) # treatment
167 for(b in 1:3) print(c(1-bootEMpreg_sort[[b]][ceiling(0.975*p)],
168                       1-bootEMpreg_sort[[b]][floor(0.025*p)])) # natural pregnancy
```

Listing D.2: Code of bootstrap procedure and computation of the 95% confidence intervals for the main estimates of the pregnancy probability after 1.5 years in the followup and the probabilities of natural pregnancies before start of the treatment in the population as described in Section 5.3

Bibliography

[1] Ørnulf Borgan. *Aalen–Johansen Estimator*. 07 2005. ISBN 9780470011812. doi: 10.1002/0470011815.b2a11001.

[2] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[3] Inge M Custers, Pieternel Steures, Jan Willem van der Steeg, Thierry JHM van Dessel, Rob E Bernardus, Petra Bourdrez, Carolien AM Koks, Wernard J Riedijk, Jan M Burggraaff, Fulco van der Veen, et al. External validation of a prediction model for an ongoing pregnancy after intrauterine insemination. *Fertility and sterility*, 88(2):425–431, 2007.

[4] Inge M Custers, Minouche ME Van Rumste, Jan Willem Van Der Steeg, Madelon Van Wely, Peter GA Hompes, Patrick Bossuyt, Frank J Broekmans, Cees NM Renckens, Marinus JC Eijkemans, Thierry JHM Van Dessel, et al. Long-term outcome in couples with unexplained subfertility and an intermediate prognosis initially randomized between expectant management and immediate treatment. *Human reproduction*, 27(2):444–450, 2012.

[5] A Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189, 2002.

[6] Bianca L De Stavola, Rhian M Daniel, George B Ploubidis, and Nadia Micali. Mediation analysis with intermediate confounding: structural equation modeling viewed through the causal inference lens. *American journal of epidemiology*, 181(1):64–80, 2015.

[7] MJC Eijkemans, FAM Kersten, AME Lintsen, CC Hunault, CAM Bouwmans, L Roijen, JDF Habbema, and DDM Braat. Cost-effectiveness of 'immediate ivf'versus 'delayed ivf': a prospective study. *Human Reproduction*, 32(5):999–1008, 2017.

[8] Jacob Farhi and Raoul Orvieto. Cumulative clinical pregnancy rates after coh and iui in subfertile couples. *Gynecological Endocrinology*, 26(7):500–504, 2010. doi: 10.3109/09513590903367036.

[9] Cynthia M Farquhar, Emily Liu, Sarah Armstrong, Nicola Arroll, Sarah Lensen, and Julie Brown. Intrauterine insemination with ovarian stimulation versus expectant management for unexplained infertility (tui): a pragmatic, open-label, randomised, controlled, two-centre trial. *The Lancet*, 391(10119):441 – 450, 2018. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(17)32406-6. URL http://www.sciencedirect.com/science/article/pii/S0140673617324066.

[10] Jon Michael Gran, Stein Atle Lie, Irene Øyeflaten, Ørnulf Borgan, and Odd O Aalen. Causal inference in multi-state models–sickness absence and work for 1145 participants after work rehabilitation. *BMC Public Health*, 15(1):1–16, 2015.

[11] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[12] David W Hosmer Jr, Stanley Lemeshow, and Susanne May. *Applied survival analysis: regression modeling of time-to-event data*, volume 618. John Wiley & Sons, 2011.

[13] Gideon Jager. Survival analysis and its applications in the evaluation of treatment timing, 2019.

[14] E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[15] Jinyong Liu, Tin-Chiu Li, Jing Wang, Wei Wang, Zhen Hou, and Jiayin Liu. The impact of ovarian stimulation on the outcome of intrauterine insemination treatment: an analysis of 8893 cycles. *BJOG: An International Journal of Obstetrics & Gynaecology*, 123:70–75, 2016.

[16] DJ McLernon, AJ Lee, A Maheshwari, R Van Eekelen, Nan van Geloven, Hein Putter, MJ Eijkemans, Jan Willem van der Steeg, F Van Der Veen, EW Steyerberg, et al. Predicting the chances of having a baby with or without treatment at different time points in couples with unexplained subfertility. *Human Reproduction*, 34(6):1126–1138, 2019.

[17] Grenier Lourdel Sanguinet Copin Merviel, Heraud. Predictive factors for pregnancy after intrauterine insemination (iui): An analysis of 1038 cycles and a review of the literature. *Fertility and Sterility*, 93(1): 79–88, 2010.

[18] Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

[19] Willem Ombelet. Evidence-based recommendations for iui in daily practice. *Middle East Fertility Society Journal*, 18(2):74 – 77, 2013.

[20] Judea Pearl. [bayesian analysis in expert systems]: Comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

[21] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[22] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 2000.

[23] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[24] H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430, 2007. doi: 10.1002/sim.2712. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2712.

[25] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393 – 1512, 1986. ISSN 0270-0255.

[26] 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, 04 1983.

[27] Kenneth J. Rothman. Causes. *American Journal of Epidemiology*, 104(6):587–592, 12 1976.

[28] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[29] Enrique F Schisterman, Stephen R Cole, and Robert W Platt. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass.)*, 20(4):488, 2009.

[30] Weichung Joseph Shih. Problems in dealing with missing data and informative censoring in clinical trials. *Trials*, 3, 2002.

[31] Peter Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11(5), 2010.

[32] Pieternel Steures, Jan Willem van der Steeg, Peter GA Hompes, J Dik F Habbema, Marinus JC Eijkemans, Frank J Broekmans, Harold R Verhoeve, Patrick MM Bossuyt, Fulco van der Veen, and Ben WJ Mol. Intrauterine insemination with controlled ovarian hyperstimulation versus expectant management for couples with unexplained subfertility and an intermediate prognosis: a randomised clinical trial. *The Lancet*, 368(9531):216 – 221, 2006.

[33] Patrick Suppes. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Pub. Co., 1970.

[34] Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313, 2000.

[35] R Van Eekelen, I Scholten, RI Tjon-Kon-Fat, Jan Willem van der Steeg, Pieternel Steures, Peter Hompes, Madelon van Wely, F Van Der Veen, BW Mol, MJ Eijkemans, et al. Natural conception: repeated predictions over time. *Human Reproduction*, 32(2):346–353, 2017.

[36] R van Eekelen, N van Geloven, M van Wely, D J McLernon, F Mol, I M Custers, P Steures, S Bhattacharya, B W Mol, F van der Veen, and M J Eijkemans. Is IUI with ovarian stimulation effective in couples with unexplained subfertility? *Human Reproduction*, 34(1):84–91, 11 2018.

[37] R van Eekelen, M J Eijkemans, M Mochtar, F Mol, B W Mol, H Groen, and M van Wely. Cost-effectiveness of medically assisted reproduction or expectant management for unexplained subfertility: when to start treatment? *Human Reproduction*, 35(9):2037–2046,, 09 2020.

[38] N van Geloven, RB Geskus, BW Mol, and AH Zwinderman. Correcting for the dependent competing risk of treatment using inverse probability of censoring weighting and copulas in the estimation of natural conception chances. *Statistics in medicine*, 33(26):4671–4680, 2014.

[39] Nan van Geloven, Theodor A Balan, Hein Putter, and Saskia le Cessie. The effect of treatment delay on time-to-recovery in the presence of unobserved heterogeneity. *Biometrical Journal*, 62(4):1012–1024, 2020.

[40] Jessica Wade, Vivien Maclachlan, and Gab Kovacs. The success rate of ivf has significantly improved over the last decade. *The Australian New Zealand journal of obstetrics gynaecology*, 55, 07 2015. doi: 10.1111/ajo.12356.

[41] Liang Wang, Ning xia Sun, Xin mei Lu, Qing Zhang, Chen Xu, Yan Cao, and Wen Li. Effect of timing and number of intrauterine insemination (iui) on the cycle clinical outcome. *Journal of Reproduction and Contraception*, 25(4):219 – 226, 2014. ISSN 1001-7844. doi: https://doi.org/10.7669/j.issn.1001-7844.2014.04.0219.

[42] James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

[43] Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20(7):557–585, 1921.