

Robust Data Mining for Medicine: Prediction of Chemotherapy Induced Amenorrhea and Menses Recovery for Breast Cancer

by

Ruoyi Gan

A thesis submitted in partial fulfillment for the
coursework degree of Master of Data Science

in the
School of Mathematics and Statistics
Faculty of Science
THE UNIVERSITY OF MELBOURNE

October 2022

THE UNIVERSITY OF MELBOURNE

Abstract

School of Mathematics and Statistics

Faculty of Science

Master of Data Science

by [Ruoyi Gan](#)

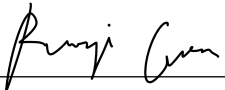
Breast cancer is the most common cancer among women and the most prevalent cancer overall. In 2019, over 30% of the 2 million new instances of invasive breast cancer identified worldwide were among women under the age of 50. One of the probable adverse effects of adjuvant chemotherapy is the suppression of ovarian function, which may lead to vasomotor symptoms, bone loss, cardiovascular illness, or sexual dysfunction (e.g. infertility). Fertility preservation is vital to therapy considerations, particularly for young premenopausal women.

This thesis presents a statistical analysis framework to investigate the influential factors associated with chemotherapy-induced amenorrhea (CIA) for older and younger patients, followed by an analysis of the likelihood of menses resumption of the CIA incident. Approaches to missing data handling has also been compared. The final missing data imputation method is the multivariate imputation by chained equations (MICE) with the random forest based on experiments.

The Cox regression model was used for predictions because of the existence of follow-up records. For the analysis of the risk of CIA, both models of the patients group with age above and below 40 achieved a concordance of approximately 0.7. The statistically significant factors are the patient's age at diagnosis, ER status, whether the cancer is invasive, and cycles of CMF received. Additional to this, for patients aged above 40 years old, the interaction terms of ER status and CMF cycles, and invasiveness of cancer and CMF cycles were also significant. Furthermore, the effect of the increase in age at diagnosis was different for the two sub-groups. Older patients will be more likely to develop CIA as their age grows, while for younger patients, their risk decreases as their age increases. We then performed further analysis on patients who developed CIA to investigate the likelihood of menses resumption, the resulting model yielded a concordance of approximately 0.7 and suggested patient's age at diagnosis and the timestamp of CIA occurrence would be the influential factors associated with the resumption.

Declaration of Authorship

I certify that this report does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text. The report is 10463 words in length (excluding text in images, tables, bibliographies and appendices).

Signed:  _____

Date: 29/10/2022 _____

Acknowledgements

I would like to express my deepest appreciation to my project supervisor Long Song for providing me with the opportunity to work with him on this project and for his invaluable patience and guidance, as well as the generous support throughout the year.

Thanks should also go to the study participants and the organisation that collected the data for this study. I would like to extend my sincere thanks to my friends for their editing help and feedback.

Contents

Abstract	i
Declaration of Authorship	ii
Acknowledgements	iii
List of Figures	vi
List of Tables	vii
Abbreviations	viii
1 Introduction	1
1.1 Background	1
1.2 Research gap	2
1.3 Contributions of this research	2
1.4 Thesis outline	3
2 Literature Review	4
2.1 CIA: Incidence and recovery	5
2.1.1 Incidence of CIA	5
2.1.2 Recovery from CIA	7
2.2 Different statistical analysis approaches	8
2.2.1 Parametric approach	9
2.2.2 Non-parametric approach	10
2.2.3 Semi-parametric approach	10
3 Data Collection	11
3.1 Data set	11
3.2 Data exploration and visualisation	12
3.3 Study populations	14
3.3.1 Part I: Incidence of CIA	14
3.3.2 Part II: Recovery from CIA	15
4 Methodology	16
4.1 Data pre-processing pipeline	16
4.2 Explanatory variables processing	18

4.2.1	Variable pre-selection	18
4.2.1.1	Irrelevant variables	19
4.2.1.2	Redundant variables	19
4.2.1.3	Variables with all identical values	19
4.2.2	Missing data imputation	20
4.2.2.1	K-Nearest Neighbours Algorithm	21
4.2.2.2	Multivariate Imputation by Chained Equations	21
4.3	Outcome variables processing	22
4.3.1	Validation pattern	23
4.3.2	Fixing incorrect records	24
4.3.2.1	Incidence of CIA	24
4.3.2.2	Recovery from CIA	25
4.4	Model fitting	25
4.5	Model evaluation	26
4.6	Model selection	27
5	Results	28
5.1	Risk of the CIA occurrence	28
5.2	Recovery from CIA	32
6	Discussion	33
6.1	Incidence of CIA	33
6.1.1	Patients with age above 40	34
6.1.2	Patients with age below 40	35
6.1.3	Baseline hazards of two groups	36
6.2	Recovery from CIA	37
6.3	Strength	38
6.4	Limitation	39
7	Conclusions and Future Work	41
7.1	Conclusion	41
7.1.1	Importance of missing-data imputation	41
7.1.2	Incidence of CIA for different age groups	42
7.1.3	Resumption of menses after developing CIA	43
7.2	Future Work	43
A	Nomograms	45
A.1	Nomograms of incidence of CIA	45
A.1.1	Patients aged above 40 years	46
A.1.2	Patients aged above 40	47
A.2	Nomogram of menstrual resumption	48
	Bibliography	49

List of Figures

3.1	Distribution of patients' age at diagnosis	12
3.2	Bar plot of age at diagnosis, for younger than 40 years old and older . . .	12
3.3	Bar plot summary of CMF cycles	13
3.4	Box plot of age at diagnosis by number of CMF cycles	13
3.5	Bar plot summary of CMF cycles, by ER status	14
3.6	Bar plot summary of CMF cycles, by not/invasive cancer	14
4.1	Data pre-processing pipeline. The label of the edges explains the number of observations excluded in each step and the corresponding reason. . . .	17
A.1	Partial nomogram of patients aged above 40 years	46
A.2	Partial nomogram of patients aged below 40 years	47
A.3	Partial nomogram of resumption of menses	48

List of Tables

4.1	Explanatory variable name and meaning in the raw data set.	18
4.2	Percentage of missing values for the selected explanatory variables	20
5.1	Patients with age above 40 years: variables and concordance of model fitted on imputed data sets, with different imputation methods.	29
5.2	Patients with age below 40 years: variables and concordance of model fitted on imputed data sets, with different imputation methods.	30
5.3	Resumption of menses for patients who developed CIA: variables and concordance of model fitted on imputed data sets, with different imputation methods.	32
6.1	Patients with age above 40 years: model variable coefficients, standard errors and variable importance.	34
6.2	Patients with age below 40 years: model variable coefficients, standard errors and variable importance.	35
6.3	Patients with age above 40 years: baseline hazards at 3, 6, 9, 12 months. .	36
6.4	Patients with age below 40 years: baseline hazards at 3, 6, 9, 12 months. .	36
6.5	Resumption of menses: model variable coefficients, standard errors and variable importance.	37
6.6	Resumption of menses: baseline hazards (likelihood) at different timestamps in 5 years.	38

Abbreviations

CIA	C hemotherapy I nduced A menorrhea
CMF	C yclophosphamide, M ethotrexate, F luorouracil treatment
ER	E estrogen R eceptors
k-NN	k Nearest N eighbours
MICE	M ultiple I mputation by C hained E quations

Chapter 1

Introduction

Within the broad field of research on breast cancer, there are several branches that focus on different and more specific topics. For example, the comparison between different therapies in terms of the success rate of treatment, the overall survival rate and disease-free survival rate of rehabilitated breast cancer patients and the rate of incidence of chemotherapy-induced amenorrhea (CIA) and the associated influential factors. This thesis focuses on the analysis of the incident rate of the CIA, and the likelihood of menstrual resumption after the CIA occurred.

1.1 Background

Although chemotherapy has been shown to reduce the risk of disease recurrence and improve the overall survival rate, the side effects may lower the patient's quality of life. Patients are interested in their risk of early menopause, which is one of the signals of infertility[1]. An informative pre-treatment discussion between physician and patient is crucial. It should include the potential adverse effects of the treatment, the risk of developing such effects in a specific patient, and the possibility of recovery or not. With this information, it is possible to benefit doctors in educating their patients on what to expect during therapy and the potential long-term consequences and help the patient construct a more carefully considered and mutually satisfactory treatment plan. As more women survive breast cancer, treatment-related morbidity and the long-term repercussions of CIA will gain importance. Therefore, it is vital to identify the key

factors associated with the risk of developing CIA and the likelihood of recovery and maintenance of healthy menstrual cycles.

1.2 Research gap

Many breast cancer patients, especially the younger patients are interested in their risk of early menopause after treatment, and the risk of infertility[2]. However, the likelihood and duration of menopause after receiving chemotherapy treatment has not been sufficiently explored in the cancer survivor community. This is a missing piece that needs to be filled in urgently for the patient's better family planning and decision-making on the treatment. Furthermore, while there have been many studies on the survival rate of breast cancer patients and several on the danger of CIA occurrences, few studies have examined the likelihood of regaining normal ovarian function. Even less research examines the distinct menstrual cycles of women of various ages or the accompanying risk of CIA at various points during the treatment.

From the statistical perspective, although the popular statistical tool, the logistic regression model, used in this area of research produced many reasonable and valuable results, it is still limited in statistical power compared to the Cox Proportional Hazard Regression model if there are follow-up records available. The lack of external or internal validation of the resultant model decreases the strength of the result and may incur over-fitting issues with a particular data set, which is possible to obtain misleading results from this defective approach, correction and improvement have to be done. This research will maximise the use of information on the follow-up records, implementing a suitable missing data imputation method that can maintain as much sample as possible and perform internal validation.

1.3 Contributions of this research

In summary, the main contributions of this research are as following:

1. Mined the statistical information in the follow-up records by using the Cox regression model, which considers the event status and event occurrence time simultaneously.

2. Assessed the quality of different missing data imputation methods, identified the most suitable one to be used in this research, and thus provided suggestions on imputing data for general survival analysis.
3. Performed sub-group analysis on the risk of CIA occurrence with selected age (40 years) as a threshold. In this way, the resulting model is more specific to each group of patients with different physical conditions.
4. Explored the influential factors associated with menses recovery after CIA and thus more comprehensive information for pre-treatment consultation could be provided.

1.4 Thesis outline

This study will focus on determining the significant variables associated with the occurrence of chemotherapy-induced amenorrhea and menses resumption among those patients who suffer from CIA. In addition, this study intends to analyse the risk of incidence of CIA for patients with different duration of treatment received and physiological conditions such as age, size of the tumour and the status of estrogen receptor (ER) et cetera.

The first objective of this study is to report the hazards of developing CIA at each three month within one year of treatment start, second is that for women who suffered from CIA, identify the factors associated with the time to return of menses.

The rest of the thesis will be organised as follows: Chapter 2 will provide more detailed background information on the previous study on the incident of CIA for breast cancer patients, followed by a brief comparison of different statistical modelling approaches that are commonly applied in survival analysis. Chapter 3 will introduce the data used for this research and present an exploratory analysis of the data. Chapter 4 describes the methods used for data pre-processing, performed separately on the predictors and the outcomes, and a detailed comparison of missing data imputation methods will be discussed. Chapter 5 will present the results for both missing data imputation and the results of the corresponding models. Chapter 6 will analyse the results obtained, and discuss the strengths and limitations of the research. Chapter 7 will summarise the main findings of this research, followed by a discussion on multiple possible future directions.

Chapter 2

Literature Review

The most prevalent cancer among women and the most prevalent cancer overall is breast cancer. In 2019, around 30% of new cases of invasive breast cancer diagnosed worldwide were among women under the age of 50[3] out of 2 million new cases. The statistics showed a growing trend that around 2.26 million new breast cancer cases were diagnosed among women in 2020. In Australia, even though the estimated number of women diagnosed with breast cancer with age under 50 was approximately 20%[3], it is still a problem that cannot neglect.

It is expected that more women will be diagnosed with early-stage breast cancer at younger ages than ever before due to increasing information, screening and self-awareness. Among the diagnoses, most breast cancers are detected at an early stage. Early detection of breast cancer increases the likelihood of long-term survival, particularly if followed by adjuvant treatment. However, the adverse effects of breast cancer treatments such as chemotherapy, radiation, and hormonal therapies may be short- and long-term.

Most national guidelines recommend adjuvant cytotoxic chemotherapy and hormone therapy for estrogen receptor-positive (ER+) tumours, while these recommendations deviate for modest ER-negative tumours. Consequently, most young women with early-stage breast cancer will require adjuvant chemotherapy.

One of the possible side-effects of the adjuvant chemotherapy is ovarian function suppression which could further induce vasomotor symptoms, bone loss, cardiovascular disease or problems with sexual activity (e.g. infertility). Especially for young premenopausal

women, fertility preservation is crucial in therapeutic decisions. As an increasing number of women with breast cancer survive the illness, treatment-related morbidity and the long-term effects of chemotherapy-induced amenorrhea (CIA) will gain increased significance. Therefore, it is essential to determine the significant factors related to the risk of occurrence of CIA, with the further estimation of the possibility of recovery and maintaining healthy menstrual cycles.

This literature review focuses on the current studies on analysing the incidence of chemotherapy-related amenorrhea in breast cancer patients as in Section 2.1. Further discussion about the statistical analysis method used in terms of the models' power, the criterion for valid observations and usage with different forms of data are in Section 2.2.

While there has been much research on the survival status of breast cancer patients and the risk of CIA incidences, few researchers have taken the probability of resuming normal ovarian function into consideration. Even fewer studies focus on the different menstruation patterns of women of different ages or the corresponding risk of CIA at different timestamps during the therapy.

2.1 CIA: Incidence and recovery

Amenorrhea produced by chemotherapy is a therapy-related side effect, yet also increases breast cancer survival. It has been suggested that the efficacy of dose-dense chemotherapy stems from the higher prevalence of chemotherapy-induced amenorrhea. After adjuvant chemotherapy, the risk of re-occurrence of breast cancer was shown to be reduced by one third, according to the findings of a major meta-analysis conducted by the EBCTCG on individual patient data from 100,000 women[4].

2.1.1 Incidence of CIA

While there is a lacking of uniform definitions of menopause and CIA, several studies defined amenorrhea as experiencing no menstruation for three or more months during the treatment period of chemotherapy, specifically six months as a standard[5]. Generally, study definitions employ shorter spans, such as 90 days, but they vary widely[6]. Other standards such as a negative pregnancy test at the 12-month evaluation point

or the level of follicle-stimulating hormone levels have also been used. Therefore, CIA could thus be determined based on these criteria, given that the subject of interest is currently undertaking chemotherapy and her menstrual functionality was normal before the treatment started.

Menopause may be a potentially adverse effect on younger women who undertake breast cancer treatment. If a woman is peri-menopausal when she begins treatment, it is possible for her to enter menopause sooner compare to not undergoing medication. Others may suffer from transient menopausal symptoms. This depends on the type of therapy and the age of the patient. From the report provided by Cancer Australia in 2013, there were about two-thirds of women diagnosed with breast cancer before age 50 will experience early or premature menopause as a result of their therapy[7]. According to the investigation of Bines et al.[8], 68% is the average CIA rate observed for regimens including cyclophosphamide, methotrexate, and fluorouracil (CMF). The incidence of chemotherapy-induced amenorrhea depends on several factors that contribute to this issue, including a woman's age, weight, lifestyle and the type of treatment. Menopausal may be temporary or permanent.

Many of the studies intend to explore the relationship between CIA and the overall and disease-free survival rate for breast cancer patients, less research focused on the analysis of factors associated with the occurrence of CIA.

Turnbull et al.[9] got the conclusion that age at the time of treatment and at menarche was connected with the risk of amenorrhea in 107 premenopausal women. The researchers were interested in discovering the component that was associated with the incidence of CIA. In addition, Liem and their colleagues[10] concur with this finding after doing an investigation on 286 patients to corroborate their claims. Both studies reported that their findings, when taken together, are compatible with those of earlier research. These investigations did not validate the resulting model in any way, either internally or externally, which is one of the shortcomings of these investigations. By analysing the performance of their model with the Leave-One-Out test, as well as 10 percent, 20 percent, and 30 percent Cross-Validation tests, as well as a back substitution check and sensitivity test, Zhang et al.[11] provided vital support to the statement that the patient's age is a significant factor.

Another model that is frequently utilised in these areas of research is known as the Cox proportional hazard model. This model is a semi-parametric technique that deals with data with censoring time in a more suitable manner. According to the findings of this study, which were derived from data collected by the International Breast Cancer Study Group Trails V and VI and which included participants ranging in age from 22 to 57 years old, menopause is more likely to occur in older women during or immediately after treatment for breast cancer[12]. At the same time, younger patients have a higher risk of going through the menopause years earlier than their peers. This discovery is essential for the use of long-term follow-up information on menstruation. However, the findings might have been more applicable if the analyses had been carried out separately with the groups of younger patients and older patients, rather than only focusing on the duration of CMF while ignoring the inconsistencies in the physiology and hormone levels of women at different ages. This would have been the case if the analyses had been carried out separately with the groups of younger patients and older patients. Lee et al.[13] utilised this sub-setting method by dividing patients into groups of age below and above 40 years and found that the older group were more likely to have permanent CIA compared with women younger than 40 years of age. This was discovered by dividing patients into groups of age below and above 40 years. The model's goodness of fit may be improved, leading to increased accuracy, by subdividing all the data into a greater number of smaller groups using objective criteria. Nevertheless, the identical problem arises in this case since there is no outcome of model validation.

To summarise, most of studies were conducted either with limited data, lack of long-term follow-up or poor categorisation of patients into different age groups or failed to present an appropriate validation of their results. This may then cause the estimation to be less accurate. Moreover, little research focused on the risk of CIA at different timestamps during the chemotherapy treatment.

2.1.2 Recovery from CIA

A study of the literature failed to demonstrate that a subsequent pregnancy increased the chance of recurrence and death in breast cancer survivors, while Gadducci et al. have identified prolonged survival in patients who became pregnant following breast cancer therapy[14]. On the contrary, Park et al. found that recovery of normal menstrual cycles

was significantly associated with shorter disease free survival in both HR-positive and HR-negative patients[15]. This finding is further agreed by Swain and his colleagues [16] with the conclusion that ovarian function recovery is associated with shorter disease-free survival of breast cancer patients. Other studies[17–19] have shown no impact of amenorrhea on survival. Since there is currently no solid evidence demonstrating that women with a history of breast cancer may have adverse consequences from future pregnancies, furthermore, early ovarian failure may be linked to increased morbidity and death.

Rosenberg et al. makes the very valid points that in addition to the loss of fertility, menopausal symptoms include night sweats, hot flashes, vaginal dryness, and weight gain. These symptoms can be incredibly upsetting for young women and have a negative impact on both health-related and psychosocial aspects of life[20]. Therefore, in a population with a high possibility of long-term survival, it is prudent to undertake measures to maintain fertility. These potential risks and issues need to be discussed carefully before the chemotherapy, and patients should have a psychological expectation of the side effects of treatment.

However, there are limited studies that intend to analyse the possible factors associated with the recovery of normal menstruation after the occurrence of CIA. Patients should not only be informed of their risk of chemotherapy-induced amenorrhea but also of the possibility of the ease of this heterogeneous effect after the treatment is finished.

2.2 Different statistical analysis approaches

In clinical trials and observational studies, the process of identifying and correcting prognostic factors are essential components of the statistical analysis. Comparisons of various treatments must always include an adjustment for relevant prognostic factors to be considered valid. It is possible for there to be estimation errors of treatment differences if important prognostic variables are neglected to be taken into consideration or if there is an insufficient use of the information that is offered by the data, particularly in observational research. Incorrect modelling of prognostic factors might also impede the discovery of nonlinear trends or threshold effects on survival, which is a potential problem.

The current research makes use of a wide variety of statistical models to identify potential influential factors that may lead to amenorrhea. The researchers also seek to predict the probability of incidence of patients based on the elements that were deemed important. Statistics may be broken down into three categories: parametric, semi-parametric, and non-parametric. These categories correspond, respectively, to well-known models such as logistic regression, Cox proportional hazard regression, and the Kaplan–Meier estimator.

This section introduces the benefits and disadvantages of different approaches, with a more in-depth discussion about the difference between them and possibly different results therefore.

2.2.1 Parametric approach

The most popular model used in this research area is the logistic regression model, which is parametric because of the finite set of parameters. Specifically, the parameters are the regression coefficients. The logistic function is defined as follow:

$$\log\left(\frac{p}{1-p}\right) = \eta = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p$$

where η is the linear predictor, p denotes the probability, and β_i is the size of influence or coefficient of the corresponding factor.

A significant benefit of logistic regression is that the exponentiated logistic regression slope coefficient (e^β) may be viewed as an odds ratio, reflecting how much the odds of a particular outcome change for a 1-unit increase in the independent variable (for continuous independent variables) or versus a reference category (for categorical variables)[21]. Studies that utilised this model followed the approach that using univariate logistic regression to select the potential factors, then perform multivariate logistic regression to determine the significance variables. Concern should be pointed out here is whether this approach is appropriate since the typical statistical analysis approach does not exclude any potential factors before fitting the model.

It is worth noting that another sub-family of the parametric approach is deep learning techniques. However, as Chen et al.[22] pointed out in their study, most the clinical statistics are a patchwork of highly specialised procedures and knowledge bases, in contrast to cognitive applications such as driving or image search. Therefore, the ambition

to develop huge, thorough models utilising large, complete patient databases without understanding the eventual use case might result in a subpar performance in practice.

2.2.2 Non-parametric approach

A commonly used non-parametric model in medical statistics is the Kaplan-Meier method, which is used to perform time-to-event statistical analysis especially the survival chance in cancer surgery. A disadvantage of the KM technique is that the log-rank test is just a significance test and thus cannot offer an estimate of the magnitude of the difference between the groups and its corresponding confidence interval. A further disadvantage of the KM technique is that it only delivers unadjusted probability of mortality (and survival)[23]. Therefore, rather than using the Kaplan-Meier model throughout the whole analysis, it could be included as a tool for investigating different survival patterns (survival curves) across different groups of factors of interest, such as treatment type or age.

2.2.3 Semi-parametric approach

Cox Proportional Hazard Regression, a semi-parametric model with an undetermined baseline hazard function, might also be utilised in this field of study as an alternative statistical model. The hazard function $h(t)$ can be estimated as follow:

$$h(t) = h_0(t) \times e^{\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}$$

where t represents the survival time, $h_0(t)$ is called the baseline hazard and corresponds to the value of the hazard if all the factors (x_i) are equal to zero. Since both $h(t)$ and $h_0(t)$ involves the survival time t , this reminds us that the hazard may vary over time.

There is no assumption regarding the distribution of the outcome in the Cox PH regression. Since Cox proportional hazards models take into consideration the time until events occur, if follow-up data are available, the Cox PH model is regarded to have more statistical power and is hence preferred above the logistic regression model. Although the underlying baseline hazard of the Cox Regression model is undefined, it is possible to approximate this value and thus allows a better interpretation of the results.

Chapter 3

Data Collection

This chapter provides the information about the data used for this research project. Initial data exploration and visualisation are also included.

3.1 Data set

This study uses data provided by the Melbourne Medical School, the University of Melbourne. The data set includes a long-term follow-up on menses information for patients who were diagnosed with breast cancer. The data consists of 767 women treated with either not receiving chemotherapy or treated with one, six or seven cycles of CMF as their chemotherapy regimen, menses information was collected for up to five years.

This data set also contains health-related information for each patient, including their age at diagnosis of breast cancer, invasiveness of cancer, stage of the size of the tumour at the time of diagnosis, identified ER status of the patient at diagnosis and whether pregnancy happened after cancer had recovered (if yes, what is the age of pregnancy was).

The menses information was recorded as normal and regular, scanty, pregnant or no menses during the follow-up period. In this study, Women with normal and scanty menses will be assumed to have regular menses cycles. The follow-up menses information was recorded from the start of the chemotherapy treatment, separated by every three months within the first year, and every twelve months for two to five years from chemotherapy. Therefore, there were eight records in total for a patient.

3.2 Data exploration and visualisation

Seven hundred and sixty-seven individuals participated in the study. The range of the age at diagnosis was from 22 years old to 57 years old, with a median of 41 years old. Furthermore, the interquartile range (IQR) of the age was (37, 45). The distribution of age at diagnosis is shown in Figure 3.1.

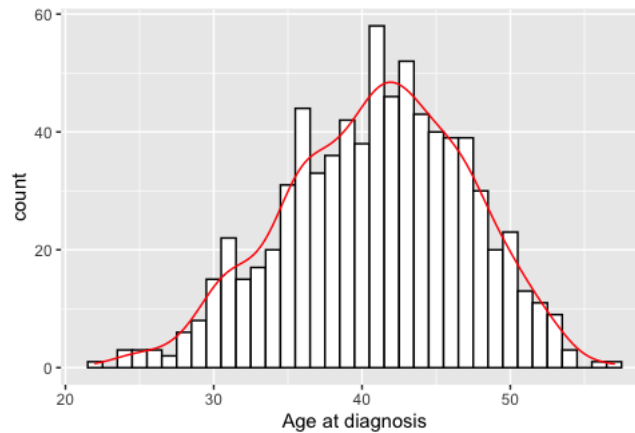


FIGURE 3.1: Distribution of patients' age at diagnosis

By dividing the patients by their median age (41 years old), the sample two groups of patients are approximately equal as shown in Figure 3.2. This provides the basis for this sub-group analysis by maintaining sufficient sample sizes for both groups.

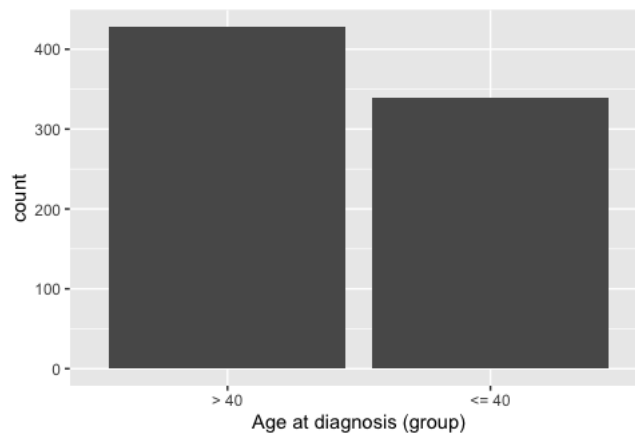


FIGURE 3.2: Bar plot of age at diagnosis, for younger than 40 years old and older

Another explanatory variable, the chemotherapy (CMF) cycles that patients received is shown in Figure 3.3. Patients were treated with none, one, six or seven cycles of chemotherapy. The majority of the subjects received one cycle of CMF, while the number

of subjects who received either six or seven CMF cycles was similar. From Figure 3.4, patients who received six or seven cycles of CMF have a lower mean of age at diagnosis than patients who received none or one cycle.

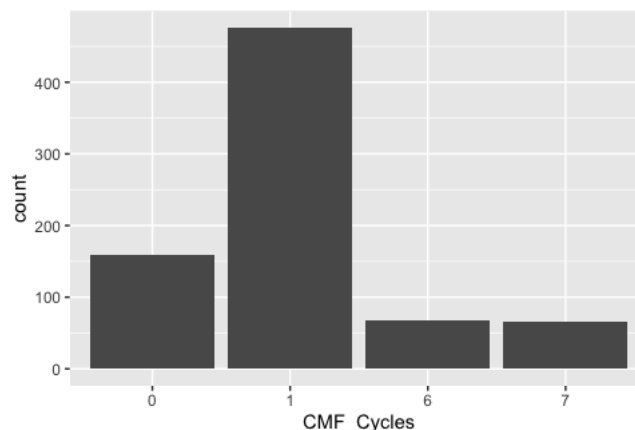


FIGURE 3.3: Bar plot summary of CMF cycles

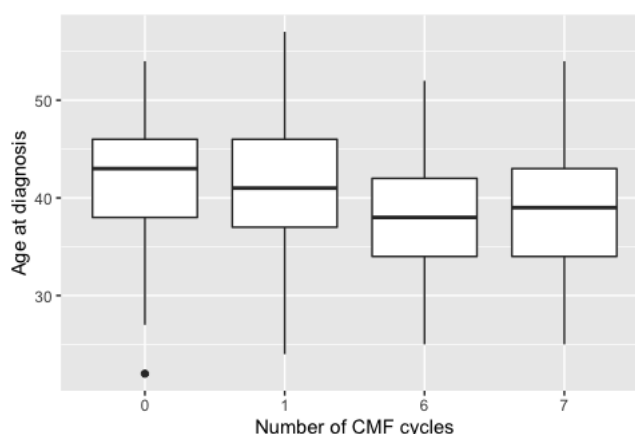


FIGURE 3.4: Box plot of age at diagnosis by number of CMF cycles

Further exploration of the relationship between the number of CMF cycles that a patient received and the ER status or the invasiveness status of the breast cancer has been performed. The ER status is one of the hormone receptor statuses, which is the estrogen receptors (ER). Women with hormone receptor-positive cancers tend to have a better outlook in the short term, but these cancers can sometimes come back many years after treatment. The presence of positive test results indicates that the patients may be suitable for a certain type of treatment. On the other hand, whether the cancer is invasive or not may also influence the decision on the duration of the CMF treatment.

From Figure 3.5 and 3.6, it is possible to see that the most of the subjects were ER-positive, and there is no clear pattern suggesting that the number of CMF cycles is

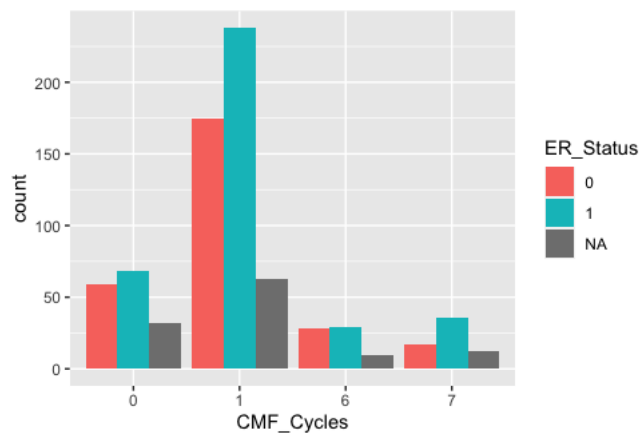


FIGURE 3.5: Bar plot summary of CMF cycles, by ER status

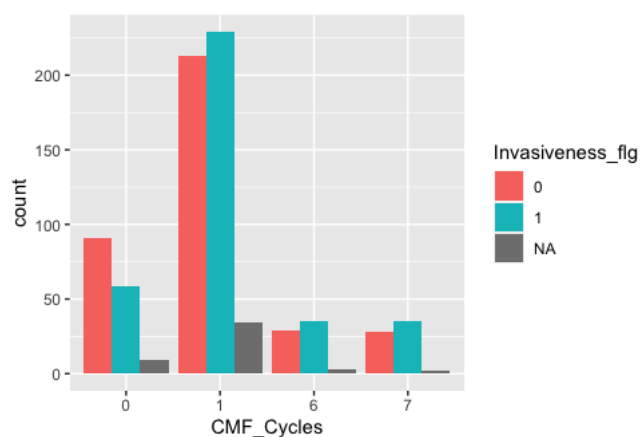


FIGURE 3.6: Bar plot summary of CMF cycles, by not/invasive cancer

related to the ER status. In contrast, subjects without invasive breast cancer were more likely to not receive the chemotherapy treatment. For patients treated with CMF (one, six or seven), the majority had invasive cancer but the difference in the count of the two groups of patients within different cycles was not significant.

3.3 Study populations

3.3.1 Part I: Incidence of CIA

The original data set consisted of 767 participants. To the first part of this study, the valid subjects will be women either in pre-menopausal or peri-menopausal states. According to the royal women's hospital, Victoria Australia, the average age of menopause

is 51[24]. Therefore, 38 subjects that did not meet this condition were excluded, resulting in a sample size of 729 ($N = 729$).

By further identifying participants with available menopause status at baseline and during the two-year follow-up period, there were 140 participants with more than four missing records out of a total of eight who were also excluded from the data set. The final sample size is 589 ($N = 589$).

3.3.2 Part II: Recovery from CIA

In the second part of this study, the purpose is to investigate the factors associated with the resumption of menses after chemotherapy-induced amenorrhea. The subjects of interest were identified with chemotherapy-induced amenorrhea. As mentioned in the previous part of the study, the number of valid participants was 589, among these participants, 99 of them were identified with the occurrence of CIA ($N = 99$).

Chapter 4

Methodology

4.1 Data pre-processing pipeline

The raw data set contains missing values for both the explanatory variables and the outcomes (recorded menstrual status) over the follow-up period.

The data pre-processing needs to be performed separately for the explanatory variables and outcomes separately. For the explanatory variables, missing values imputation was done by both the K-Nearest Neighbours algorithm (k-NN) and Multivariate Imputation by Chained Equations (MICE) via R package `mice`[\[25\]](#). On the other hand, the pre-processing of outcomes (validation and fixing missing values) was performed using the original record for each individual patient as explained below.

Starting with the initial 767 patients in the raw data set, for the first part of the analysis which focuses on analysing the influential variables related to the incidence of CIA, the number of valid observations after the data cleaning is 588 (Figure [4.1](#)). The data cleaning process first filters out the patients who did not receive CMF, then validate the outcomes by pre-defined patterns (records that fluctuate frequently will be viewed as invalid), and lastly, filter out the patients with an age greater than 50 years old as they are highly likely to be in postmenopausal status. These observations were divided further into two subgroups based on their age at diagnosis. One of the groups contains observations that were younger than or equal to 40 years old ($N = 246$), and the other group were patients who were older than 40 years old ($N = 315$). This threshold was suggested by Fornier et al.[\[26\]](#) and Lee et al.[\[13\]](#) in their studies.

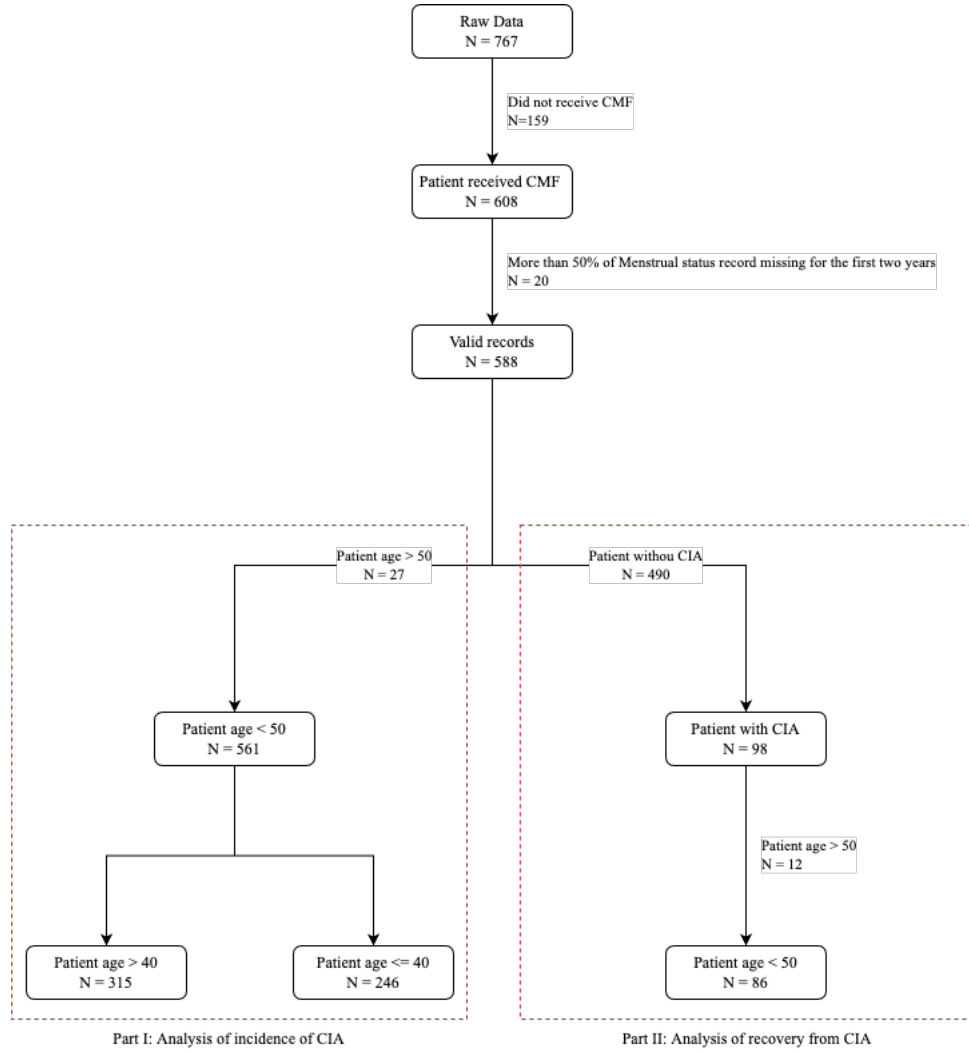


FIGURE 4.1: Data pre-processing pipeline. The label of the edges explains the number of observations excluded in each step and the corresponding reason.

In the second part of the analysis, the aim is to investigate the relationship between the features and the recovery from CIA, thus the subjects of interest are patients who developed CIA during the two-year follow-up period after receiving their first CMF treatment. There were 98 patients who meet these criteria, while 86 of them were aged below 50 years old (Figure 4.1).

Figure 4.1 illustrates the overall pre-processing steps of data pre-processing.

4.2 Explanatory variables processing

The original data set contains 21 variables excluding the outcome variables. The variable names and meanings are summarised in Table 4.1.

Variable name	Meaning	Type	Value
ID	Assigned ID	Numeric	
Data_track	Name of the dataset	Alphabet	
Age	Age, all compiled	Numeric	
Age_diagnosis	Age in years at diagnosis	Numeric	
Agediag_category	Age in years at diagnosis	Categorical	0 = ≥ 40 1 = ≤ 40
Prg1_age	Age at 1st term pregnancy (live birth or gestation > 24 weeks)	Numeric	
Preg_PostCA	Did you get pregnant after breast cancer?	Boolean	0 = No 1 = Yes
Invasiveness_flg	Was cancer invasive	Boolean	0 = No 1 = Yes
Tumor_Size	What stage was the size of tumour at the time of diagnosis	Boolean	0 = ≤ 2 cm 1 = ≥ 2 cm
Nodal_Status	Nodal status at the time of diagnosis	Boolean	0 = Negative 1 = Positive
ER_Status	Identified ER status of patient at diagnosis	Boolean	0 = Negative 1 = Positive
CT_flg	Did you receive chemotherapy	Boolean	0 = No 1 = Yes
Neoadj_CT	Did you receive chemotherapy before surgery	Boolean	0 = No 1 = Yes
Chemoregim	What was the chemotherapy regimen patient treated with	Categorical	Text description
Chemo_Cat	What was the chemotherapy regimen patient treated with	Categorical	1 = Anthracycline and taxane-based; 2 = Anthracycline based only 3 = Taxane-based only; 4 = CMF; 5 = Others
AC	Did you receive adriamycin cyclophosphamide (AC) treatment	Boolean	0 = No 1 = Yes
CMF	Did you receive cyclophosphamide, methotrexate, fluorouracil (CMF) treatment	Boolean	0 = No 1 = Yes
CMF_Cycles	Number of CMF cycles received	Integer	
AC+CMF	Did you receive a combination of AC and CMF treatment	Boolean	0 = No 1 = Yes
AC+T	Did you receive a combination of AC and Taxel treatment	Boolean	0 = No 1 = Yes
Taxel	Did you receive Paclitaxel (Taxel) chemotherapy drug	Boolean	0 = No 1 = Yes

TABLE 4.1: Explanatory variable name and meaning in the raw data set.

4.2.1 Variable pre-selection

Variable pre-selection was done before the missing value imputation, which contains three parts:

- Excluding irrelevant variables.
- Excluding redundant variables.

- Excluding variables with the same value for every observation.

4.2.1.1 Irrelevant variables

The ID of patients and the name of the data set are irrelevant to the analysis, and hence should be excluded. Furthermore, there were only two observations who were pregnant after being diagnosed with breast cancer, which cannot provide sufficient information for analysis.

Therefore, the irrelevant variables are ID, `Data_track`, `Prg1_age` and `Preg_PostCA` and should be excluded.

4.2.1.2 Redundant variables

Variable `Age` and `Age_diagnosis` are identical to each other, there the variable `Age` was excluded. The `Agediag_category` was a grouping of `Age_diagnosis` but with overlapping intervals (both 0 and 1 can contain patients aged 40), thus was excluded.

Variables `CMF`, `CT_flg` and `CMF_Cycles` provide similar meanings. `CMF` and `CT_flg` are identical to each other. A non-zero value of `CMF_Cycles` indicates receiving CMF and a missing value of `CMF_Cycles` means did not receive CMF. Therefore, `CMF` and `CT_flg` can be excluded.

Variable `Nodal_Status` and `ER_Status` are identical to each other for every observation, thus `Nodal_Status` was excluded from the data set.

Variable `Invasiveness_flg` and `Tumor_Size` are identical to each other, `Tumor_Size` was excluded from the data set.

4.2.1.3 Variables with all identical values

All entries of `Neoadj_CT`, `AC`, `AC+CMF`, `AC+T` and `Taxel` are 0, which do not provide any information that can be used in the analysis and hence should be excluded.

After the pre-selection of the explanatory variables, the remaining variables are:

`Age_diagnosis`, `Invasiveness_flg`, `ER_Status` and `CMF_Cycles`.

4.2.2 Missing data imputation

Full case analysis is an acceptable method to handle missing data in certain circumstances, for example, when the missingness of data is all at random, without dependence on observed or unobserved values, and the amount of missingness is less than 5%[\[27\]](#). These stringent missing assumptions may lead to skewed estimates and reduced power. An improvement has therefore been developed, which is the single imputation approach. However, once this family of methods is finished, analyses proceed as if the imputed values were known, actual values do not account for imputation uncertainty. Multiple imputations are thus more advantaged by filling in the missing values multiple times and creating multiple complete data sets. It should be noticed that for the multivariate method to have a good performance, there should be a dependency between the missing and non-missing values.

Variable name	Percentage of missing values
Age_diagnosis	0%
Invasiveness_flg	6.414%
ER_Status	13.980%
CMF_Cycles	0%

TABLE 4.2: Percentage of missing values for the selected explanatory variables

In this analysis, among the four selected explanatory features, **Invasiveness_flg** and **ER_Status** contain missing values. The percentage of missing values for each of the variables are 6.414% and 13.980% respectively. Considering this amount of missing data, it may incur serious issues for many machine learning algorithms. In this analysis, there may be more than one feature have missing values for each patient, the imputation method used should be capable to impute multiple missing values at the same time.

Different missing value imputation methods are available, with a different focus on qualitative or quantitative data. The methods tested were K-Nearest Neighbours Algorithm (k-NN) and Multivariate Imputation by Chained Equations (MICE), both approaches use a model to predict the missing values. Multiple models were developed based on these methods, with different parameters used in imputation.

4.2.2.1 K-Nearest Neighbours Algorithm

One popular technique for imputation is a K-nearest neighbour model. A new sample is imputed by finding the samples in the training set “closest” to it and averaging these nearby points to fill in the value[28].

The k-NN imputation was performed using the R package `recipes`. All the non-missing features in one row of the data frame are treated as the predictor which then is used to predict the missing value. The function `step_impute_knn` which uses Gower’s distance as the distance measurement is then being used to impute the missing value[29]. Considering the size of the data set ($N=767$), the number of neighbours was selected to be 10.

4.2.2.2 Multivariate Imputation by Chained Equations

The `mice` package in R implements a method to deal with missing data. This method is based on Fully Conditional Specification, where each incomplete variable is imputed by a separate model[25], and the imputation model is fitted to each variable as needed. The MICE algorithm is described below[30]:

1. Build a basic imputation for every missing value in the data set.
2. Set back missing values for one feature (F_x).
3. The observed values of F_x are used to train a prediction model in which F_x is a dependent variable, and the other features are independent.
4. The missing values for F_x are replaced with the predictions calculated by the model built in step 3.
5. For each feature with missing values, steps 2–4 are repeated. When a prediction model has imputed all features with missing values, one cycle or iteration is finished.
6. Steps 2–5 are repeated for n iterations, and the imputations are updated at each cycle. The objective is to use the number of iterations to achieve a stable imputation. The imputed data set is obtained in the last iteration.

Both `Invasiveness_flg` and `ER_Status` are categorical variables, in particular, binary variable. Therefore, the imputation methods to be used for each column in data are classification and regression trees (`cart`), random forest imputations (`rf`), and polytomous logistic regression (`polyreg`). The former two could be applied to any data type, while `polyreg` is suitable for nominal categorical data.

There are some hyper-parameter that need to be specified when using MICE imputation. First is the number of iterations (`maxit`) to be performed. Literature suggested that in general, 10 iterations for each imputation is sufficient[31]. The second hyper-parameter is the number of imputations, which is the number of imputed data sets (`m`) generated. The difference between each imputed data sets will only occur if the values were originally missing. According to [27], the power of MICE will be increased by setting `m=40`, however, in practice, the size of the data set, the amount of missing values and computational resources need to all be considered in order to determine a suitable value. In this analysis, since the number of variables (`n=4`) and the number of observations are both small, it is possible to set `m` to be a larger number, specifically, `m=50`. Furthermore, the last hyper-parameter `na.allow` is used to control the validation of each observation, the input value is 0, 1 or 2. For example, if `na.allow=1`, then only original observations with no more than one missing value will be included in the data set, and then be imputed.

In summary, there were 10 different imputed data sets generated in total with different methods of missing data imputation method. One of them was done by using k-NN algorithm, the other data sets were based on MICE imputation, with hyper-parameters `maxit=10, m=50, na.allow=(0, 1, 2), method=('cart', 'rf', 'polyreg')`, results in 9 different data sets.

4.3 Outcome variables processing

To ensure the validity of the data, patients with more than four out of eight missing records will be excluded from the study. Additional menstrual status records are available to be used if necessary. The maximum follow-up time of a patient was up to 29 years, while most of the follow-ups cease at ten years. For the simplicity of analysis and

to perform binary classification, normal menses and scanty menses were grouped to have menses, resulting in two levels of outcome records: no menses and have menses.

There were two different outcome variables available to be used in the data set, one is the menstrual status of a patient (`menstatus_Mo`), and the other is the corresponding amenorrheic status from chemotherapy (`Amen_ST`). In this analysis, the primary outcome variable is decided to be `Amen_ST`, while `menstatus_Mo` is used as the extra information for corrections and fixing of missing values.

Each individual outcome could be seen as an integer list, for example, `[0, 0, 0, 0, 1, 1, 1, 1]`, where 0 means no CIA occurs, and 1 means CIA occurs. The records are done at the third, sixth, ninth, twelfth, twenty-fourth, thirty-sixth, forty-eighth and sixtieth months. The pre-processing of the amenorrheic status could be described as the following two steps:

1. Validation of the pattern of records list.
2. Fix incorrect records by menstrual status information.

4.3.1 Validation pattern

To perform the validation on the pattern, the integer list could be shrunk, in other words, remove the consecutive and repeat integer. For example:

- `[1, 1, 1, 1, 1, 1, 1, 1]` to `[1]`
- `[0, 0, 1, 1, 1, 1, 1, 1]` to `[0, 1]`
- `[1, 1, 0, 0, 0, 1, 1, 1]` to `[1, 0, 1]`
- `[1, 1, 0, 0, 1, 1, 0, 0]` to `[1, 0, 1, 0]`

The valid patterns are:

- `[0]` and `[1]`
- `[0, 1]` and `[1, 0]`
- `[0, 1, 0]` and `[1, 0, 1]`

- [0, 1, 0, 1] and [1, 0, 1, 0]

Apart from these patterns, other patterns represent changes in the record that is too frequent, which may have been recorded with errors, or the individual's physical condition may be too uncertain for the study to be feasible.

To further ensure the stability of the records, the last two integers of the original integer list should be the same. Since the last two records correspond to the forty-eighth and sixtieth months, thus it is reasonable to consider the patient's physical condition is stable.

4.3.2 Fixing incorrect records

After the validation, the next step is to check if the records of menstrual status and amenorrheic status are matched for each individual. If the two records do not match, then the amenorrheic status should be replaced by the menstrual status information. For every missing value in amenorrheic status, the corresponding observation of menstrual status will be checked. If the latter one is non-missing, then the amenorrheic status will be fixed by the menstrual status. If both are missing, then they will be replaced by the previous value. After iterating through the integer list, if the first record (at the third month) is missing, it should be replaced by the value of the second record.

The above progress was done automatically by the program, after this, manual checking and correction were performed. Three records were modified manually.

After the validation and correction, the number of remaining observations is 588 (N = 588).

4.3.2.1 Incidence of CIA

Since amenorrhea generally occurs within two years from chemotherapy, amenorrheic status after one year will be neglected. The amenorrheic status from chemotherapy status was originally recorded for the third, sixth and twelfth months. To ensure the consistency of time interval, the status at the ninth month was added, using the information of corresponding menstrual status at the ninth month.

In order to fit the Cox proportional hazard model, event and occurrence time should be extracted. In this analysis, the occurrence of the event is defined to be the incidence of CIA. For a patient to be classified as suffering from CIA, the duration of missing menses should be at least three months, in other words, there should be at least two consecutive 1's presented in the first four values of the integer list. The time of the event is then determined by the first month of amenorrhea occurrence.

4.3.2.2 Recovery from CIA

Among patients who developed amenorrhea within the first year from chemotherapy ($N = 86$), their menstrual statuses after the first occurrence of CIA have been examined to see if their menses recovered or not. The event in this part of the analysis is hence the returning of menses, the time is calculated by subtracting the month of recovery by the time of CIA occurrence.

4.4 Model fitting

In this analysis, the Cox Proportional Hazard (Cox PH) model was used for both parts (incident of CIA and recovery of menstrual functionality after CIA). The Cox PH model[32] is a commonly used regression model in medical research, which investigate the association between the survival time of patients and predictor variables. There are two assumptions made by the Cox PH model: (1) survival curves for different strata must have hazard functions that are proportional over the time t and (2) the relationship between the log hazard and each predictor is linear[33]. This model is capable of both qualitative and quantitative predictors and enhances the power of survival analysis by examining the effect of multiple risk factors simultaneously.

The general form of the Cox PH model is:

$$h(t) = h_0(t)\exp(\beta_1x_1 + \dots + \beta_kx_k)$$

where the response variable (outcome) $h(t)$ is a hazard function which is the instantaneous event rate at time t . This hazard function has been modelled as an exponential

function of an arbitrary baseline hazard $h_0(t)$, without assuming a particular parametric form of $h_0(t)$. x_i and β_i are the predictors and the associated regression coefficient respectively. Compared with other survival analysis methods such as the Kaplan-Meier method, and the logistic regression, which is not an exclusive tool but commonly used in survival analysis, the Cox PH model considers the time until events occur, and is thus regarded to have more statistical power. Details about other methods and comparisons were illustrated in Chapter 2.

In the analysis of the incidence of chemotherapy-induced amenorrhea, patients are grouped into two different age groups (below 40 years and above 40 years) and models are fitted separately. The predictors used in this section are `Age_diagnosis`, `Invasiveness_flg`, `ER_Status` and `CMF_Cycles`. The analysis of the recovery from CIA does not split patients into different age group, and uses predictors `Age_diagnosis`, `Invasiveness_flg`, `ER_Status`, `CMF_Cycles`, and `Amen_time` (month of CIA occurred).

4.5 Model evaluation

The concordance statistic C become an increasingly popular and major summary statistic for survival analyses. Harrell et al.[34] provided a definition of the concordance for proportional hazard model, as the fraction of all the ordered time pairs (i.e., all i and j such that $K(i, j) = 1$ or $K(j, i) = 1$) in which the risk score, x , correctly predicts the order. Let τ be an upper time limit for comparison, the concordance statistic C is:

$$C = \frac{\sum_{i \neq j} \mathbb{1}\{t_i < \tau\} K(i, j) [\mathbb{1}\{x_i > x_j\} + \mathbb{1}\{x_i = x_j\}/2]}{\sum_{i \neq j} \mathbb{1}\{t_i < \tau\} K(i, j)}$$

For the usage in the Cox model, higher risk scores predict shorter event times, so C inverts the standard definition of concordance. Values of C range from 0 to 1, indicating a perfectly discordant to the concordant risk score.

Another way of defining the concordance is as the probability that the prediction x goes in the same direction as the actual data y , $P(x_i > x_j | y_i > y_j)$. The concordance is the fraction of concordant pairs[35]. C value of 0.5 means that the model is no better at predicting an outcome than random chance. Values over 0.7 indicate a good model. Values over 0.8 indicate a strong model.

4.6 Model selection

As previously stated in section 4.2.2, there are 10 data sets generated using different imputation methods and parameters. For each data set, 3 different models are fitted for analysis of the incidence of CIA (above 40 years and below 40 years), and the analysis of recovery from CIA. The concordance statistics are calculated for each model and sum up the statistics to obtain an overall score. These scores are then used to select the final missing data imputation method and hyper-parameters, hence the final data set and model formula.

Chapter 5

Results

This chapter provides the results of this analysis, produced based on the methodology discussed in Chapter 4. The results of the risk of chemotherapy-induced amenorrhea (CIA) occurrence and menses resumption from CIA are presented, detailed discussion about the results will be provided in Chapter 6.

5.1 Risk of the CIA occurrence

Multiple missing data imputation methods with corresponding hyper-parameters introduced in Chapter 4 were assessed by the resulting models' concordance. There were ten different imputed data sets generated for each group of patients (age above and below 40 years), with results shown in Table 5.1 and 5.2 for patients with age above 40 years and below 40 years respectively.

Risk of CIA occurrence: patients age above 40 years

Imputation method	Selected variables	Concordance, C	Number of imputation, m	MICE method	Number of NA allowed in one row
MICE	Age.diagnosis Invasiveness_flg ER_Status CMF.Cycles Invasiveness_flg*CMF.Cycles ER_Status*CMF.Cycles	0.6976	50	Classification and regression trees	0
MICE	Age.diagnosis Invasiveness_flg ER_Status CMF.Cycles Invasiveness_flg*CMF.Cycles ER_Status*CMF.Cycles	0.7143	50	Classification and regression trees	1
MICE	Age.diagnosis Invasiveness_flg ER_Status CMF.Cycles Invasiveness_flg*CMF.Cycles ER_Status*CMF.Cycles	0.7149	50	Classification and regression trees	2
MICE	Age.diagnosis Invasiveness_flg ER_Status CMF.Cycles Invasiveness_flg*CMF.Cycles ER_Status*CMF.Cycles	0.6976	50	Polynomial regression	0
MICE	Age.diagnosis Invasiveness_flg ER_Status CMF.Cycles Invasiveness_flg*CMF.Cycles ER_Status*CMF.Cycles	0.7097	50	Polynomial regression	1
MICE	Age.diagnosis Invasiveness_flg ER_Status CMF.Cycles Invasiveness_flg*CMF.Cycles ER_Status*CMF.Cycles	0.7096	50	Polynomial regression	2
MICE	Age.diagnosis Invasiveness_flg ER_Status CMF.Cycles Invasiveness_flg*CMF.Cycles ER_Status*CMF.Cycles	0.6976	50	Random forest	0
MICE	Age.diagnosis Invasiveness_flg ER_Status CMF.Cycles Invasiveness_flg*CMF.Cycles ER_Status*CMF.Cycles	0.7118	50	Random forest	1
MICE	Age.diagnosis Invasiveness_flg ER_Status CMF.Cycles Invasiveness_flg*CMF.Cycles ER_Status*CMF.Cycles	0.7102	50	Random forest	2
k-NN	Age.diagnosis Invasiveness_flg ER_Status CMF.Cycles Invasiveness_flg*CMF.Cycles ER_Status*CMF.Cycles	0.7090	-	-	-

TABLE 5.1: Patients with age above 40 years: variables and concordance of model fitted on imputed data sets, with different imputation methods.

Risk of CIA occurrence: patients age below 40 years

Imputation method	Selected variables	Concordance, C	Number of imputation, m	MICE method	Number of NA allowed in one row
MICE	Age.diagnosis Invasiveness_flg ER.Status CMF.Cycles Invasiveness_flg*CMF.Cycles	0.6976	50	Classification and regression trees	0
MICE	Age.diagnosis Invasiveness_flg ER.Status CMF.Cycles Invasiveness_flg*CMF.Cycles	0.6927	50	Classification and regression trees	1
MICE	Age.diagnosis ER.Status	0.6190	50	Classification and regression trees	2
MICE	Age.diagnosis Invasiveness_flg CMF.Cycles Invasiveness_flg*CMF.Cycles	0.6976	50	Polynomial regression	0
MICE	Age.diagnosis Invasiveness_flg ER.Status CMF.Cycles Invasiveness_flg*CMF.Cycles	0.6970	50	Polynomial regression	1
MICE	Age.diagnosis ER.Status	0.6287	50	Polynomial regression	2
MICE	Age.diagnosis Invasiveness_flg ER.Status CMF.Cycles Invasiveness_flg*CMF.Cycles	0.6976	50	Random forest	0
MICE	Age.diagnosis Invasiveness_flg ER.Status CMF.Cycles Invasiveness_flg*CMF.Cycles	0.7022	50	Random forest	1
MICE	Age.diagnosis Invasiveness_flg ER.Status CMF.Cycles Invasiveness_flg*CMF.Cycles Invasiveness_flg	0.6939	50	Random forest	2
k-NN	ER.Status CMF.Cycles Invasiveness_flg*CMF.Cycles	0.7003	-	-	-

TABLE 5.2: Patients with age below 40 years: variables and concordance of model fitted on imputed data sets, with different imputation methods.

Recall that the final imputation method is selected based on the average concordance achieved by both groups, and higher concordance corresponds to better model performance, hence a better quality of the imputation methods and parameters. As a result, the final selected missing data imputation method is Multivariate Imputation by Chained Equations (MICE), with the random forest as the imputation method to be used for each column with missing data occurring, and the maximum number of missing values in one record is 1.

The model selection was performed with R package `psfmi`, which is a package that provides functions to apply pooling, backward and forward selection of linear, logistic and Cox regression models across multiply imputed data sets using Rubin's Rules (RR)[36].

Using this final configuration, the model formula obtained with in-built for patients aged above 40 years is:

$$\begin{aligned} h_{above\ 40}(t) = & h_{0,above\ 40}(t) \exp (0.1281 \times \text{Age_diagnosis} + 0.8620 \times \text{Invasiveness_flg} \\ & + 0.6784 \times \text{ER_Status} + 0.5141 \times \text{CMF_Cycles} \\ & - 0.2566 \times \text{Invasiveness_flg} : \text{CMF_Cycles} \\ & - 0.2831 \times \text{ER_Status} : \text{CMF_Cycles} \end{aligned}$$

and the model formula for patients aged below 40 years is:

$$\begin{aligned} h_{below\ 40}(t) = & h_{0,below\ 40}(t) \exp (-0.0192 \times \text{Age_diagnosis} + 1.0156 \times \text{Invasiveness_flg1} \\ & + 1.1828 \times \text{ER_Status1} + 0.2080 \times \text{CMF_Cycles} \\ & - 0.3146 \times \text{Invasiveness_flg1} : \text{CMF_Cycles6} \end{aligned}$$

where $h_{0,above\ 40}(t)$ and $h_{0,below\ 40}(t)$ represent the unknown baseline hazard functions for the Cox proportional hazard models of patient groups of above and below 40 years old respectively.

5.2 Recovery from CIA

The final missing data imputation method of the analysis of the resumption of menses was selected together with the models' concordance as in Section 5.1, which is MICE with random forest imputation method, with a maximum of 1 NA value allowed in one record of the data set. The model concordance of this setting is 0.6994.

Resumption of menses					
Imputation method	Selected variables	Concordance, C	Number of imputation, m	MICE method	Number of NA allowed in one row
MICE	Age_diagnosis Amen_time	0.7074	50	Classification and regression trees	0
MICE	Age_diagnosis Amen_time	0.6994	50	Classification and regression trees	1
MICE	Age_diagnosis Amen_time	0.6927	50	Classification and regression trees	2
MICE	Age_diagnosis Amen_time	0.7074	50	Polynomial regression	0
MICE	Age_diagnosis Amen_time	0.6994	50	Polynomial regression	1
MICE	Age_diagnosis Amen_time	0.6927	50	Polynomial regression	2
MICE	Age_diagnosis Amen_time	0.7074	50	Random forest	0
MICE	Age_diagnosis Amen_time	0.6994	50	Random forest	1
MICE	Age_diagnosis Amen_time	0.6927	50	Random forest	2
k-NN	Age_diagnosis Amen_time	0.6927	-	-	-

TABLE 5.3: Resumption of menses for patients who developed CIA: variables and concordance of model fitted on imputed data sets, with different imputation methods.

The associated model formula is:

$$h_{recovery}(t) = h_{0,recovery}(t) \exp (-0.0505 \times \text{Age_diagnosis} - 0.2371 \times \text{Amen_time})$$

Chapter 6

Discussion

This chapter examines the key findings and implications of the findings, followed by a review of some of the study's strengths and limitations.

6.1 Incidence of CIA

In Chapter 5, the variable selection and the final models are obtained by R package `psfmi`, which did not provide complete information about the models. For example, the multi-level categorical variable `CMF_Cycles` with levels 1, 6, and 7 were concatenated into a single variable, with one coefficient. Furthermore, the interaction terms `Invasiveness_flg:CMF_Cycles` and `ER_Status:CMF_Cycles` should also be presented separately for each level of `Invasiveness_flg`, `CMF_Cycles` and `ER_Status`. Therefore, models were fitted again with function `coxph()` in R package `survival`[\[37\]](#) to the same imputed data sets, with the same selected variables.

With the use of `coxph()`, the concordance of the model for patients above 40 years is 0.710, while the concordance of the model for patients below 40 is 0.682. Compared with the statistics obtained in Section 5.1 (above 40 years: 0.7118, below 40 years: 0.7022), there is a slight decrease in the concordance of the latter model. The differences in the model concordances and the coefficient estimates were caused by the different packages used, however, the differences were not significant and thus acceptable.

6.1.1 Patients with age above 40

Variable	Coefficient	exp(coefficient)	se(coefficient)	Pr(> z)
Age_diagnosis	0.1439	1.1548	0.0076	< 0.001
Invasiveness_flg1	0.4732	1.6052	0.0487	< 0.001
ER_Status1	0.5601	1.7509	0.0529	< 0.001
CMF_Cycles6	2.9075	18.3108	0.0879	< 0.001
CMF_Cycles7	2.4294	11.3521	0.1375	< 0.001
Invasiveness_flg1*CMF_Cycles6	-0.9312	0.3941	0.0995	< 0.001
Invasiveness_flg1*CMF_Cycles7	-1.1621	0.3128	0.1354	< 0.001
ER_Status1*CMF_Cycles6	-1.3036	0.2716	0.1070	< 0.001
ER_Status1*CMF_Cycles7	-1.3752	0.2528	0.1371	< 0.001

TABLE 6.1: Patients with age above 40 years: model variable coefficients, standard errors and variable importance.

The model formula for patients aged above 40 years is:

$$\begin{aligned}
 h_{above\ 40}(t) = & h_{0,above\ 40}(t) \exp (0.1439 \times \text{Age_diagnosis} \\
 & + 0.4732 \times \text{Invasiveness_flg1} + 0.5601 \times \text{ER_Status1} \\
 & + 2.9075 \times \text{CMF_Cycles6} + 2.4294 \times \text{CMF_Cycles7} \\
 & - 0.9312 \times \text{Invasiveness_flg1:CMF_Cycles6} \\
 & - 1.1621 \times \text{Invasiveness_flg1:CMF_Cycles7} \\
 & - 1.3036 \times \text{ER_Status1:CMF_Cycles6} \\
 & - 1.3752 \times \text{ER_Status1:CMF_Cycles7})
 \end{aligned}$$

From the results, variables `Age_diagnosis`, `Invasiveness_flg1`, `ER_Status1`, `CMF_Cycles6`, and `CMF_Cycles7` all have a positive effect on the hazard ratio, which means that for a patient who aged above 40 years old, with retaining other variable constant, her risk of developing CIA increases as her age increases, or if she is experiencing invasive breast cancer, or is ER-positive, or if she is treated with 6 or 7 chemotherapy cycles. However, it should also be noticed that all the interaction terms (`Invasiveness_flg1:CMF_Cycles6` and `ER_Status1:CMF_Cycles7`) introduce negative effects on the risk of CIA.

To further illustrate the interpretation of these coefficients, suppose there is a patient who is aged 42 years old with invasive cancer, is identified with negative ER status at disease diagnosis, and received 6 CMF cycles as her treatment. Her hazard ratio of

developing CIA at the 6th month is:

$$\begin{aligned}
 h(6) &= h_0(6) \exp (0.1439 \times (42 - 41) + 0.4732 \times 1 + 0.5601 \times 0 + 2.9075 \times 1 + 2.4294 \times 0 \\
 &\quad - 0.9312 \times 1 - 1.1621 \times 0 - 1.3036 \times 0 - 1.3752 \times 0 \\
 &= h_0(6) \exp (2.5934) \\
 &= h_0(6) \times 13.37517
 \end{aligned}$$

Thus she is 13.3752 times more likely to develop CIA at the 6th month, compared to a patient aged 41 years old, without invasive cancer and is ER-negative, and receives 1 cycle of CMF, whose hazard of developing CIA is equal to the baseline hazard $h_0(6)$.

Even though all the interaction terms will decrease the hazard ratio of developing CIA, the positive effect introduced by the first-order main effects cannot vanish. For example, to enable the maximum negative effects from the interaction terms, the patient should be ER-positive and receiving 7 CMF cycles, and the cancer should be invasive, together these will create a decrease in hazard ratio with a factor of $\exp(-1.1621 - 1.3752) = \exp(-2.5373)$, however, the main effects will introduce a factor of $\exp(0.4732 + 0.5601 + 2.4294) = \exp(3.4627)$, therefore, the hazard ratio will still increase by $\exp(-2.5373 + 3.4627) = 2.522877$.

6.1.2 Patients with age below 40

Variable	Coefficient	exp(coefficient)	se(coefficient)	Pr(> z)
Age_diagnosis	-0.0093	0.9908	0.0068	0.173
Invasiveness_flg1	0.9643	2.6229	0.0701	< 0.001
ER_Status1	0.8097	2.2473	0.0562	< 0.001
CMF_Cycles6	1.8661	6.4628	0.0843	< 0.001
CMF_Cycles7	0.9579	2.6062	0.1013	< 0.001
Invasiveness_flg1:CMF_Cycles6	-2.3235	0.0979	0.1296	< 0.001
Invasiveness_flg1:CMF_Cycles7	-2.1326	0.1185	0.1777	< 0.001

TABLE 6.2: Patients with age below 40 years: model variable coefficients, standard errors and variable importance.

The model formula for patients aged below 40 years is:

$$\begin{aligned}
 h_{\text{below } 40}(t) = & h_{0,\text{below } 40}(t) \exp (-0.0093 \times \text{Age_diagnosis} \\
 & + 0.9643 \times \text{Invasiveness_flg1} + 0.8097 \times \text{ER_Status1} \\
 & + 1.8661 \times \text{CMF_Cycles6} + 0.9579 \times \text{CMF_Cycles7} \\
 & - 2.3235 \times \text{Invasiveness_flg1:CMF_Cycles6} \\
 & - 2.1326 \times \text{Invasiveness_flg1:CMF_Cycles7})
 \end{aligned}$$

Compared with the above 40 years old group, the below 40 years group has one less interaction term `ER.Status:CMF.Cycles`, other than this, the other selected predictors are the same. The interaction term `Invasiveness_flg:CMF.Cycles` still provides a negative effect on the hazard ratio, while the main effect terms all imply a positive influence on the ratio except for `Age_diagnosis`. This finding suggests that for patients aged equal to or below 40 years old, every one-year increment of age at diagnosis will decrease the hazard ratio by a factor of $\exp(-0.0093) = 0.9908$. Although this is only a slight decrease, it illustrates a different trend to the above 40 years group.

Similar to the other group, the decrease caused by the interaction term and `Age_diagnosis` on hazard ratio cannot vanish the increase created by the main effects.

6.1.3 Baseline hazards of two groups

Table 6.3 and 6.4 present the baseline hazards $h_0(t)$ of patient groups aged above and below 40 years, the estimations were obtained by using the `basehaz()` function from the `survival` package.

Time	3	6	9	12
Hazard	0.0232	0.0479	0.0722	0.0791

TABLE 6.3: Patients with age above 40 years: baseline hazards at 3, 6, 9, 12 months.

Time	3	6	9	12
Hazard	0.0105	0.0217	0.0325	0.0363

TABLE 6.4: Patients with age below 40 years: baseline hazards at 3, 6, 9, 12 months.

An important finding is that patients aged equal to or below 40 years old have lower baseline hazards than the group above 40 years old, more specifically, the baseline hazard

of developing CIA for the former group of patients is approximately half that of the latter group at all timestamps (at 3th, 6th, 9th, 12th months). This result demonstrates the importance of sub-setting patients by age at diagnosis, and further suggest that when discussing the possible risks and side effects of chemotherapy for cancer with a patient, it is important to treat patients separately according to their age. 40 years could be an informed cut-off point. Moreover, since the average baseline hazards for older patients are higher than the younger patients, the treatment duration and strength should be considered more carefully in order to minimise the influence of the potential side effects.

6.2 Recovery from CIA

Variable	Coefficient	exp(coefficient)	se(coefficient)	Pr(> z)
Age_diagnosis	-0.0527	0.9487	0.0031	< 0.001
Amen_time	-0.2320	0.7930	0.0078	< 0.001

TABLE 6.5: Resumption of menses: model variable coefficients, standard errors and variable importance.

The concordance C of this model (with `coxph()`) is 0.693, which is similar to the concordance obtained in Chapter 5 (0.6994).

The model formula is:

$$h_{\text{recovery}}(t) = h_{0,\text{recovery}}(t) \exp(-0.0527 \times \text{Age_diagnosis} - 0.2320 \times \text{Amen_time})$$

In contrast to the analysis of the incidence of CIA, there were fewer predictors selected. Only 2 selected predictors in the model of this section, which were **Age_diagnosis** and **Amen_time** (the month of CIA occurred after CMF regimen). Both predictors have a negative coefficient which illustrates the negative effect on the likelihood of menstrual resumption. For every 1-year increase in patient's age at diagnosis, the hazard (event) ratio of menses recovery is decreased by 0.9487. On the other hand, for every 1-month increase in the lasting time of CIA, the event ratio of menses recovery is decreased by 0.7930.

Time	6	9	15	18	21	24
Hazard	0.1526	0.2157	0.3005	0.4259	0.6462	0.6756
Time	27	30	33	42	60	
Hazard	0.7055	0.7358	0.8017	0.8381	0.8381	

TABLE 6.6: Resumption of menses: baseline hazards (likelihood) at different timestamps in 5 years.

Table 6.6 presents the baseline hazards at each timestamp. It is shown that as time increase, the baseline likelihood of recovery increases, which could be interpreted as the instantaneous baseline likelihood of recovery is higher at later month after CMF. Furthermore, since the recovery time recorded and input into the model was recorded from the time chemotherapy began and the most patient experienced CIA within one year of CMF began, therefore it is reasonable that the baseline likelihood of menses recovery is low before the 15th month as there was overlap of `Amen_time` and `Recovery_time`. After the 15th month, there is a significant increase of the baseline hazard from the 18th month to 21th month, potentially advises that most patients will be more likely to have their menses recovered after 18 months of CMF regimen.

6.3 Strength

Firstly, the approach to handle the missing data in the original data set is to impute the missing data based on other observed data. As the original data was collected by surveying patients, the recorded data is likely to be incomplete. Moreover, as there was follow-up information collected for over more than 5 years, all participants cannot have the same duration of follow-up. Some might exit the survey earlier than others. Therefore, this missing value imputation method overcomes the drawback of reducing sample size by implementing complete-case analysis, which excludes all records with a missing value. The original data set contains only 767 records, and there were 114 incomplete cases (approximately 14.86%).

Secondly, by comparing different missing data imputation methods, the final selected method, multivariate imputation by chained equations (MICE) offers a great advantage over other missing data techniques in terms of its flexibility and independent models for columns in the data set that needs to be imputed. It also generates several complete data sets, with missing values imputed based on observed values for a specific individual

and observed data relations for other participants, providing the observed variables are included in the imputation model. Because multiple imputations entail making numerous estimates for each missing value, studies that make use of multiply imputed data account for imputation uncertainty and produce appropriate standard errors.[38].

The last and most important strength is the approach of sub-group analysis. This approach successfully captures the different characteristics of patients older than 40 years old and younger than 40 years old. The success was demonstrated by the different baseline hazards (younger patients have approximately half of the baseline hazard than the older patients), and the opposite sign of the coefficient of predictor `Age_diagnosis`. Therefore, the sub-setting strategy enables more targeted models to be developed.

6.4 Limitation

The original data set only contain 767 observations, and there were only 561 patients remaining after excluding the post-menopausal patients. This sample size could be considered relatively small which accompanies the limitation of being more likely to produce false-positive results or over-estimating the magnitude of an association[39]. Because of the small sample size, the predictors will have a larger standard error and hence be considered statistically insignificant. In this way, some important predictors might be excluded in the variable selection process, especially in the analysis of menstrual resumption.

One major concern is the disadvantages of MICE. MICE does not have the same theoretical justification as other imputation approaches. In particular, fitting a series of conditional distributions, as is done using the series of regression models, may not be consistent with proper joint distribution. Furthermore, clustering is not always automatically incorporated by the MICE procedures, but it is important to address in both data analyses and when imputing missing data[38]. The risk of neglecting the clustering nature may incur inaccurate model fitting. Some research found that this may not be a large issue in applied settings[40, 41], but further research of its suitable application is needed for better use.

For the analysis of the incidence of CIA (Section 6.1), it should be noticed that since the baseline hazards were calculated based on the column means of the data set, the

information provided by these estimates is limited because of the interaction terms (`Invasiveness_flg:CMF_Cycles` and `ER_Status:CMF_Cycles`). However, since the baseline hazard function is analogous to the intercept term in a multiple regression or logistic regression model and the Cox regression model does not require this to be specified, these estimations are not usually important in the model interpretation. These are rather used to demonstrate the importance of sub-group analysis. One other limitation is that for the menses resumption part of the analysis, there were only 98 patients who developed CIA in the whole data set, 12 of them aged above 50 years old and been considered as not eligible for the analysis because of the high likelihood of being post-menopausal, thus the sample size ($N=86$) is very limited. There might be other potential significant predictors that have not been included in the model due to insufficient data, further experiments and analyses are required to obtain a more precise result.

Although all models fitted for the three analysis achieved concordance statistics of approximately 0.7, there are still leaves room for improvement. A more comprehensive analysis should include more candidate predictors, which requires to be recorded when performing data collection. Some other variables that might influence the outcome could be patient's body mass index (BMI), smoking status, previous medical history, menarche time, et cetera. More participants, that is, a larger sample size will be better for minimise false positive.

Chapter 7

Conclusions and Future Work

This chapter concludes the research with a summary of the main finding, followed by suggestions on the future work.

7.1 Conclusion

In this research, we investigated the significant influential factors associated with the risk of developing chemotherapy-induced amenorrhea for breast cancer patients who receive chemotherapy as their treatment, followed by an analysis of the likelihood of menstrual resumption of patients who suffered from CIA. Multivariate imputation by chained equations (MICE) missing-data imputation was used for replacing the missing values of the predictors in the data, while pattern-matching and correction with relevant information were applied to the outcome variables (menstrual status) pre-processing.

7.1.1 Importance of missing-data imputation

In Chapter 4, we discussed different approaches for handling missing data, and the algorithms of missing value imputations such as multivariate imputation by chained equations (MICE) and k-nearest-neighbours (k-NN).

With the original 767 records of patients, and 561 out of them were in pre-monopausal or peri-monopausal status, we considered it important to maintain the sample size as large as possible to minimise the disadvantages of small data set, which are the higher

probability of producing false-positive results, and the chance of excluding predictors that is, in fact, significant but estimated to have the large standard error. By using the MICE algorithm with different imputation methods (random forest, classification and regression tree) and the different number of missing values in one row, 50 imputed data sets were generated and merged to form a large data set with each combination of settings. The k-nearest-neighbour algorithm further generates another imputed data set. Therefore, there were 10 imputed data sets created in total.

By fitting a Cox regression model on each of the imputed data sets, the quality of the imputation methods and hyper-parameters were assessed based on the concordance, C -statistics of the corresponding model. The final selected method was MICE with random forest, and the number of missing values allowed in each record is 1. This model with this method achieved an average concordance of 0.7045. A detailed comparison was presented in Chapter 5.

7.1.2 Incidence of CIA for different age groups

In the analysis of the risk of CIA occurrence, we divided the patients into two groups by their age at cancer diagnosis. The first group contains patients who aged above 40 years old, and the second group contains patients who were 40 years old or younger. Two models were fitted separately for the groups.

For the first group, the predictors that were considered as significantly associated with the risk of CIA are age at diagnosis, whether the cancer is invasive or not, ER status of the patient, number of cycles of chemotherapy that the patient received with two more interaction terms between invasiveness of cancer and CMF cycles, and ER status with CMF cycles. The first-order main effect terms all imply an increase in the risk of developing CIA, while the interaction terms all lower the risk. The model achieved a concordance of 0.7100.

For the second group, the significant predictors are age at diagnosis, whether the cancer is invasive or not, ER status of the patient, number of cycles of chemotherapy that the patient received with an extra interaction term between invasiveness of cancer and CMF cycles, that is, no interaction between ER status with CMF cycles as compared to the first group of older patients. In addition, the age at diagnosis has a negative coefficient

hence the risk of a patient developing CIA will decrease as her becoming older, until she reaches the 40 years old threshold. The model for the second group has a concordance of 0.6820.

The different trends in the effect of age on the risk of CIA demonstrated the importance of sub-group analysis, and this can be further proved by the different significant variables selected for each group. Both concordances of the models can be considered as a common result for survival data.

7.1.3 Resumption of menses after developing CIA

There were 86 patients developed CIA, which were the subjects of this part of the analysis. Similar to the previous analysis, a Cox regression model was fitted and variables selection was performed using the backward-elimination method with the variables' p-values. The final selected significant variables were . Both variables had negative coefficients and thus imply negative effect on the likelihood of menses recovery. The concordance of this model was 0.6930, which is also a acceptable result.

It should be noted that because of the very limited sample size, the model selection process might have excluded some important variables. The reason is that small sample size always associated with large standard errors of the variable's coefficient estimate, and hence been considered as not statistically significant if the corresponding confidence interval includes zero. However, this analysis still provides an insight of the influential factors of the likelihood of menses resumption.

7.2 Future Work

It is advised that future research focus on the following areas:

1. **Perform analysis with data of larger sample size.** This could decrease the probability of excluding important predictors and the probability to producing false-positive result.

2. **Perform data collection that obtains a more comprehensive record** of the patient's physical characteristics, medical history and measure patient's hormone levels at follow-ups.
3. **Experimenting more missing data imputation method.** Pre-select methods based on the preliminary knowledge on the type of missing data. For example, missing completely at random (MCAR), missing at random (MAR), missing not at random (MNAR), and structurally missing. Different types of missing data may require different approaches.
4. **Validates the model on external data set.** The model always has the probability of over-fitting to a particular data set, therefore, it is important to examine the model performance with unobserved data to the model.

Appendix A

Nomograms

A.1 Nomograms of incidence of CIA

This section provides partial nomograms that can be used to manually obtain predicted values from the fitted regression models. Two nomograms for patients with age above and below 40 years old are provided separately, with another nomogram for the recovery from CIA.

The nomogram does not have lines representing sums, but it has a reference line for reading scoring points (range 0-100). Once the reader manually totals the points, the predicted values can be read at the bottom[\[42\]](#).

A.1.1 Patients aged above 40 years

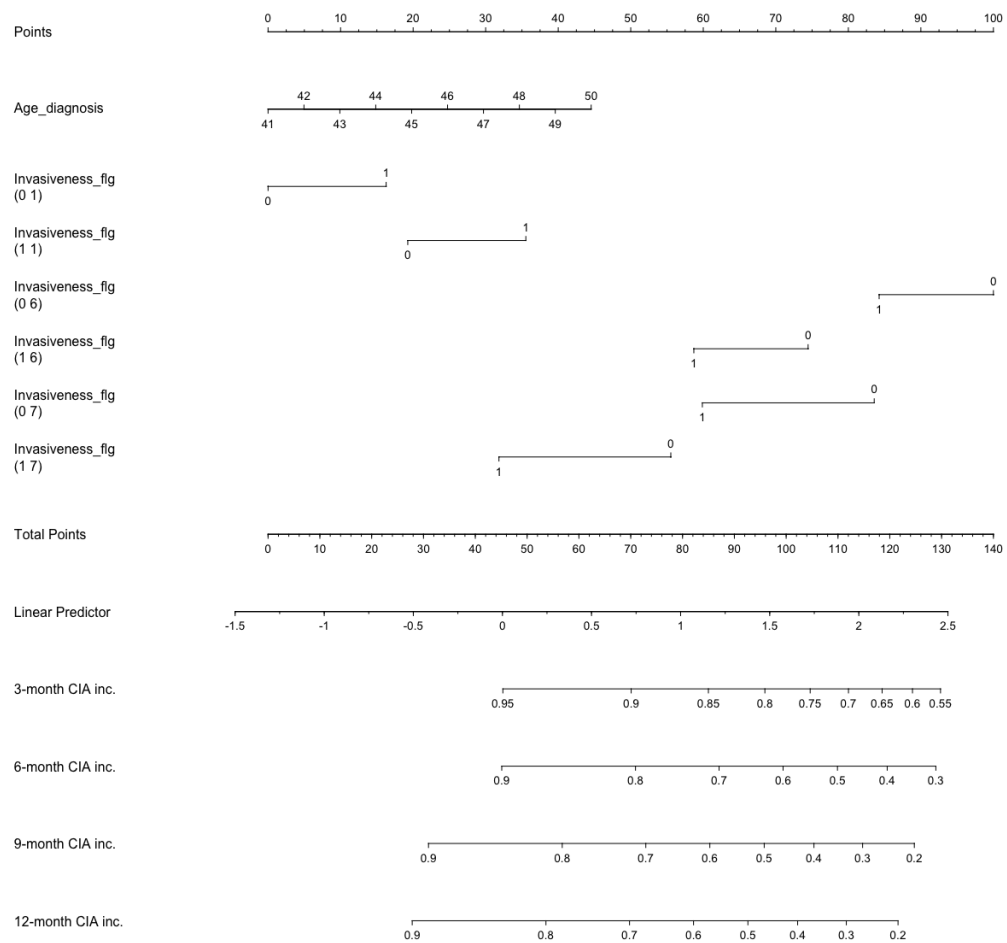


FIGURE A.1: Partial nomogram of patients aged above 40 years

A.1.2 Patients aged above 40

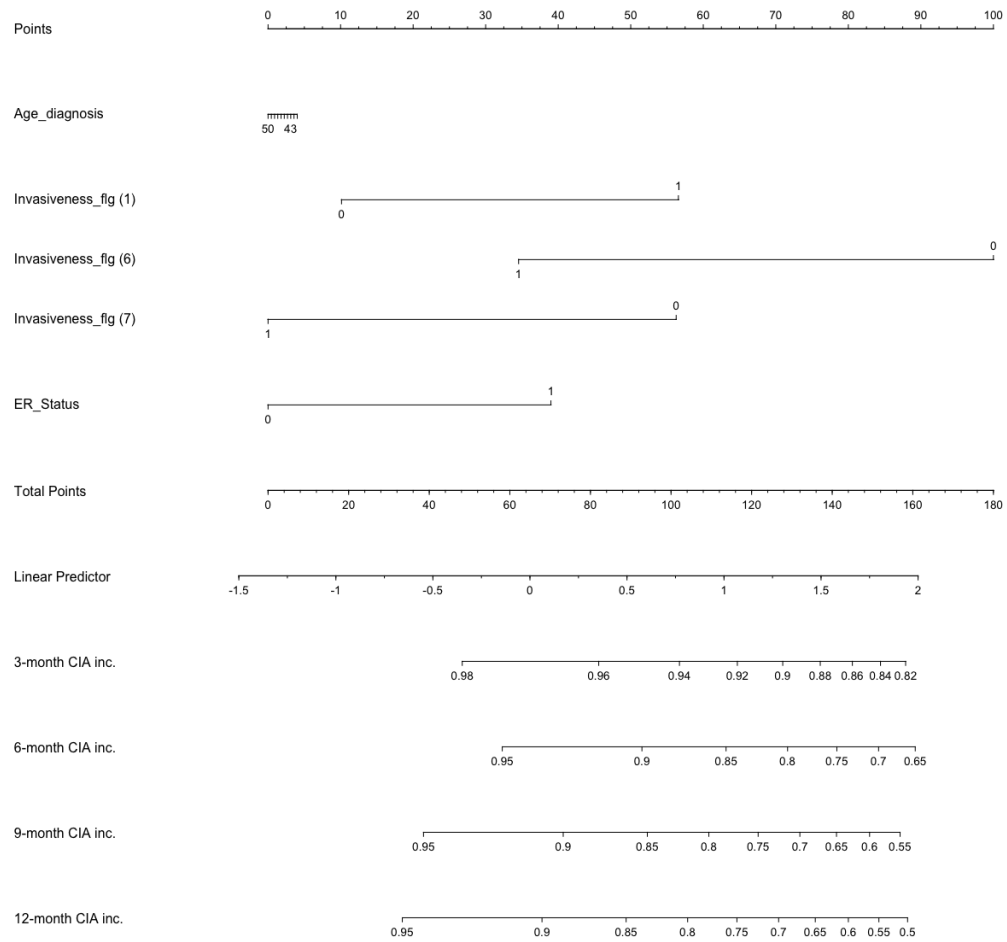


FIGURE A.2: Partial nomogram of patients aged below 40 years

A.2 Nomogram of menstrual resumption

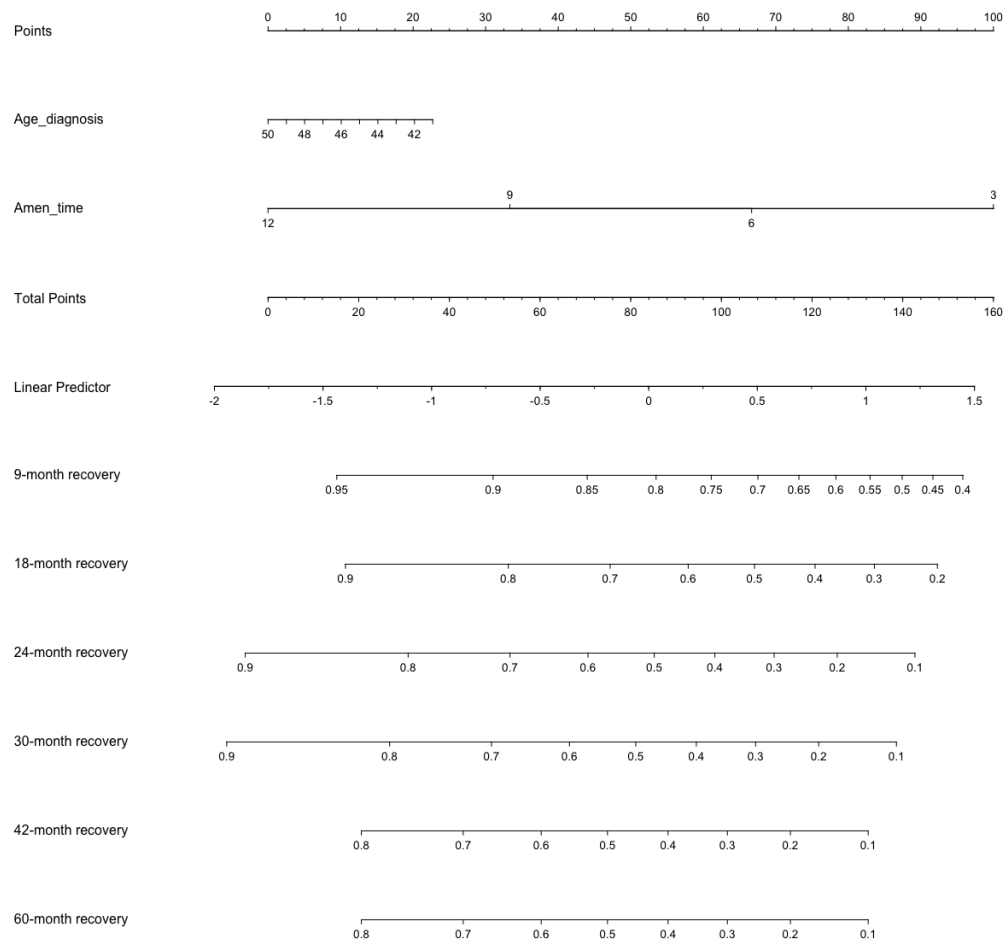


FIGURE A.3: Partial nomogram of resumption of menses

Bibliography

- [1] Ann Partridge, Shari Gelber, Richard D. Gelber, Monica Castiglione-Gertsch, Aron Goldhirsch, and Eric Winer. Age of menopause among women who remain premenopausal following treatment for early breast cancer: Long-term results from international breast cancer study group trials v and vi. *European Journal of Cancer*, 43(11):1646–1653, Jul 2007.
- [2] Ann H. Partridge, Shari Gelber, Jeffrey Peppercorn, Ebonie Sampson, Katherine Knudsen, Marc Laufer, Randi Rosenberg, Michele Przypyszny, Alison Rein, Eric P. Winer, and et al. Web-based survey of fertility issues in young women with breast cancer. *Journal of Clinical Oncology*, 22(20):4174–4183, 2004.
- [3] Global Burden of Disease Collaborative Network. Global burden of disease study 2019 (gbd 2019) reference life table, 2021.
- [4] Early Breast Cancer Trialists’ Collaborative Group (EBCTCG), R Peto, C Davies, J Godwin, R Gray, H C Pan, M Clarke, D Cutter, S Darby, P McGale, C Taylor, Y C Wang, J Bergh, A Di Leo, K Albain, S Swain, M Piccart, and K Pritchard. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet*, 379(9814):432–444, February 2012.
- [5] S D Harlow and S A Ephross. Epidemiology of menstruation and its relevance to women’s health. *Epidemiol. Rev.*, 17(2):265–286, 1995.
- [6] S D Harlow and S L Zeger. An application of longitudinal methods to the analysis of menstrual diary data. *J. Clin. Epidemiol.*, 44(10):1015–1025, 1991.
- [7] Cancer Australia. *Breast cancer and early menopause — a guide for younger women*. Cancer Australia, 3 edition, 2013.

- [8] J Bines, D M Oleske, and M A Cobleigh. Ovarian function in premenopausal women treated with adjuvant chemotherapy for breast cancer. *J. Clin. Oncol.*, 14(5):1718–1729, May 1996.
- [9] Arran K Turnbull, Samir Patel, Carlos Martinez-Perez, Anne Rigg, and Olga Oikonomidou. Risk of chemotherapy-related amenorrhoea (CRA) in premenopausal women undergoing chemotherapy for early stage breast cancer. *Breast Cancer Res. Treat.*, 186(1):237–245, February 2021.
- [10] Giok S Liem, Frankie K F Mo, Elizabeth Pang, Joyce J S Suen, Nelson L S Tang, Kun M Lee, Claudia H W Yip, Wing H Tam, Rita Ng, Jane Koh, Christopher C H Yip, Grace W S Kong, and Winnie Yeo. Chemotherapy-related amenorrhea and menopause in young chinese breast cancer patients: Analysis on incidence, risk factors and serum hormone profiles. *PLoS One*, 10(10):e0140842, October 2015.
- [11] Bailin Zhang, Jinqi Wu, Rongshou Zheng, Qian Zhang, Margaret Zhuoer Wang, Jun Qi, Haijing Liu, Yipeng Wang, Yang Guo, Feng Chen, Jing Wang, Wenyue Lyu, Jidong Gao, Yi Fang, Wanqing Chen, and Xiang Wang. Evaluation of menopausal status among breast cancer patients with chemotherapy-induced amenorrhea. *Chin. J. Cancer Res.*, 30(4):468–476, August 2018.
- [12] Ann Partridge, Shari Gelber, Richard D Gelber, Monica Castiglione-Gertsch, Aron Goldhirsch, and Eric Winer. Age of menopause among women who remain premenopausal following treatment for early breast cancer: long-term results from international breast cancer study group trials V and VI. *Eur. J. Cancer*, 43(11):1646–1653, July 2007.
- [13] Sunyoung Lee, Whoon Jong Kil, Mison Chun, Yong-Sik Jung, Seok Yun Kang, Seung-Hee Kang, and Young-Taek Oh. Chemotherapy-related amenorrhea in premenopausal women with breast cancer. *Menopause*, 16(1):98–103, January 2009.
- [14] Angiolo Gadducci, Stefania Cosio, and Andrea Riccardo Genazzani. Ovarian function and childbearing issues in breast cancer survivors. *Gynecological endocrinology*, 23(11):625–631, November 2007.
- [15] I H Park, H S Han, H Lee, K S Lee, H S Kang, S Lee, S W Kim, S Jung, and J Ro. Resumption or persistence of menstruation after cytotoxic chemotherapy is

- a prognostic factor for poor disease-free survival in premenopausal patients with early breast cancer. *Ann. Oncol.*, 23(9):2283–2289, September 2012.
- [16] Sandra M Swain, Jong-Hyeon Jeong, Charles E Geyer, Jr, Joseph P Costantino, Eduardo R Pajon, Louis Fehrenbacher, James N Atkins, Jonathan Polikoff, Victor G Vogel, John K Erban, Priya Rastogi, Robert B Livingston, Edith A Perez, Eleftherios P Mamounas, Stephanie R Land, Patricia A Ganz, and Norman Wolmark. Longer therapy, iatrogenic amenorrhea, and survival in early breast cancer. *N. Engl. J. Med.*, 362(22):2053–2065, June 2010.
- [17] M Vanhuyse, C Fournier, and J Bonnetterre. Chemotherapy-induced amenorrhea: influence on disease-free survival and overall survival in receptor-positive premenopausal early breast cancer patients. *Ann. Oncol.*, 16(8):1283–1288, August 2005.
- [18] Qiong Zhou, Wenjin Yin, Yueyao Du, Zhenzhou Shen, and Jingsong Lu. Prognostic impact of chemotherapy-induced amenorrhea on premenopausal breast cancer. *Menopause*, 22(10):1091–1097, October 2015.
- [19] Wendy R Parulekar, Andrew G Day, Jon A Ottaway, Lois E Shepherd, Maureen E Trudeau, Vivien Bramwell, Mark Levine, Kathleen I Pritchard, and National Cancer Institute of Canada Clinical Trials Group. Incidence and prognostic impact of amenorrhea during adjuvant therapy in high-risk premenopausal breast cancer: analysis of a national cancer institute of canada clinical trials group Study–NCIC CTG MA.5. *J. Clin. Oncol.*, 23(25):6002–6008, September 2005.
- [20] Shoshana M Rosenberg and Ann H Partridge. Premature menopause in young breast cancer: effects on quality of life and treatment interventions. *J. Thorac. Dis.*, 5 Suppl 1:S55–61, June 2013.
- [21] Patrick Schober and Thomas R Vetter. Logistic regression in medical research. *Anesth. Analg.*, 132(2):365–366, February 2021.
- [22] David Chen, Sijia Liu, Paul Kingsbury, Sunghwan Sohn, Curtis B Storlie, Elizabeth B Habermann, James M Naessens, David W Larson, and Hongfang Liu. Deep learning and alternative learning strategies for retrospective real-world clinical data. *NPJ Digit. Med.*, 2(1):43, May 2019.

- [23] Vianda S Stel, Friedo W Dekker, Giovanni Tripepi, Carmine Zoccali, and Kitty J Jager. Survival analysis i: the Kaplan-Meier method. *Nephron Clin. Pract.*, 119(1):c83–8, June 2011.
- [24] The Royal Women’s Hospital. About menopause.
- [25] Stef van Buuren. Package ‘mice’, Nov 2021.
- [26] Monica N Fornier, Shanu Modi, Katherine S Panageas, Larry Norton, and Clifford Hudis. Incidence of chemotherapy-induced, long-term amenorrhea in patients with breast carcinoma age 40 years and younger after adjuvant anthracycline and taxane. *Cancer*, 104(8):1575–1579, October 2005.
- [27] John W Graham, Allison E Olchowski, and Tamika D Gilreath. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prev. Sci.*, 8(3):206–213, September 2007.
- [28] Max Kuhn and Kjell Johnson. *Applied predictive modeling*. Springer, 2019.
- [29] Max Kuhn and Hadley Wickham. Impute via k-nearest neighbors - step_impute_knn.
- [30] Maritza Mera-Gaona, Ursula Neumann, Rubiel Vargas-Canas, and Diego M. López. Evaluating the impact of multivariate imputation by mice in feature selection. *PLOS ONE*, 16(7), 2021.
- [31] Trivellore E Raghunathan, Peter W Solenberger, and John Van Hoewyk. Iweware: Imputation and variance estimation software. *Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan*, 2002.
- [32] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.
- [33] Dana Hashim and Elisabete Weiderpass. Cancer survival and survivorship. In Paolo Boffetta and Pierre Hainaut, editors, *Encyclopedia of Cancer (Third Edition)*, pages 250–259. Academic Press, Oxford, third edition edition, 2019.
- [34] F E Harrell. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, May 1982.

-
- [35] Terry M. Therneau and Elizabeth Atkinson. Concordance. *The Lancet*, 342, 1993.
- [36] Martijn Heymans. *psfmi: Prediction Model Pooling, Selection and Performance Evaluation Across Multiply Imputed Datasets*, 2022. R package version 1.0.0.
- [37] Terry M Therneau. Survival analysis [r package survival version 3.4-0], Aug 2022.
- [38] Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 2011.
- [39] A. Hackshaw. Small studies: Strengths and limitations. *European Respiratory Journal*, 32(5):1141–1143, 2008.
- [40] Joseph L. Schafer and John W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.
- [41] Jaap P.L Brand. *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. Erasmus Universiteit Rotterdam, 1999.
- [42] Frank E Harrell. Package rms, Apr 2022.