UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2015:4

Reasons for Missing Data of Risk Categorisation in NPCR

Marcus Westerberg

Examensarbete i matematik, 15 hp
Handledare: Hans Garmo, Regionalt cancercentrum Uppsala Örebro
Ämnesgranskare och examinator: Rolf Larsson
Maj 2015

Department of Mathematics
Uppsala University

Reasons for Missing Data of Risk Categorisation in NPCR

Marcus Westerberg

June 1, 2015

Abstract

The risk of prostate cancer (PCa) can be described using five risk categories based on clinical assessment involving the risk of cancer and metastasis level. In some cases the information needed to calculate the risk is not registered. The aim is to assess potential differences in properties like age, treatment, comorbidity and survival, between men with a defined risk stage categorisation of prostate cancer compared to men lacking information to calculate the risk stage category. The main measures involved in the risk stage assessment are Gleason score, prostate specific antigen (PSA) value and T-stage. Men missing data for risk stage categorisation may be lacking one of the three or a combination of at least two out of the three. Subgroups of these men will be analysed in a similar way, in order to understand the reasons to why they are missing data for risk stage categorization.

Statistical analysis involves univariate and multivariate logistic regression, along with survival and competing risk analysis. Data will be presented in tables, figures and forest plots, including odds ratios, 95% confidence intervals and p-values.

According to the study, men missing data of risk stage categorization were about 2.6% of all men in the data base, and had most likely a low risk PCa. They had higher comorbidity levels but the overall probability of death was the same compared to other men. In addition, they had significantly lower proportion of death by PCa and experienced a large proportion of death by other cancer, which concurs with the previous conclusions about comorbidity and low risk PCa, indicating that they had another disease, possibly cancer, that required more attention.

Considering only men missing data for risk categorisation, a large proportion were missing PSA level (58.3%), and these men had higher comorbidity, were older, and had a large proportion of death by other cancer. Surprisingly, men missing Gleason level (20.3%) had increased odds ratios for lower comorbidity levels, were younger at time of diagnosis, and had a higher survival probability in general. Unexpectedly, men missing T-stage (32%) were more likely to being treated by Radio Theraphy (RT), were less likely to attend university hospitals and more likely to attend private physicians. Men missing a combination of at least two out of three of Gleason, PSA, T-stage (19.6%) had higher comorbidity levels and were more likely to be treated by RT, less likely to attend university hospitals, had a large proportion of death by other cancer, and a larger proportion of death closer to the time of diagnosis.

Lastly, there were some indications of variations of the proportions of missing data of risk stage categorisation when dividing it into the subgroups mentioned above, and viewed over year of diagnosis. There was an increase in missing data of risk stage categorization around 2006 and an explanation of this could be the change of IT-system for registration, leaving a general increase of missing data behind it, perhaps due to a looser control during the transition and unfamiliarity of the new system.

The main conclusion was that the reasons for missing data of risk stage categorisation are most likely high comorbidity levels, probably including another cancer in combination with a low risk PCa. It was most common to be missing data of risk stage categorisation due to missing PSA level, and those men had high comorbidity and were older. Surprisingly, private physicians and/or treatment by RT were more likely to be missing T-stage, and younger men with low comorbidity were more likely to be missing Gleason score.

Foreword and Acknowledgements

Special acknowledgements to my supervisors Rolf Larsson, Uppsala University, guidance and advice, and Hans Garmo, Regional Cancer Centre Uppsala Örebro, for leadership and for making this thesis possible. Additional thanks to Yasin Folkvaljon, Regional Cancer Centre Uppsala Örebro, for feedback and support.

Nomenclature

Medical Terms

AA	Antiangrogens
CCI	Charlson's Comorbidity Index
GnRH	Gonadotropin-releasing hormone
LISA	Longitudinal Integration Database for Health Insurance and Labour Market Studies
NPCR	National Prostate Cancer Register of Sweden
PCa	Prostate Cancer
PCBaSe	Prostate Cancer Data Base Sweden
PSA	Prostate-Specific Antigen
RP	Radical Prostatectomy
RT	Radio Therapy
WHO	World Health Organization grading scheme

Mathematical Notation

$\mathbf{X} = (X_1, \ldots, X_k)$	a vector of k explanatory variables
X_i	explanatory variable
$\mathbf{Y} = (Y_1, \ldots, Y_k)$	a vector of k response variables
Y_i	response variable
$\hat{\theta}$	estimate of θ
CI	Confidence Interval
OR	Odds Ratio

Contents

1 Introduction

When assessing the risk stage of prostate cancer (PCa) five risk categories based on clinical assessment are considered. The first three consider the risk of cancer localised to the prostate and the two later describe the metastasis level of the cancer. In some cases the information needed to calculate the risk is lost or not listed, and the men with insufficient information are considered as missing data of risk stage categorisation, Sandin and Wigertz (2013). According to this definition of risk categories men can be missing data of risk stage categorisation due to missing specified Gleason score, PSA level and/or T-stage. The reasons may be different and little is known about these men, especially if they are distinguishable as a group, and in that case how and why, compared to men with a specified risk stage. The objective is to understand why men are missing data for risk stage categorisation, in order to maintain quality of records and possibly for improvement of quality on different levels.

1.1 Purpose

The aim of the thesis is to understand why men are missing data of risk stage categorisation. In order to do this a statistical analysis of the men with missing data of risk stage categorisation will be done to describe the group, and to assess potential differences between them and men that do not miss data for risk categorisation. In addition, subgroups of men missing data for risk categorisation will be considered and compared against each other to understand whether subgroups differ regarding missing data of risk stage categorisation and if there are any patterns. The analysis will consider properties such as age at time of diagnosis, treatment, survival after diagnosis and comorbidity.

The results from the study could be used for further analysis of the data, data quality assessment and potentially quality assessment of the registration process for hospitals or regions.

1.2 Hypotheses

For each property i that is possible to assess via the data, like comorbidity, age when diagnosed, survival after diagnosis, or treatment:

H_{i_0} : no difference between men missing data and other men for property i

H_{i_1} : difference between men missing data and other men for property i

In addition, when investigating men missing data for risk categorisation, there are subgroups of interest that are missing data of different reasons, for example only due to missing PSA, or subgroups with other properties. For these we have the hypotheses:

H_{jk_0} : no difference between the subgroups j and k, $j \neq k$

H_{jk_1} : difference between the subgroups j and k, $j \neq k$

1.3 Expectations and Challenges

Due to the large number of observations it is expected that the statistical analysis will be reliable, but it may be that the potential differences are difficult to spot for other reasons. For example, it will be a challenge, at least initially, to choose which variables to focus on.

A personal challenge, except from the required computer coding, will be to be a part of the statistical reasoning and decision making of the process, and to interpret the results.

1.4 Delimitations

The study is limited to the time period 1998-2012 when NPCR was nation wide, even though there is more data available, Hemelrijck (2013). The main reason is that not all data is available during earlier time periods when NPCR was not nation wide.

1.5 Method

Initially, data collected by several relevant institutions and gathered in a database called PCBaSe, Hemelrijck et. al. (2013), will be accessed, in comma-separated values format (CSV). It will be presented in a table displaying the covariates of interest, splitting them over columns to explore subsets of the men. The choice of relevant covariates and subsets will depend on the results revealed during the study. This will give an overview of the data and may expose information about which hypotheses that should be formulated for later analysis.

After that, logistic regression, both univariate and multivariate, will be used to produce odds ratios and confidence intervals, indented to reveal potential differences in distribution between groups defined in the previous step.

Then survival analysis in terms of Kaplan-Meier estimation of survival curves, both for overall survival and the hypothetical survival where death by PCa is the only cause of death, and the other competing risks are considered as censored. Competing risk analysis in terms of cumulative incidence will be used to compare survival, divided into different causes of death, for the different subgroups looked at in the first stage. The competing risks are death by prostate cancer, other cancer, cardiac/vascular and other causes. Survival time is defined as the time from diagnosis until death, emigration or end of follow-up on 31 December 2012, whichever event came first.

The level of statistical significance is set to 0.05. Statistical computer software used: R version 3.1.2 (http://www.r-project.org/)

1.6 Conclusions

Men missing data for risk stage categorisation had generally high comorbidity, were more likely to die by other cancer and less likely to die by prostate cancer, and had most likely a PCa that could have been categorised as low risk. It is most common to be missing data due to missing PSA value (possibly assessed but not registered), and those men had high comorbidity and were older. Surprisingly, private physicians and/or treatment by RT are more likely to be missing T-stage (possibly assessed but not registered), and younger men with low comorbidity are more likely be missing Gleason score (probably not always assessed). There were some variations of the proportion of missing data over year of diagnosis and indications of a connection between missing data of risk stage categorisation, age and year of diagnosis. The variations seen when looking at the proportions of missing data over time could partly be explained by the IT-system change of registration.

2 Background

The risk of prostate cancer is described using five risk stage categories based on clinical assessment. Men can be missing data for risk categorisation due to missing specified Gleason score, PSA value and/or T-stage. These three measures are taken using different methods and a short description follows below.

2.1 Medical Procedures

2.1.1 Measures Used for Risk Categorisation

A Gleason grade is obtained with a needle core biopsy taken from the prostate. The procedure is performed by entering the rectum with a machine equipped with 6-12 hollowed needles, and then inserting the needles into the prostate to extract a tissue sample containing cells from the prostate. With some luck, the sample contains cells from the part that contains cancer. Then the pathologist evaluates the level of cell mutations in the cancer cells and uses this to assess the severity of the cancer, which is measured in Gleason grades (1-5). Then Gleason score is calculated by summing the most common Gleason grade and the highest Gleason grade. The procedure might be painful and there is a risk of side effects, of which the most common severe side effect is infection, Sandin and Wigertz (2013). The World Health Organisation differentiation grade (WHO) is obtained by fine needle aspirates from the prostate, Stattin et al. (2013). Around 2000 there was a change towards using the Gleason grading system instead of the previously used WHO grading, Hemelrijck et. al. (2013).

To retain a PSA value an ordinary blood sample is taken and the level of prostate specific antigen (PSA) in $\mu g/L$ is measured.

The T-stage is assessed by the urologist, who inserts a finger in the rectum to examine the part of the prostate that is accessible. If it is possible to feel tumours growing outside of the prostate and into the vas deferens the stage is classified as T4. If the tumour is outside the capsule it is classified as T3, if there are lumps that seem to be inside the capsule it is classified as T2, and if no lumps are found it is classified as T1. Tumours classified as T1 and discovered by transurethral resection of the prostate (TURP), which basically is a reduction of the prostate by planing to simplify micturition, are classified as T1a or T1b depending on the amount of cancer cells found in the planed sample. If not discovered by TURP it is classified as T1c., Sandin and Wigertz (2013).

2.1.2 Other Variables

Of the stages in the TNM classification of malignant tumours T-stage is the primary stage used in the risk categorisation, but the other stages are also of importance. N-stages describe the stages of regional lymph nodes, where N0 equals no signs of regional lymph node metastases, N1 equals signs of region lymph node metastases and NX equals not possible to assess. Similary, the M-stages describe distant metastases, where M0 equals no sings, M1 equals signs, and MX equals not possible to assess (available option only before 2011), Stattin et al. (2013).

In addition, there are several treatment methods for the prostate cancer, for example Radical Prostatectomy (RP) which is a surgery applied to any patient with clinically localized prostate cancer that can be completely excised surgically, who has a life expectancy of at least 10 years, and has no serious comorbid conditions that would contraindicate an elective operation.

Another example is radio therapy (RT). Brachy therapy is a kind of radio therapy that involves placing radioactive sources into the prostate tissue. External radio therapy uses external beams and is one of the principle treatment options for clinically localized prostate cancer.

Active surveillance involves actively monitoring the course of disease with the expectation to intervene if the cancer progresses, with a risk of missing the opportunity for cure, and getting pro-

gression and/or metastases. Similarly, watchful waiting is used with the intention of symptomatic therapy at progression in men with PCa.

Lastly, hormonial treatments may include antiangrogens (AA) or Gonadotropin-releasing hormone (GnRH), which are used for retaining an incurable PCa, Mohler et al. (2010).

Mode of detection indicates the main reason of how the PCa was revealed, where symptomatic includes lower urinary tract symptoms (LUTS), other symptoms such as hematuria, or remote symptoms such as back pain when having distant metastases or other cancer, and non-symptomatic indicates health assessment including PSA testing and without urinary tract symptoms, Stattin et. al. (2013).

The comorbidity level, which predicts the ten-year mortality for a patient who may have a range of comorbid conditions, is measured by Charleson's Comorbidity Index (CCI), which assigns weights to 17 medical conditions, including diabetes and hypertension, allowing for a final comorbidity score to be calculated for each individual. Each condition is assigned a score of 1, 2, 3, or 6, and the final CCI score is given as the sum of these scores. Individuals are grouped into CCI categories for nal scores of 0,1, 2, or 3+. The prostate cancer is not included, and one is not given a CCI level of 6 for metastatic PCa. CCI levels indicate 0 = no comorbidity, 1 = mild comorbidity, 2 = medium comorbidity, and 3+ = severe comorbidity, Charlson et. al. (1987).

2.2 Data Collection

2.2.1 PCBaSe and NPCR

Data was retrieved from the Prostate Cancer Data Base Sweden (PCBaSe), Hemelrijck et. al. (2013), which is a nationwide population-based research database of men diagnosed with prostate cancer, and considered all records between 1998 and 2012, at total of 129 389 men. PCBaSe was created by record linkage between the National Prostate Cancer Register of Sweden (NPCR) and several of other national registers, like the Prescribed Drug Register, In-Patient Register, Cause of Death Register, National Population Register and the Longitudinal Integration Database for Health Insurance and Labour Market Studies (LISA).

NPCR is a tool for documentation and assessment of quality of the health care of men with prostate cancer. It is also used for research and to improve the treatment of the disease. The register contains data including diagnosis procedure, spread and type of cancer, waiting times, treatment and its outcome. The participation in the register is voluntary and it is not possible to track or identify specific individuals in the compiled material.

2.2.2 Data Retrieved from PCBaSe

For all men information were retrieved on age, highest educational level[1], year of diagnosis, diagnosis performed by private or public physician, diagnosis performed at university hospital or not, and mode of detection, as well as survival times and censoring indication.

In addition, clinical data were retrieved on serum levels of PSA, tumour differentiation by Gleason score/WHO-grade[2], cancer stage according to the tumour-node-metastasis classification, and primary treatment. Men were classified into five prostate cancer risk categories. These were defined similar to the National Comprehensive Cancer Network (NCCN) guidelines, Mohler et. al. (2010) but altered to distinguish between regionally metastatic and distant metastatic disease.

- Low-risk localized prostate cancer was defined as clinical local stage T1/T2, Gleason score 6 or below and PSA less than 10 $\mu g/ml$

[1] low (compulsory school, $\leq$9 years), middle (upper secondary school, 10 - 12 years), high (college and university, $\geq$13 years)

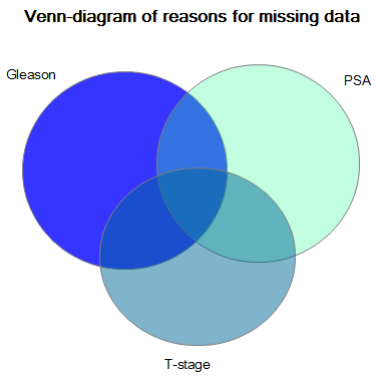
[2] referred to as Gleason score only

- Intermediate risk localized prostate cancer was defined as stage T1/T2, Gleason score[3] 7 and/or PSA level of 10 - 20 $\mu g/ml$
- High-risk localized prostate cancer was defined as stage T3 and/or Gleason score 8 - 10 and/or PSA levels of 20 - 50 $\mu g/ml$
- Regionally metastatic or locally advanced prostate cancer was defined as stage T4 and/or N1 disease and/or PSA levels of 50 - 100 $\mu g/ml$ without distant metastasis (M0 or MX disease)
- Distant metastatic disease was defined as M1 disease and/or PSA at least 100 $\mu g/ml$

The comorbidity burden was assessed using data from the National Patient Register and classified according to Charlsons comorbidity index (CCI). Data on treatment modalities included radical prostatectomy, radiotherapy, and hormonal therapy. Information on alternative treatment approaches like active surveillance and watchful waiting was also obtained.

2.2.3 Subsets of Data

In some analyses men missing data for risk stage categorisation will be divided into subsets according to the following scheme



- Missing data of risk stage categorisation
- Not missing data of risk stage categorisation
- Missing data of risk stage categorisation due to missing Gleason score
- Missing data of risk stage categorisation due to missing PSA value
- Missing data of risk stage categorisation due to missing T-stage
- Missing data of risk stage categorisation due to combinations of 2-3 of missing Gleason score, PSA value and T-stage

The choice of these subsets is motivated by the definition of the risk stage categories above, and the subsets will be called reasons for missing data of risk stage categorisation. They are by definition overlapping.

2.3 Previous Research

No previous research has been done that covers the same hypotheses and data material. The most relevant study is *Capture rate and representativity of The National Prostate Cancer Register of Sweden*, Tomic, et. al. (2015), which concludes that the data in NPCR is generalizable to all men in Sweden with prostate cancer.

[3]Beginning in 2007, data were available to further categorize Gleason 7 cancers into 3 + 4 versus 4 + 3

3 Theory

Unless stated otherwise, the sections 3.1-4 and 3.5.1 are based on material from Dobson (2002), and the remaining part of section 3.5 is based on material from Kleinbaum et. al. (2010).

3.1 Introduction

The theory upon which the statistical analysis is used in this thesis involves the analysis of relationships between measurements made on groups of subjects. The terms response, outcome or dependent variable are used for those variables that may vary in response to other variables that are called explanatory, predicator or independent variables. Products of independent variables are called interaction terms. It is important to note that the responses are regarded as random variables and the predictors are regarded as non-random measurements or observations. The dependent and independent variables may be measured in several ways.

1. Nominal - categories, like yes/no or colours. If there are two categories then the variable is called dichotomous. If there are several categories the variable is called polychotomous.
2. Ordinal - categories where there is some natural ordering, like age grouped into intervals.
3. Continuous - measurements that may attain any real value, on a specific interval, like time or length.

Nominal and ordinal variables are called categorical or qualitative and continuous variables are often called quantitative. A qualitative explanatory variable is called a factor and its categories are called levels.

Before exploring logistic regression some knowledge of the fundamental theory concerning the exponential family, inferences and tests is needed. The exponential family is a family of distributions with certain properties with many important implications. Most relevant in this particular case is the occurrence of the exponential family in the definition of the Generalised linear model, of which the logistic regression is a special case.

3.2 The Exponential Family

Definition 1. Let Y be a random variable with distribution function $p(y;\theta)$ depending on the single parameter θ. Then we say that the class of probability measures $\mathbb{P} = \{P_\theta : \theta \in \Theta\}$ is an exponential family if

$$p(y;\theta) = t(\theta)\exp[b(\theta)a(y)]s(y)$$

where the functions a, b, s, t are known real valued functions, $a(y)$ is a statistic, $s(y) \geq 0$, $t(\theta) > 0$. We call $b(\theta)$ the natural parameter. If more parameters, other than θ are present then we call those the nuisance parameters. If $a(y) = y$ then the distribution is said the be in canonical form.

Example 1. Let $Y \sim Bin(n, \theta)$ then

$$p(y,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = (1-\theta)^n \exp[\ln(\frac{\theta}{1-\theta})y]\binom{n}{y} \quad , \quad \theta \in (0,1)$$

and hence the class of all binomial distributions belong to an exponential family with natural parameter $b(\theta) = \ln(\frac{\theta}{1-\theta})$, $t(\theta) = (1-\theta)^n$, $s(y) = \binom{n}{y}$ and $a(y) = y$, i.e. the distribution is in canonical form.

3.3 Generalised Linear Models

In order to properly define the logistic model and its background some understanding of generalised linear models is needed. Let $Y_1, \ldots, Y_N$ be a set of independent random variables whose distributions belong to the same exponential family in canonical form

$$f(y_i; \theta_i) = t(\theta_i) \exp[y_i b(\theta_i)] s(y_i)$$

then the joint density function is of the form

$$f(y_1, \ldots, y_N; \theta_1, \ldots, \theta_N) = \prod_{i=1}^{N} t(\theta_i) \exp[y_i b(\theta_i)] s(y_i) = \exp[\sum_{i=1}^{N} \ln(t(\theta_i)) + \sum_{i=1}^{N} \ln(s(y_i)) + \sum_{i=1}^{N} y_i b(\theta_i)]$$

When the parameters θ_i are not of direct interest there may be a smaller set of parameters $\beta_1, \ldots, \beta_p$, $p < N$ used as coefficients in a sum of independent variables, as will be further discussed below.

When considering linear regression models the mean of the distribution of the outcome variable is often a key property to model, and is usually what we use our data to describe. To connect the data and the parameters of interest with the mean we use a special function called the link function.

Definition 2. Let $E(Y_i) = \mu_i$ and x_i a column vector of explanatory variables and β a vector of the parameters of interest β_i. Let $g(\mu_i)$ be a monotone differentiable function such that $g(\mu_i) = x_i^T \beta$, then we call g the link function.

Definition 3. Generalized Linear Model Let $\mathbf{Y} = (Y_1, \ldots, Y_N)$ be a set of independent random variables whose distributions belong to the same exponential family in canonical form, $\beta = (\beta_1, \ldots, \beta_p)$ a set of parameters and $\mathbf{X} = (X_1, \ldots, X_p)$ a set of explanatory variables such that

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{N1} & \ldots & x_{Np} \end{pmatrix}$$

In addition, let $E(Y_i) = \mu_i$ and g be a link function such that $g(\mu_i) = x_i^T \beta$. Then we call $\{\mathbf{Y}, \mathbf{X}, \beta, g\}$ the generalized linear model. Hence the model is primarily described by the distribution and mean structure. Usually each row of $\mathbf{X}$ represents an individual and each column represents an independent variable.

3.4 Inferences

Confidence intervals for, and testing hypotheses about, parameters are two important concepts when performing regression. Then knowledge of the sampling distribution of the particular statistic in use is crucial, but first we need some basic concepts about inferences.

The material below, until and including the Fisher information matrix, is based on material from Zwanzig and Liero (2012).

Definition 4. We call the function $L(\theta; \mathbf{X}) = P(\mathbf{X}; \theta)$ the likelihood function of the parameter $\theta = (\theta_1, \ldots, \theta_k)$. The likelihood function is the joint probability of observing the data, and therefore the maximum of the likelihood function gives the preferred estimate of θ, denoted $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$, and is called the maximum likelihood estimate (MLE) of θ.

Remark 1. The maximum is found by solving the system of equations $\frac{\partial L}{\partial \theta_i} = 0 \quad \forall j \in 1, \ldots, k$. It is common to maximize the logarithm of the likelihood function instead, which gives the same estimates and is denoted as $l(\theta; \mathbf{X}) = \ln(L(\theta; \mathbf{X}))$ and called the log-likelihood function.

Usually it is not only the value of the point-wise estimation of a parameter that is of interest, but also an estimation of the variance. In the general case, for several parameters, we use the covariance matrix to describe this.

Definition 5. We call the matrix

$$\mathbf{V}(\hat{\theta}) = [v_{ij}] = cov(\hat{\theta}_i, \hat{\theta}_j)$$

the covariance matrix of the estimator $\hat{\theta}$. It contains the covariances between parameters, and also the variances, which are found on the diagonal. The standard error $s.e$ of the estimator $\hat{\theta}_i$ is $s.e_{\hat{\theta}_i} = \sqrt{Var(\hat{\theta}_i)}$

Definition 6. Regularity Conditions 1-4 In order to define the following notions and to be able to conclude some general results regarding the test statistics below we need to impose some regularity conditions.

Reg 1: the distributions $\{P_\theta, \theta \in \Theta\}$ have common support $\mathbb{A}$

Reg 2: the parameter space $\Theta \subseteq \mathbb{R}^p$ is an open set

Reg 3: $\forall x \in \mathbb{A}$ the likelihood function has finite partial derivatives

Reg 4: $\forall x \in \mathbb{A}$ the likelihood function has second partial derivatives and $\forall \theta \in \Theta$ the order of integration (and summation) and differentiation of second order is interchangeable, for all second partial derivatives.

Definition 7. The gradient of the log-likelihood function describes the relative rate at which the likelihood function changes with respect to the parameters. Under regularity conditions 1-3 the vector of partial derivatives of the log-likelihood function is called the score function and is denoted as

$$S(\theta; y) = (\frac{\partial}{\partial \theta_1} l(\theta; y), \ldots, \frac{\partial}{\partial \theta_p} l(\theta; y))^T$$

Regularity conditions 1-3 imply that the order of integraton (and summation) and differentiation can be interchanged. Using this one may prove the following.

Proposition 1. Under regularity conditions 1-3 then

$$E[S(\theta; Y)] = 0 \quad \forall \theta \in \Theta$$

The information contained in the data can be measured with the Fisher information matrix defined below. When the sample size increases the information increases.

Definition 8. Under regularity conditions 1-3 the matrix $I_{\mathbf{Y}}(\theta) = Cov(S(\theta; \mathbf{Y}))$ is called the Fisher information matrix.

Now, assume that the response variables Y_i $i \in 1, \ldots, N$ are independent random variables in a generalised linear model. If one imposes regularity condition 4 it is possible to prove the following.

Proposition 2. Under regularity conditions 1-4, with $J(\theta, \mathbf{Y})$ being the matrix with elemtents $J(\theta, \mathbf{Y})_{j,k} = -\frac{\partial^2 l(\theta; \mathbf{Y})}{\partial \theta_j \partial \theta_k}$, the Fisher information matrix may be written as

$$I_{\mathbf{Y}}(\theta) = E[J(\theta, \mathbf{Y})]$$

Lemma 1. Under regularity conditions 1-4 then the following approximation holds

$$\mathbf{S}(\theta, \mathbf{Y}) \approx \mathbf{S}(\hat{\theta}, \mathbf{Y}) - I(\hat{\theta})(\theta - \hat{\theta})$$

Proof: Using the previous proposition, approximating the second derivative with the Fisher information matrix, a Taylor series expansion of the log-likelihood function near the estimate yields the approximation.

Proposition 3. Under certain regularity conditions, including 1-3, the maximum likelihood estimators are consistent, asymptotically normal and efficient, Zwanzig and Liero (2009).

Thus it is possible to approximate the distribution of MLE with the normal distribution, given that it satisfies the regularity conditions. This approximation becomes better for larger sample sizes. From now on we assume that the MLE in consideration satisfies the regularity conditions.

Proposition 4. The Wald Statistic The statistic $(\hat{\theta}-\theta)^T I(\hat{\theta})(\hat{\theta}-\theta)$ is approximately $\chi^2(p)$ distributed, where p is the number of parameters considered, i.e the length of θ. For the one parameter case $\hat{\theta} \sim N(\theta, I(\hat{\theta})^{-1})$ approximately.

Proof: By definition $\hat{\theta}$ maximizes $l(\theta)$ so $\mathbf{S}(\hat{\theta}; \mathbf{Y}) = 0$. Hence

$$\mathbf{S}(\theta; \mathbf{Y}) = -I(\hat{\theta})(\theta - \hat{\theta})$$

by lemma 1, if I is invertible. If we regard $I(\hat{\theta})$ as a constant then $E[\hat{\theta}-\theta] = I^{-1}(\hat{\theta})E[\mathbf{S}] = 0$, by the proposition 1, so $\hat{\theta}$ is at least asymptotically consistent. Now, since $I(\hat{\theta})$ is symmetric $(I^{-1}(\hat{\theta}))^T = I^{-1}(\hat{\theta})$ and $I(\hat{\theta}) = E[SS^T]$ the covariance matrix for $\hat{\theta}$ is

$$E[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T] = I^{-1}(\hat{\theta})E[\mathbf{S}\mathbf{S}^{\mathbf{T}}]I = I^{-1}(\hat{\theta})$$

Hence, by the previous proposition,

$$(\hat{\theta}-\theta)^T I(\hat{\theta})(\hat{\theta}-\theta) \sim \chi^2(p) \quad \textit{approximately}$$

and for the one-parameter case

$$\hat{\theta} \sim N(\theta, I^{-1}(\hat{\theta})) \quad \textit{approximately}$$

Definition 9. A model that contains the maximum amount of parameters that can be estimated is called a saturated model, or sometimes the reference model, or the ideal model.

Remark 2. Let m be the maximum number of parameters that can be estimated in the model, and let θ_{max} be the vector of parameters in the saturated model, and $\hat{\theta}_{max}$ its estimator. Then $L(\hat{\theta}_{max}, \mathbf{y})$ will be larger than any other likelihood function for these observations and provides the most complete description of the data. If $L(\hat{\theta}, \mathbf{y})$ is the maximum value of the likelihood function corresponding to the model of interest, then the likelihood ratio

$$\lambda = \frac{L(\hat{\theta}_{max}, \mathbf{y})}{L(\hat{\theta}, \mathbf{y})}$$

provides a way of assessing the goodness of fit for the model.

Definition 10. We call

$$D = \ln(\lambda) = 2(l(\hat{\theta}_{max}; \mathbf{y}) - l(\hat{\theta}; \mathbf{y}))$$

the deviance, or log-likelihood ratio statistic.

Proposition 5.

$$D = \ln(\lambda) = 2(l(\hat{\theta}_{max}; \mathbf{y}) - l(\hat{\theta}; \mathbf{y}))$$

is approximately $\chi^2(m-p, v)$ distributed, where v is the non-centrality parameter which is close to zero if the model fits the data approximately as good as the saturated model, and m is the number of parameters in the saturated model and p is the number of parameters in the model of interest.

3.5 Logistic Regression

3.5.1 The Model

The data that we want to fit a model to is in its essence a vector $\mathbf{Y}$ of dichotomous variables, usually coded as $Y_j = 1$ if outcome $j \in \{1, \ldots, n\}$ does not have the disease or is a success, and $Y_j = 0$ if it has the disease or is a failure. We may assume that the Y_j are independent $\sim Ber(\pi_j)$ for each j. Then the probability function is $\pi_j^{y_j}(1-\pi_j)^{1-y_j}$ which belongs to an exponential family with natural parameter $ln(\frac{\pi_j}{1-\pi_j})$ and expectation $E[Y_j] = \pi_j$.

We want to know the probabilities π_j but we only have one observation of each individual. To solve this we assume that $\pi_j = \pi$ for all j, then we have more observations for π. Secondly, we want to use explanatory variables and model π, but to fit a line to this kind of data would in general not be a good idea. Instead we use a transformation of the data to model the expected value of the Y's using a special case of a general linear model, the logistic model.

First, $E[Y_j] = \pi$ so we are modelling the probabilities π with the link function g such that $g(\pi) = x_j^T\beta$. Now we have an expression which is linear in the coefficients, expressing a function of π using the explanatory variables. If g is an appropriate link function, then we have a logistic regression model, which also is a general linear model $\{\mathbf{Y}, \mathbf{X}, \beta, g\}$, and hence we may use the theory of the previous chapter regarding inferences. But first we explore the contents of the logistic model.

Remark 3. If some x_i are factors then they may be decoded as a sum of indicators, "dummy" variables. In the continuous case the variable may first be discretised and coded in an appropriate manner.

Definition 11. To be sure that π is restricted to $[0, 1]$ it is often modelled using a cumulative probability distribution $\pi(t)$. So let f be a function such that

$$\pi(t) = \int_{-\infty}^{t} f(s)ds$$

where $f(s) \geq 0$ and $\int_{-\infty}^{\infty} f(s)ds = 1$. Then we call the probability density function $f(s)$ the tolerance distribution.

Remark 4. For simplicity, assume that we have only one explanatory variable x, such that $g(\pi) = \beta_1 + \beta_2 x$. As we shall see later, we want a linear model in log-odds. Therefore the tolerance distribution for the logistic regression model has density

$$f(s) = \frac{\beta_2 \exp(\beta_1 + \beta_2 s)}{[1 + \exp(\beta_1 + \beta_2 s)]^2}$$

and if we consider π as a function of x then

$$\pi(x) = \int_{-\infty}^{x} f(s)ds = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)}$$

which gives the link function

$$g(\pi(x)) = ln(\frac{\pi(x)}{1 - \pi(x)}) = \beta_1 + \beta_2 x$$

Definition 12. The link function g above is called the logit function,

$$\text{logit}P(x) = \beta_1 + \beta_2 x$$

and in general we write

$$\text{logit}\pi(\mathbf{x}_i) = ln(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}) = \mathbf{x}_i^T\beta$$

Proposition 6. Rewriting $\pi(x)$ using the logit transformation we obtain

$$\pi(x) = \frac{1}{1 + e^{-(\beta_1 + \beta_2 x)}}$$

which we call the logistic function.

Proposition 7. If we write $\text{logit}(P(X)) = \alpha + \beta_1 X_1 + \dots \beta_k X_k$, we may obtain the function

$$P(Y = 1|X_1, \dots, X_k) = \frac{1}{1 + e^{-(\alpha + \sum_{i=1}^{k} \beta_i X_i)}}$$

The coefficient α is called the intercept, α and β_i are unknown parameters, and we will, for simplicity, call it the logistic model.

Remark 5. The logit transformation of the logistic model is

$$\text{logit}P(X) = \ln(\frac{P(X)}{1 - P(X)}) = \alpha + \sum_{i=1}^{k} \beta_i X_i$$

Essentially we have a linear model in the log-odds, as defined below.

3.5.2 Odds Ratios

Odds contain information about whether the successful event is more probable or not, compared to the unsuccessful event. Odds ratios are used to compare if a state of an explanatory variable is more probable to cause a successful event than another state. This may be used to reveal structural differences in distributions, as we will see later.

Definition 13. Let π be a probability, then we call the ratio $\frac{\pi}{1-\pi} = Odds$ and by rewriting we may obtain the probability $\pi = \frac{Odds}{1+Odds}$. For the logistic model the odds is defined as

$$O(X) = \frac{P(X)}{1 - P(X)}$$

which describes the risk for individual X.

Definition 14. When we want to compare the odds of two groups of individuals, for example $X_1 = (X_{1,1}, \dots, X_{1,k})$ and $X_2 = (X_{2,1}, \dots, X_{2,k})$, then we may compute the ratio of their corresponding odds, the odds ratio

$$OR_{X_1, X_2} = \frac{O(X_1)}{O(X_2)}$$

Now we know that that the logit function describes log odds for individual X and that the odds satisfy

$$O(X) = \frac{P(X)}{1 - P(X)} = e^{\alpha + \sum_{i=1}^{k} \beta_i X_i}$$

Using this, the odds ratio between two groups in the logistic model becomes

$$OR_{X_1, X_2} = \frac{O(X_1)}{O(X_2)} = e^{\alpha + \sum_{i=1}^{k} \beta_i X_{1,i} - (\alpha + \sum_{i=1}^{k} \beta_i X_{2,i})} = e^{\sum_{i=1}^{k} \beta_i (X_{1,i} - X_{2,i})}$$

Hence each $X_{j,i}$ contributes to the odds ratio in a multiplicative way. Assume all components of X_1, X_2 are the same, except for $X_{1,j}, X_{2,j}$ for some j, then the odds ratio describes the odds of having the property $X_{1,j}$ versus $X_{2,j}$ when all other terms are fixed. In this case the only surviving coefficient will be β_j. e^{β_j} is called the adjusted odds ratio since it adjusts for the effects of the other variables which were held fixed.

Example 2. Consider a simple model where X is a variable representing individuals with high and low income, attaining $X = 0$ for low and $X = 1$ for high. Let the outcome variable Y represent if the individual is ill (1) or not ill (0). The logistic model will be

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\alpha+\beta X)}}$$

with corresponding $\text{logit}P(\mathbf{X}) = \alpha + \beta X$. If a, b, c, d represent the frequencies of individuals in respective category, then we may obtain the following table.

	$X = 1$	$X = 0$
$Y = 1$	a	b
$Y = 0$	c	d

Now it is possible to calculate the odds of being ill ($Y = 1$) versus not ill ($Y = 0$)

$$X = 1 \text{ gives } \frac{\hat{P}(Y = 1|X = 1)}{\hat{P}(Y = 0|X = 1)} = \frac{a}{c} \qquad X = 0 \text{ gives } \frac{\hat{P}(Y = 1|X = 0)}{\hat{P}(Y = 0|X = 0)} = \frac{b}{d}$$

This gives the odds ratio $\hat{OR}_{X=1,X=0} = \frac{ad}{bc}$. Using the logistic model we would instead obtain

$$\hat{OR}_{X=1,X=0} = \frac{\hat{O}(X = 1)}{\hat{O}(X = 0)} = e^{\hat{\alpha}+\hat{\beta}X_1-(\hat{\alpha}+\hat{\beta}X_0)} = e^{\hat{\beta}}$$

Example 3. Inspired by Sainani (2014).

The general interpretation of the coefficients is that α is the "baseline" log odds, where baseline refers to a model that is empty, i.e. ignores all $X's$. The coefficients β_i represent the change in the log odds, that would result from a one unit change in the variable X_i, when all other X_i's are fixed. For example, if we are given a model in logit form with estimated parameters

$$\text{logit(T)} = -1.1 + 0.5 \cdot \text{H} + 1.5 \cdot S$$

Where $T =$ (1 if an individual passed the test, 0 if not), $H =$ number of studying hours and $S =$ gender (1 if female, 0 if male). Then if $\hat{OR}_{\text{women vs men}} = e^{1.5} = 4.48$ it means that women more than four times higher odds for passing the test than men, and $OR_{\text{Studying hours}} = e^{0.5} = 1.65$ means that for each additional studying hour the odds for passing the test increases with about 65 %.

Sometimes we want to compare odds ratios for a given factor's levels to find which are more likely to cause a success for the outcome. To compare odds ratios in a structured way a reference is chosen, that is one of the factor levels to which all other odds ratios are computed against. This will make all odds ratios comparable and the analysis intuitive.

Definition 15. Let $X = (X_1, \ldots, X_k)$. Without loss of generality, assume X_1 is the reference, then the odds ratios will be

$$OR_j = \frac{O(X_j)}{O(X_1)} \quad \text{and naturally} \quad OR_1 = \frac{O(X_1)}{O(X_1)} = 1$$

So the reference odds ratio is by definition always 1.

Remark 6. When performing multiple logistic regression each odds ratio is computed with respect to the reference for the specific factor. In the example above this would produce two references, one for the odds ratios for the factor S and one for the factor G, and the odds ratios are considered as adjusted, where odds ratios for one factor are adjusted for the confounding effect of the other factors, called confounders.

3.5.3 Testing Hypotheses and Confidence Intervals

A natural question would be how the coefficients in the logistic model are estimated. The answer is the maximum likelihood method, see below. The tests that follow after this are special cases of the tests mentioned in the section about inferences and generalised linear models.

Definition 16. If n is number of observations, where P is as defined above, m_1 is the number of diseased and $n - m_1$ is the number of not diseased, then the likelihood function, when the variables are few compared to the number of observations, is

$$L = \prod_{i=1}^{m_1} P(X_i) \prod_{j=m_1+1}^{n} (1 - P(X_j))$$

Proposition 8. The likelihood function above satisfies the regularity conditions and hence the maximum likelihood estimators (MLE) obtained are consistent, asymptotically normal and efficient, Rashid, Shifa, (2009).

Proposition 9. Wald's Test To test the hypothesis $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$ we use the the test statistic $Z = \frac{\beta_i}{s.e_{\beta_i}}$ which is approximately $N(0, 1)$ distributed, or Z^2 which is approximately χ^2 distributed with 1 df, where $s.e_{\hat{\beta}_i} = \sqrt{Var(\hat{\beta}_i)}$. The asymptotic normality follows from the proposition above.

Definition 17. Likelihood Ratio Test To test the hypothesis $H_0 : \beta_k = 0 \; \forall \; k \in K$, for some subset K of the indices of the coefficients, vs $H_1 : \beta_k \neq 0$, i.e to compare the reduced model against the full model, we compute the likelihood ratio of the models $-2\ln(\frac{\hat{L}_{red}}{\hat{L}_{max}})$. This test statistic is approximate χ^2 with $m - p$ degrees of freedom if n is large, by proposition 16 and above, where m is the number of parameters in the full model and p is the number of parameters in the reduced model.

Definition 18. Confidence Intervals To estimate intervals of the parameter estimates we proceed more or less as usual. The $100(1 - \alpha)\%$ confidence interval for one parameter β_i is defined as

$$\hat{\beta}_i \pm \lambda_{1-\frac{\alpha}{2}} s.e_{\hat{\beta}_i}$$

where λ is the corresponding quantile of the standard normal distribution, which follows from proposition 29. In the case of logistic regression, a $100(1-\alpha)\%$ confidence interval for the parameter β_i is defined just as above, but remembering that β_i is the logarithm of the odds ratio, we may obtain a $100(1 - \alpha\%)$ non-symmetric confidence interval for the odds ratio e^{β_i}

$$e^{(\hat{\beta}_i \pm \lambda_{1-\frac{\alpha}{2}} s.e_{\hat{\theta}_i})}$$

Example 4. If we consider the dichotomous outcome variable Y taking values $0, 1$ and the independent variable E (education level) taking values low, middle, high and missing. If we choose the lower education level as reference and decompose E into three dichotomous variables E_1, E_2, E_3 respectively, then the model in logit form is

$$\text{logit}(P(Y)) = \alpha + \beta_1 E_1 + \beta_2 E_2 + \beta_3 E_3$$

Now, lets us say that the statistical software produces the following output

	Estimate	Std. Error	z value	Pr($\geq \lvert z \rvert$)	
(Intercept)	-3.68427	0.02791	-132.008	$<$ 2e-16	***
Middle	0.08811	0.04006	2.199	0.0279	*
High	0.03870	0.04763	0.813	0.4165	
Missing	0.30865	0.12825	2.407	0.0161	*

where the second column specifies the estimates of each corresponding coefficient, and the fifth gives the p-values from Wald's test. The factor level low is not shown since it is the reference. From this table we may compute the odds ratio

$$\hat{OR}_{Middle,Low} = \exp\left[\hat{\alpha} + \hat{\beta}_1 1 + \hat{\beta}_2 E_2 + \hat{\beta}_3 E_3 - (\hat{\alpha} + \hat{\beta}_1 0 + \hat{\beta}_2 E_2 + \hat{\beta}_3 E_3)\right] = e^{\hat{\beta}_1} \approx e^{0.88} \approx 1.09$$

Similarly, the corresponding confidence intervals may be computed using the standard error specified in the third column

$$95\%\text{CI:} \quad \exp[0.088 \pm \lambda_{0.025} 0.04] \approx (1.01, 1.18)$$

3.6 Survival Analysis

In survival analysis the outcome variable of interest is time until an event occurs, called the survival time. The event may be death, disease etc. and is called failure. In competing risk analysis several events c_k are considered at the same time. Let $T \geq 0$ be the random variable for an individual X's survival time, and t be a realisation of T. The goal of survival analysis is to estimate, interpret and compare survivor and/or hazard functions derived from the survival data. The content of this section is based on Moeschberger et. al. (1997), unless stated otherwise.

3.6.1 Censoring

Censoring occurs when the true survival time is not known exactly. It may take several forms, of which the most common are defined below.

Definition 19. When the true survival time T_{true} is greater or equal to the observed survival time T_{obs} we call this data right censored. In other words, let C_r be a fixed right censoring time, and assume for each X the corresponding T's are i.i.d distributed, then the exact life time of X is known if and only if $T \leq C_r$, and if $T > C_r$ then the corresponding event is right censored.

Definition 20. Let T be the true lifetime associated with a specific individual X and C_l be a censoring time. If $T < C_l$ then we do not know the true value of T and we call T left censored. This means that the event of interest has already occurred before X is observed in the study.

Example 5. If we are following persons until they become HIV positive we may record a failure when a subject first tests positive for the virus. However we may not know the exact time of exposure to the virus, hence the data is left censored. If an individual drops out of the study, due to migration or an accident, or if the study ends before the individual produces a positive test for the virus, then the data is right censored.

Definition 21. When the knowledge of censoring time for an individual provides no further information about the person's likelihood of survival at a future time, if the individual had continued the study, then we call the censoring non-informative.

Remark 7. Assumptions From now on we assume that the censoring is non-informative and independent. Independent censoring means that the censoring is the same in each subgroup. Furthermore we assume that the probability of experiencing a certain event-type is the same during the time period i.e. time independent. As we shall see later, no data will be left censored in this paper, and therefore left censoring will not be considered.

3.6.2 Kaplan Meier estimator and the Log-Rank test

When we want to assess the survival time for different groups we look at their corresponding survival curves. These describe how members of the groups survived over time and may reveal a lot of information. Essentially, the survival curves estimate the probability of survival over the time period considered.

Definition 22. The function $S(t) = P(T > t)$ describes the probability of an individual surviving beyond the time t and is called the survivor function. When T is discrete, assuming it takes values t_j, $j = 1, 2, \ldots$, having probability mass function $p(t_j) = P(T = t_j)$, where $t_1 < t_2 < \ldots$, then the survival function is defined as

$$S(t) = P(T > t) = \sum_{t_j > t} p(t_j)$$

Definition 23. The function

$$h(t) = \lim_{\epsilon \to 0} \frac{P(t \leq T < t + \epsilon | T \geq t)}{\epsilon}$$

is called the hazard rate. It is the conditional probability that an individual who survives just to prior to time t experiences the event at time t. When T is discrete the hazard function is given by

$$h(t_j) = P(T = t_j | T \geq t_j) = \frac{p(t_j)}{S(t_{j-1})}$$

Remark 8. For discrete T the survival function may be written as the product of conditional survival probabilities

$$S(t) = \prod_{j:t_j \leq t} \frac{S(t_j)}{S(t_{j-1})}$$

and since $p(t_j) = S(t_{j-1}) - S(t_j)$ hence the survival function is related to the hazard function by the following equality

$$S(t) = \prod_{j:t_j \leq t} (1 - h(t_j))$$

Definition 24. Kaplan Meier Estimator Assume the events occur at n distinct times $t_1 < \cdots < t_n$ and that at time t_f there are m_f events, or deaths, and n_f number of individuals at risk. The estimator

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t_1 > t \\ \prod_{f:t_f \leq t}(1 - \frac{m_f}{n_f}) & \text{if } t_1 \leq t \end{cases}$$

is called the Kaplan Meier or Product-limit estimator. It is a step function with jumps at the observed event times. The quantity $\frac{m_f}{n_f}$ provides an estimate of the hazard rate.

Remark 9. Under certain regularity conditions the Product-limit estimator, for fixed t, has an approximate normal distribution.

The rest of this section is based on material from Kleinbaum et. al. (2012).

Remark 10. Right censoring is dealt with in the following way: if time r is right censored then the corresponding individual has been counted towards the numbers at risk up until time r but will not contribute to the numbers at risk after that, hence the number of individuals at risk after time r will be reduced. The censoring will not contribute to the number of events or deaths. Had the individual experienced death then the overall survival probability would have been smaller and had the individual lived the overall survival probability would have been larger, compared to the censored case.

In some cases it is relevant to test whether two Kaplan-Meier curves are similar over time, or not. For this one may use the Log-Rank test of equality. In other cases it might be more interesting to look at and compare the survival probability after a certain time, and for this it is possible to use confidence intervals for point estimates.

Definition 25. Let $t_{(f)}, \quad f \in \{1, \ldots, n\}$ be the ordered failure times where n is the total number of failure times. Let m_{if} be the number of subjects failing and n_{if} be the number of individuals at risk, at time f in group $i \in \{1, 2\}$. Then the total number of observed failures is $O_i = \sum_{f=1}^{n} m_{if}$ and the expected number of failures is $e_{if} = n_{if} \frac{m_{1f}+m_{2f}}{n_{1f}+n_{2f}}$ at time f and the expected total number of failures is $E_i = \sum_{f=1}^{n} e_{if}$.

Proposition 10.

$$Var(O_i - E_i) = \sum_{f=1}^{n} \frac{n_{1f} n_{2f} (m_{1f} + m_{2f})(n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2 (n_{1f} + n_{2f} - 1)}$$

Proposition 11. The Log-Rank test When testing H_0 : no difference between Kaplan-Meier curves of two independent groups against H_1 : Kaplan-Meier curves differ, we may use the Log-Rank test statistic

$$\frac{(O_i - E_i)^2}{Var(O_i - E_i)} \sim \chi^2_{df=1} \quad i = 1, 2$$

Proposition 12. Confidence Intervals The $100(1-\alpha)\%$ confidence interval for $\hat{S}(t)$ at time t is

$$\hat{S}(t) \pm \lambda_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}[\hat{S}(t)]}$$

where Greenwood's formula for the variance is given by

$$\hat{Var}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{f:t_{(f)} \leq t} \frac{m_f}{n_f(n_f - m_f)}$$

and $t_{(f)} = f$-ordered failure time, $m_f =$ number of failures at $t_{(f)}$ and $n_f =$ number at risk at $t_{(f)}$. The usage of the quantile of the normal distribution is motivated by remark 9.

3.6.3 Competing Risk Analysis and Cumulative Incidence Curves

When looking at several competing events of failure, event-types, that are mutually exclusive, like for example different causes of death, the goal is to assess and compare the death rates between the various competing events. One such way is to find the marginal probability of each competing event and use the Kaplan-Meier estimator for the total amount of failures independent of type of failure, and calculate the cumulative incidence for each event-type. This approach does not require that the competing risks are independent.

Definition 26. For the event type c_k we call

$$\hat{h}_{c_k}(t_i) = \frac{m_{c_k,i}}{n_i}$$

the estimated hazard, where $m_{c_k,i}$ is the number of events for event-type c_k at time t_i, n_i is the number of subjects at risk at time t_i, and there are g different event-types $k = 1, \ldots, g$.

Definition 27. The product

$$\hat{I}_{c_k}(t_i) = \hat{S}(t_{i-1}) \hat{h}_{c_k}(t_i)$$

is called the estimated incidence of failing from event-type c_k at time t_i.

Definition 28. The sum

$$\text{CIC}_{c_k}(t_i) = \sum_{j=1}^{i} \hat{I}_{c_k}(t_j)$$

is called the cumulative incidence for the event-type c_k at time t_i. It requires that the overall hazard is the sum of the individual hazards for all the risk types

$$h(t) = h_{c_1}(t) + \cdots + h_{c_k}(t) + \cdots + h_{c_g}(t)$$

which is satisfied if the events are mutually exclusive and non-recurrent. Plotting the curve for each time t_i will show the proportion of dying at each time, for the specific event-type c_k, and hence produces the estimated marginal probabilities.

Proposition 13. The cumulative incidence curve and Kaplan-Meier curve are closely related to each other in the following way. Assume that there is no competing risk, then

$$CIC(t_i) = 1 - \hat{S}(t_i)$$

Proof: (by the author)

A simple proof by induction on i. For $i = 1$, then $\hat{S}(t_0) = 1$ by definition of $\hat{S}$, so

$$1 - \hat{S}(t_1) = 1 - (1 - \hat{h}(t_1)) = h(t_1) = CIC(t_1)$$

Now assume that $CIC(t_i) = 1 - \hat{S}(t_i)$ is true $\forall i \leq p$, then

$$1 - \hat{S}(t_{p+1}) = 1 - \prod_{j=1}^{p+1}(1 - \hat{h}(t_j)) = 1 - \hat{S}(t_p)(1 - \hat{h}(t_{p+1}))$$

$$= 1 - \hat{S}(t_p) + \hat{S}(t_p)\hat{h}(t_{p+1}) = [\text{ind. hyp.}] = \sum_{j=1}^{p}[\hat{S}(t_{j-1})\hat{h}(t_j)] + \hat{S}(t_p)\hat{h}(t_{p-1}) = CIC(t_{p+1})$$

and the proposition follows.

Remark 11. Sometimes one might want to estimate the survival probability in the hypothetical scenario where disease A exists and all other competing risk do not. This is generally not possible, since there is likely an unknown dependence between the competing risks, but we may consider death by a competing disease as a censored event, and the overall survival function $\hat{S}_A$ instead of $\hat{S}$. Plotting $\hat{S}_A(t)$ will give the corresponding Kaplan-Meier curve. This is called the censoring method. Under the hypothetical assumption of no competing risks, by proposition 53, the cumulative probability of death may be estimated by

$$1 - \hat{S}_A(t_p) = \sum_{j=1}^{p} \hat{S}_A(t_{j-1})\hat{h}_A(t_j)$$

Observe that

$$\sum_{j=1}^{p} \hat{S}_A(t_{j-1})\hat{h}_A(t_j) \neq \sum_{j=1}^{p} \hat{S}(t_{j-1})\hat{h}_A(t_j) = CIC_A(t_p)$$

Dignam and Kocherginsky (2008) explain that $1 - \hat{S}_A(t_p)$ will produce an overestimation of the cumulative event-specific probability since each death by competing risk will not be counted as a death. The choice of whether to use cumulative incidence or the hypothetical Kaplan Meier curve depends on the question of interest. This phenomenon can be seen in the next example.

Example 6. This example is based on an example from Kleinbaum et. al. (2012), page 449. Consider a study involving 24 individuals receiving radiotherapy (RT) for the treatment of prostate cancer. Patients may either die of the disease (cancer), other causes or still be alive at the time of analysis, time is given in weeks. The data are shown in the table below, including the calculations required for the cumulative incidence curve for the event-type death from cancer (ca). D indicates the type of death, where d is death by the disease of interest, c is censored and o is death by other cause.

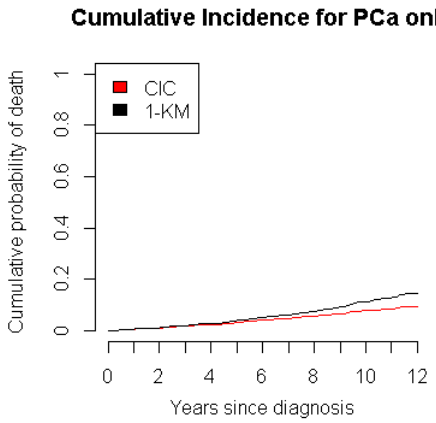
f	t_f	n_f	D	$m_{ca,f}$	$\hat{h}_{ca}(t_f)$	$\hat{S}(t_{f-1})$	$\hat{I}_{ca}(t_f)$	$\text{CIC}_{ca}(t_f)$
1	0	24	-	0	0	-	0	0
2	0.7	24	d	1	0.042	1.000	0.042	0.042
3	1.5	23	o	0	0	0.958	0	0.042
4	2.8	22	o	0	0	0.916	0	0.042
5	3.0	21	d	1	0.048	0.875	0.042	0.083
6	3.2	20	c	0	0	0.833	0	0.083
7	3.8	19	o	0	0	0.833	0	0.083
8	4.7	18	o	0	0	0.789	0	0.083
9	4.9	17	d	1	0.059	0.745	0.044	0.127
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Steps:

1. $\hat{h}_{ca}(t_f) = \frac{m_{ca,f}}{n_f}$
2. $\hat{S}(t_{f-1}) = \prod_{j:t_j \leq t_f}(1 - \frac{m_j}{n_j})$
3. $\hat{I}_{ca}(t_f) = \hat{S}(t_{f-1})\hat{h}_{ca}(t_f)$
4. $\text{CIC}_{ca}(t_f) = \sum_{j=1}^{f} \hat{I}_{ca}(t_j)$

The overall survival function $\hat{S}$ is directly affected by death by other causes since that is an event, but not by censoring since censoring is not an event, and only affects the numbers at risk n_f. For example, at time $t_f = 1.5$ there is a death from other causes, such that $1 - \frac{m_f}{n_f} = 1 - \frac{1}{23} = 0.9565$ and hence the next entry is $0.9565 \times 0.958 = 0.916$. On the other hand, when $t_f = 3.2$ there is censoring and $\hat{S}$ is constant. Had all the deaths by other causes been considered as censored then $\hat{S}$ would have been greater. This explains why the censoring method overestimates the probability of failure for the event-type considered.

An example of this can be seen in the figure below, where men missing data for risk stage categorisation were analysed for death by prostate cancer (PCa). The corresponding 1-Kaplan-Meier curve is shown for death by PCa only.



Here we see that, under the hypothesis that no other cause of death exists, then death by PCa is more likely.

As the log-rank test was used to compare Kaplan-Meier curves there is a similar test to compare several cumulative incidence curves, for the same competing risk, between populations. There are several way to do this, but the most popular is Gray's test, which concludes the theory section.

Remark 12. Gray's test To test if two or several CIC's for the same competing risk, from different and independent subsets, or groups, are similar or differ, we may use a certain k-sample test, called Gray's test, Gray (1988). It does not require independence of competing risks. The details of the test is beyond the scope of this thesis and are best described in its original form. Intuitively it works similarly to the log-rank test but for CIC's.

4 Statistical Analysis

Initially, a table was created to overview the data, including each factor and various subsets of the data, to illuminate potential differences and anomalies, which in turn may indicate where to perform a deeper analysis. A graphical representation of the proportions of missing data and a graph over the proportion of men in the subsets over the years of diagnosis will be produced.

4.1 Data Cleaning and Logistic Regression

The status of the risk category, i.e. risk assessed versus information missing to assess risk will first be assessed with univariate and multivariate logistic regression, resulting in adjusted odds ratios, 95% confidence intervals, and p-values from Wald's test for the coefficients, using the function *glm*. Then, for each of the four subsets of missing data, univariate and multivariate logistic regression will be performed to produce similar output.

Some outcomes were excluded or combined with others, in cases where no individuals were present in that particular cell. This was done to avoid errors, division by zero and/or other complications that would not add anything to the model. Likelihood-ratio tests were computed with the R-package *lmtest*.

4.1.1 Models

Logistic regression requires models, and all models for this will be expressed in terms of the logit function. In univariate models there will be one model for each factor, and these will be indexed by k. The factors considered are age when diagnosed (intervals), level of education, year of diagnosis (intervals), treatment, hospital, physician, mode of detection and CCI - a total of eight factors. The following outcomes will be considered, where U_i denotes outcome indexed i.

U_1: Missing data

U_2: Missing data due to missing Gleason

U_3: Missing data due to missing PSA

U_4: Missing data due to missing T-stage

U_5: Missing data due to combinations of 2-3 of missing Gleason, PSA, T-stage

There is one univariate model for each factor, and for the multivariate there is only one model containing a sum of factors. For all outcomes except missing data (missing data for risk stage categorisation), only men missing data for risk categorisation are included. Each factor $1 \leq k \leq 8$ has l_k levels and each level is a dichotomous variable $X_{j,k}$, where $1 \leq j \leq l_k$.

Univariate logistic regression for outcome: U_i

$$\text{logit}(P(U_i|X))_k = \alpha + \sum_{j=1}^{l_k} \beta_{j,k} X_{j,k} \quad , \quad 1 \leq k \leq 8$$

Multivariate logistic regression for outcome: U_i

$$\text{logit}(P(U_i|X)) = \alpha + \sum_{k=1}^{8} \sum_{j=1}^{l_k} \beta_{j,k} X_{j,k}$$

4.2 Survival Analysis

No modeling is required for the survival analysis applied in this thesis. The competing risks are divided into four categories: death by prostate cancer, other cancer, cardiac/vascular and other causes. The reason being that cardiac/vascular is the most common cause of death, tumours being the second, and among tumors PCa is the most common cause of death for men, as of 2012, Statistics Sweden (2013).

The assumptions made are: no left censoring (since it is not possible to assess this via the data), independent and non-informative censoring. The survival time starts when the individual is diagnosed with prostate cancer. Furthermore we assume that the events are time independent.

Kaplan-Meier curves are estimated for overall survival and survival of prostate cancer (PCa) where other causes of death were considered as censored. The Log-Rank test is performed for testing H_0 : no difference between the Kaplan-Meier curves vs H_1 : difference, using *survdiff()*, where appropriate (subsets of data are independent/disjoint), and 95% confidence intervals, presented in brackets, are computed for the probability of survival after 12 years from diagnosis.

In addition, a competing risk analysis is performed to analyse differences between the subsets above, and some other subsets of men, with the purpose to explain the differences seen in the previous tables. Competing risk analysis is performed using the R-package called *cmprsk*, which among other things produces cumulative instance curves. Gray's test is applied to assess H_0 : resp. risk CIC's are the same between two independent/disjoint subsets, against H_1 : difference.

5 Results

5.1 Overview of data

The proportion of men missing data for risk stage categorisation in the register was about 2.56%, or 3315 men out of 129 389. An overview of the data can be seen in the appendix (table 1), which indicates that men missing data have the same age and education as other men. About 98.2% of men missing data have at least one of Gleason, PSA or T-stage specified and these men have generally lower Gleason score, PSA levels and T-stages, indicating that men missing data of risk stage categorisation have mostly low risk prostate cancer. Regarding treatments, RT, RP and GnRH+AA, are all under-represented for MD while expectance/surveillance are over-represented. Again, this indicates that men missing data of risk stage categorisation mainly consist of men with low risk PCa.

As seen in figure 1a below MD was be split into four subsets: missing data due to missing Gleason, PSA, T-stage and combinations, called reasons for missing data of risk stage categorisation.

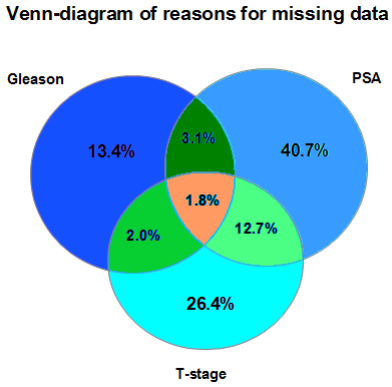


Figure 1a The proportions of reasons of all men missing data.

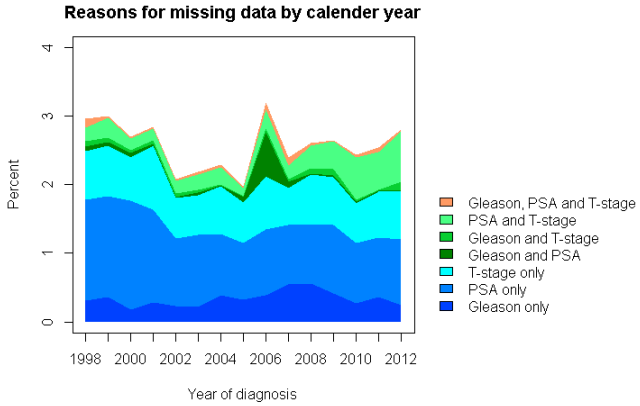


Figure 1b The proportions of missing data by year of diagnosis of all men.

Figure 1b shows that the proportions vary over the years of diagnosis, with a decrease during 2002-2005 and an increase around 2006, especially for men missing data due to missing a combination of Gleason and PSA. When considering these results it is relevant to know that at around 2006-2007 there was a change of system of registration to INCA[4].

5.2 Logistic Regression

Changes made to factor treatment: Non-curative now includes AA, GnRH, surgical castration and non-curative other/missing. RT includes RT brachy, external, external + brachy and unknown, Missing includes missing, curative therapy other/missing and death before treatment decision/initiation.

The p-values for most coefficients are close to zero, indicating that the corresponding factor level should be included in the model. When comparing the uni- and multivariate models it is important to note changes in factor levels, which often occur for several factors simultaneously, and which could indicate collinearity.

[4]http://www.cancercentrum.se/sv/INCA/om-inca/

5.2.1 Missing data

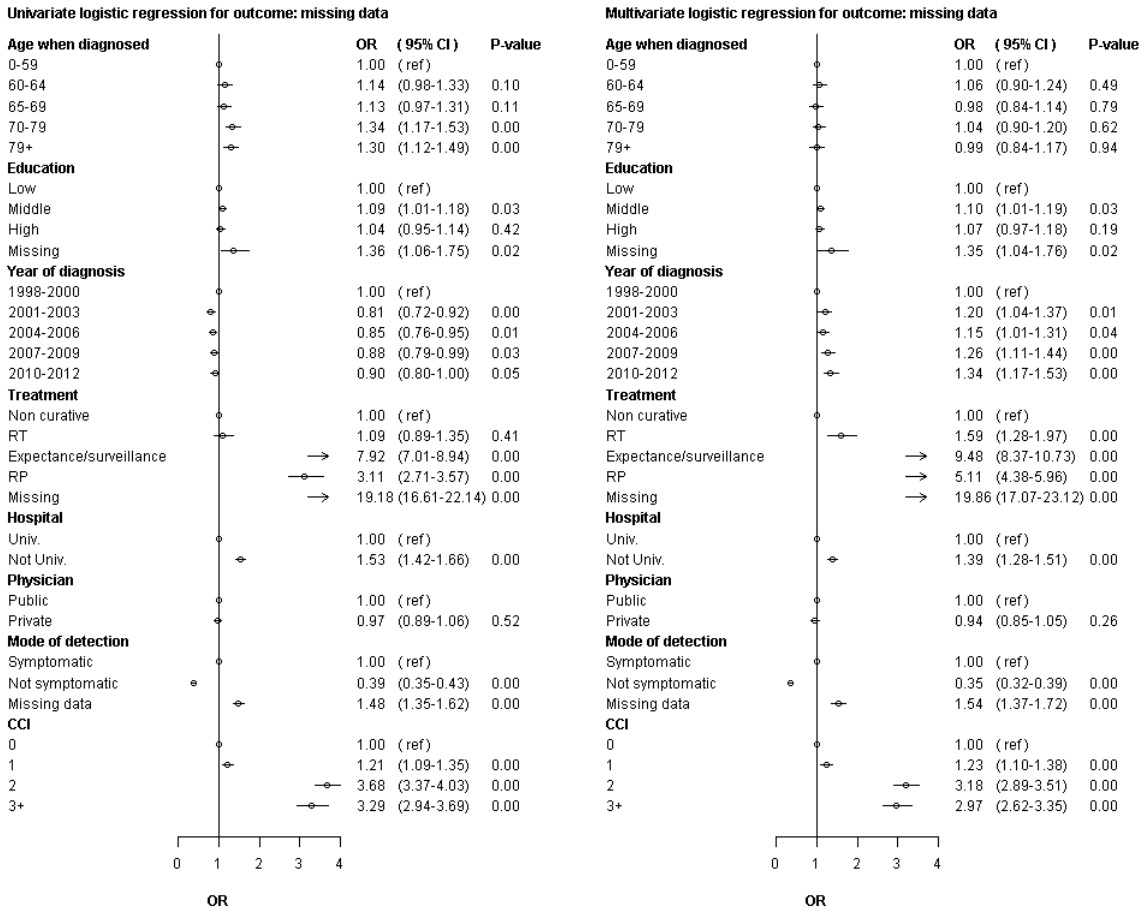


Figure 2 Likelihood of missing data. Univariate logistic regression (left): no adjustments made, multivariate logistic regression (right): adjustments made for age of diagnosis, education, year of diagnosis, treatment group, type of hospital, type of physician, method of discovery and CCI. CCI=Charlson's comorbidity index; OR=odds ratio; P-values by Wald test.

Figure 2 shows that men with high comorbidity (CCI 2, 3+), attending other hospitals than university, missing education, mode of detection or treatment data have large odds ratios and are therefore are more likely to miss data. The same applies to men treated by RP or expectance/surveillance. This can be seen in the univariate model, and in the multivariate model where adjustments are made for the other variables. Hence the results are not spurious findings but can be explained by collinearity. Men discovered by non-symptomatic reasons are less likely to miss data, and there are no differences between private and public physicians. In the univariate model older men (70+) seem to have a larger likelihood of missing data, but the effect vanishes when adjustments are made for other variables. On the other hand, the odds ratios for year of diagnosis change from ≤ 1 to ≥ 1 when adjusting for other variables, indicating that year of diagnosis could be affected by other variables, most likely age at diagnosis. Considering the likelihood ratio test, for the univariate models: Education (p=0.03), Year of diagnosis (p=0.01), Physician (p=0.52) and p<0.001 for all other models, and the multivariate model.

5.2.2 Reasons for Missing Data of Risk Stage Categorisation

Men with high comorbidity have a larger likelihood of miss data due to missing PSA, combinations or possibly T-stage (fig. 3b-d). Surprisingly, men missing Gleason (fig. 3a) are younger and have less comorbidity. There are also some variations of odds ratios over year of diagnosis, which corresponds to what can be seen in figure 1b.

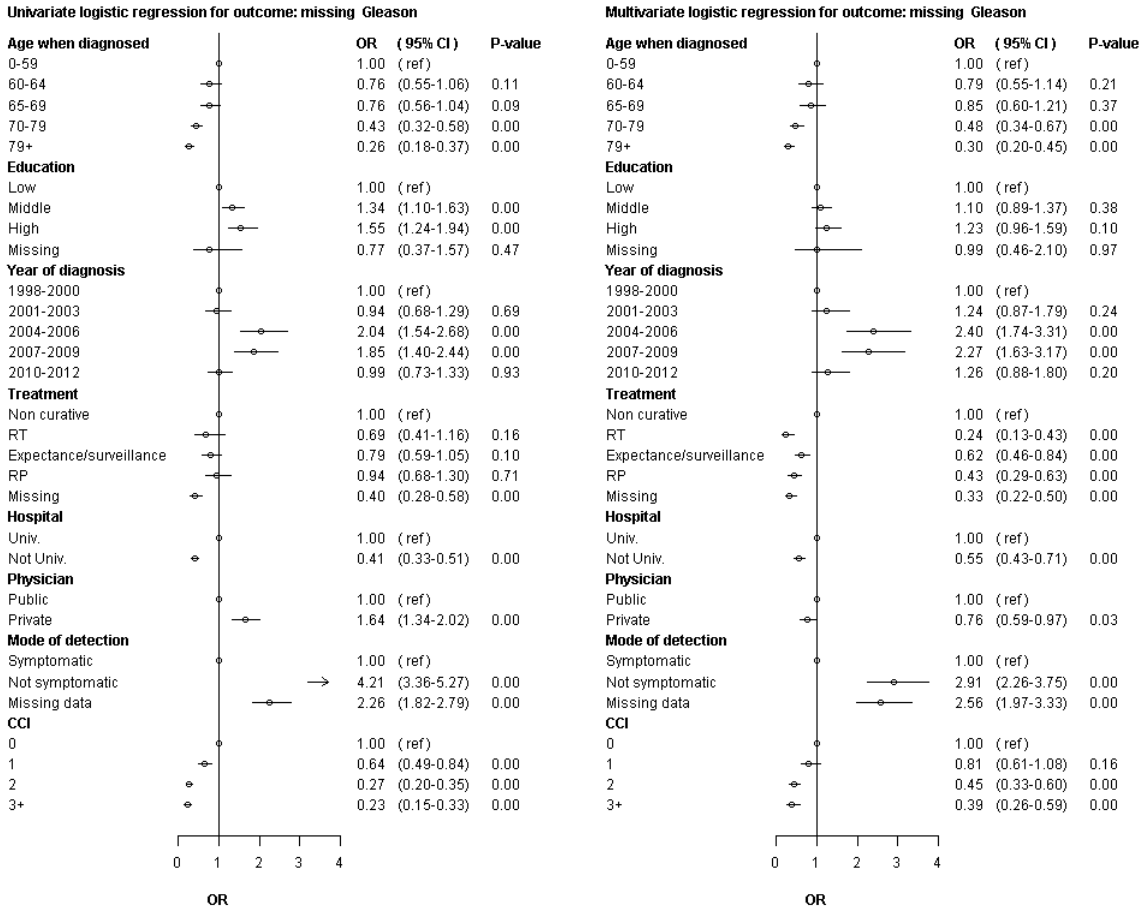


Figure 3a Likelihood of missing data. Univariate logistic regression (left): no adjustments made, multivariate logistic regression (right): adjustments made for age of diagnosis, education, year of diagnosis, treatment group, type of hospital, type of physician, method of discovery and CCI. CCI=Charlson's comorbidity index; OR=odds ratio; P-values by Wald test.

In contrast to men missing data for risk stage category in general, men that are younger, have lower CCI, and/or are discovered by non symptomatic reasons, attending university hospitals, attending private physicians and/or have a higher education level, are more likely be missing risk stage category due to missing Gleason score, as seen in figure 3a. There was an increased amount of these men during the years 2004-2009. Men missing treatment are less likely to miss data, yet men missing mode of detection are more likely. This can be seen in both models, but when the adjustments for other variables are made the differences in education vanish and men attending

private physicians are less likely to miss data. On the other hand, the effect if treatment strategy on risk of missing Gleason score is more pronounced in the multivariate analysis indicating a correlation between treatment and physicians and education, possibly with a link between private physicians and RP. Considering the likelihood ratio test $p<0.001$ for all models.

Men missing data of risk stage categorisation due to missing PSA value are more similar to men missing data for risk categorisation in general. High comorbidity characterises these men, and so does the high age when diagnosed.

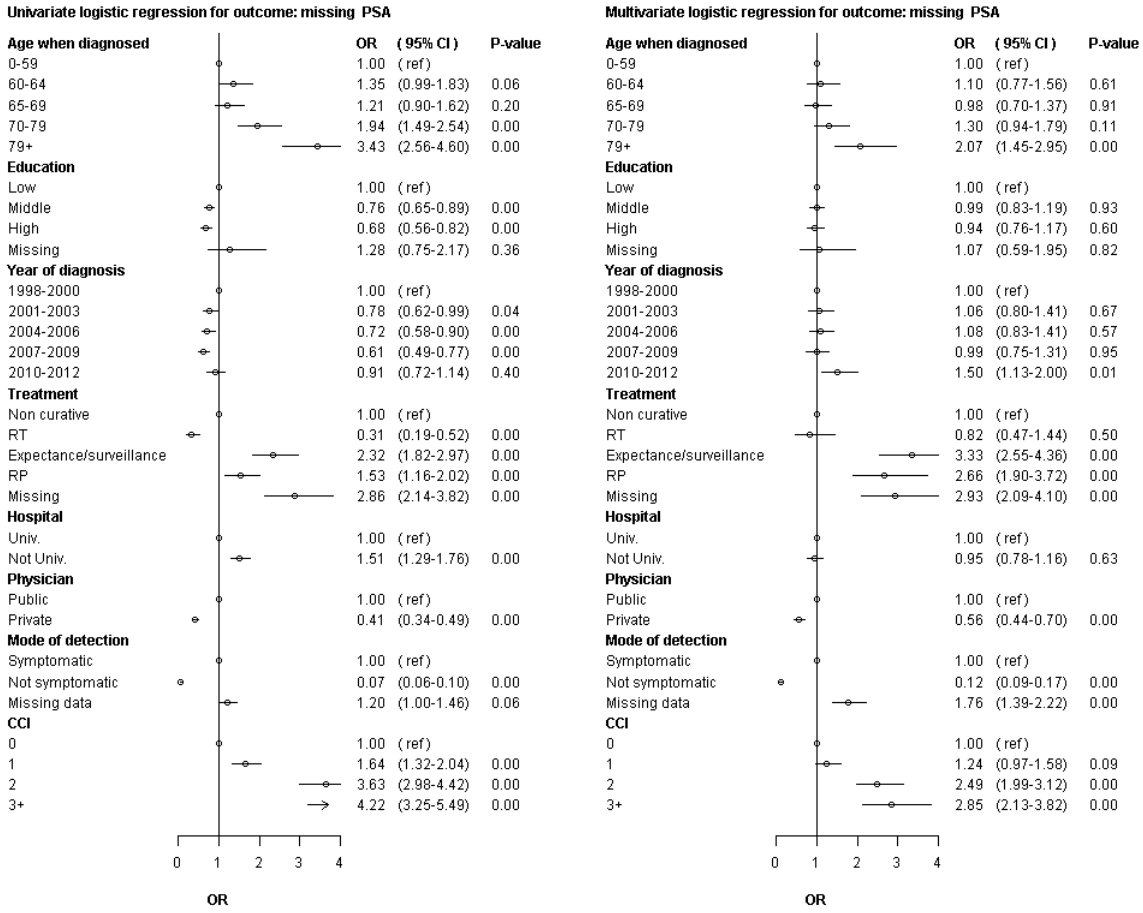


Figure 3b Likelihood of missing data. Univariate logistic regression (left): no adjustments made, multivariate logistic regression (right): adjustments made for age of diagnosis, education, year of diagnosis, treatment group, type of hospital, type of physician, method of discovery and CCI. CCI=Charlson's comorbidity index; OR=odds ratio; P-values by Wald test.

Figure 3b shows that older men (70+), with high comorbidity (CCI 2, 3+), attending public physicians and other hospitals are more likely to be missing data of risk stage categorisation due to missing PSA value. They have a slightly lower level of education but the differences disappear in the multivariate model, as well as the difference between hospitals. Men missing mode of detection or treatment data are more likely to be missing PSA value. It is less likely that men missing data for risk categorisation are missing PSA value between 2001-2009, but when adjustments are made for the other variables the differences vanish and instead the years 2010-2012 show an increased likelihood. This again indicates a relationship between year of diagnosis and age when diagnosed.

The odds ratios for RP and RT increase in the multivariate model while there is a small increase of odds ratio for private physician. Considering the likelihood ratio test $p<0.001$ for all models.

In contrast to the previous results, men missing data of risk stage categorisation due to missing T-stage have an ambiguous relationship with comorbidity and age when diagnosed, as seen below.

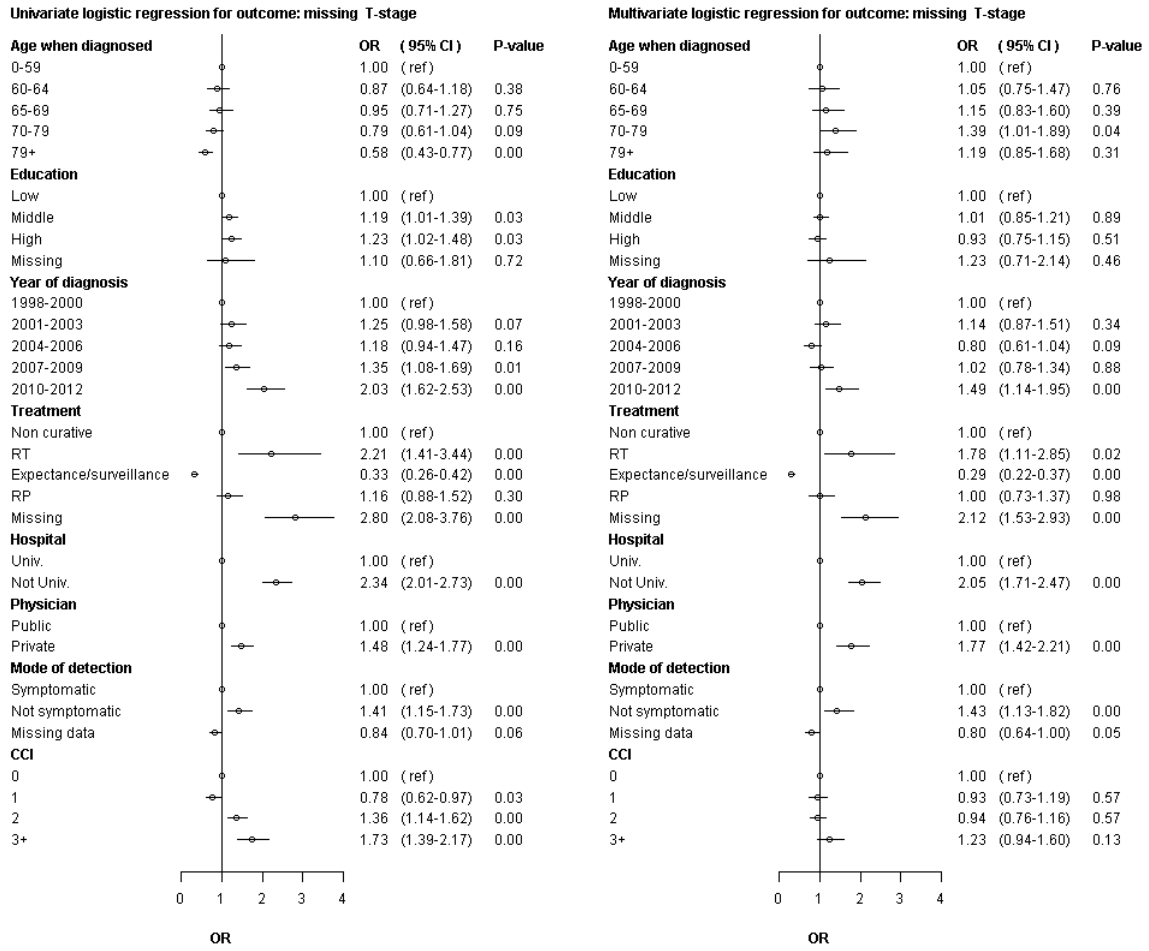


Figure 3c Likelihood of missing data. Univariate logistic regression (left): no adjustments made, multivariate logistic regression (right): adjustments made for age of diagnosis, education, year of diagnosis, treatment group, type of hospital, type of physician, method of discovery and CCI. CCI=Charlson's comorbidity index; OR=odds ratio; P-values by Wald test.

In figure 3c we see that there are small differences in age when diagnosed, comorbidity and education, but they tend to disappear when adjusting for other variables. There are also some differences in year of diagnosis, with an increased likelihood of missing data during 2007-2012, yet this effect is subdued when adjustments are made, indicating a collinearity between year of diagnosis and age when diagnosed. Most intriguing is that men treated by RT or missing treatment data, attending other hospitals and/or private physicians are more likely to be MDTS in both models. Considering the likelihood ratio test $p<0.001$ for all models except the univariate model including Education ($p=0.08$).

Men missing due to combinations (fig. 3d) share similar properties with the other subgroups, yet there are some interesting differences. While high comorbidity, other hospitals, and missing

treatment and/or discovery data seem to be contributing factors, there are also some major fluctuations over year of diagnosis.

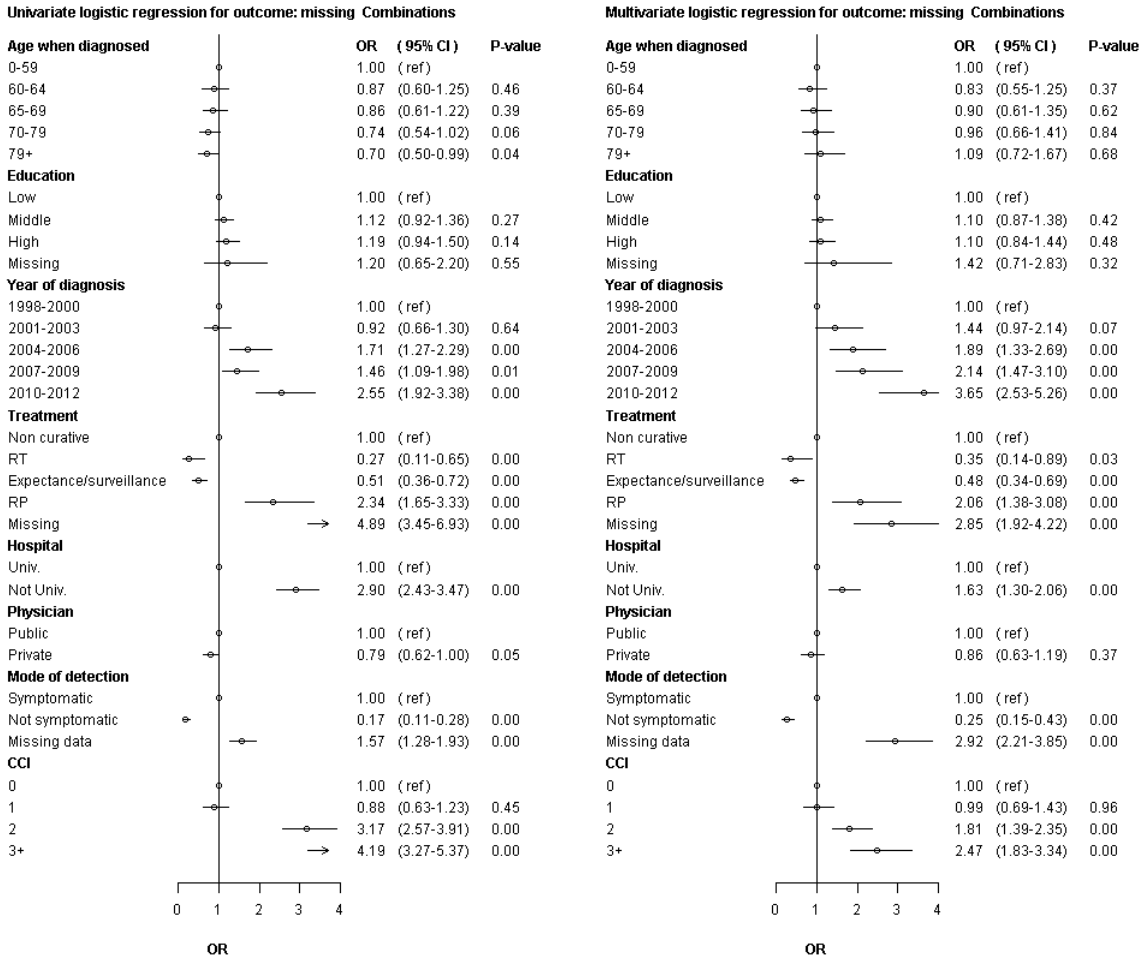


Figure 3d Likelihood of missing data. Univariate logistic regression (left): no adjustments made, multivariate logistic regression (right): adjustments made for age of diagnosis, education, year of diagnosis, treatment group, type of hospital, type of physician, method of discovery and CCI. CCI=Charlson's comorbidity index; OR=odds ratio; P-values by Wald test.

Figure 3d shows that there is an increased likelihood of men missing data of risk stage categorisation due to combinations of missing Gleason score, PSA value and T-stage during 2004-2012, with a peak at 2010-2012, as seen in both models. This is also reflected with more detail in figure 1b. No differences in ages when diagnosed, education or for physicians. Men with higher comorbidity (CCI 2, 3+) and/or attending other hospitals have an increased likelihood of missing data of risk stage categorisation due to missing combinations, as well as men missing treatment or mode of detection data. Men treated with RP have an increased likelihood, while men treated with RT and expectance/surveillance have a lower likelihood compared to non-curative. No major differences between the uni- and multivariate models. The likelihood ratio test gives, for the univariate models: Education ($p=0.47$), Age when diagnosed ($p=0.18$), Physician ($p=0.04$) and $p<0.001$ for all other models, including the multivariate.

5.3 Survival Analysis

The relationship between comorbidity and men missing data of risk stage categorisation in general, and some of the reasons, further motivates to assess the survival and competing risks of death.

First a comparison between men missing data of risk stage categorisation and others is made (fig. 4a-d), and then we make comparisons between the reasons of missing data of risk stage categorisation (fig. 5a-f). The most interesting difference is the reduced probability of death by PCa for men missing data of risk stage categorisation, agreeing with previous results that these men mainly have low risk PCa, compared to others. The large proportion of death by other cancer in almost all comparisons agrees with the generally higher comorbidity levels of men missing data of risk stage categorisation.

Remembering that the logistic regression revealed interesting results considering men missing treatment or mode of detection data, as well as some differences between physicians and between hospitals, a comparison for these subsets is also made, showing the corresponding survival curves and cumulative incidence curves. All men considered in this case are missing data for risk categorisation. Missing treatment data includes missing, curative therapy other/missing and death before treatment decision/initiation.

The most noticeable differences are those between private and public physicians and between men missing treatment data and men not missing treatment data (fig. 6a-d). No major differences were found for men missing mode of detection data compared to men not missing mode of detection data (fig. 7a-d). More detailed results follow below.

5.3.1 Missing Data of Risk Stage Categorisation

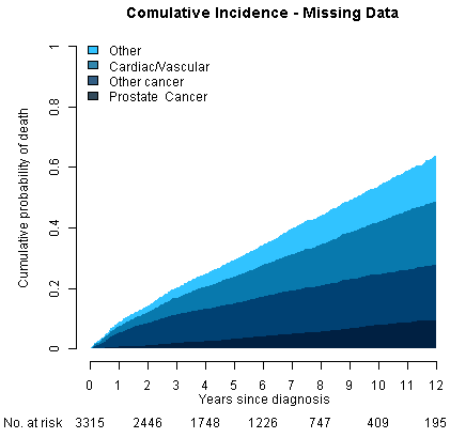


Figure 4a

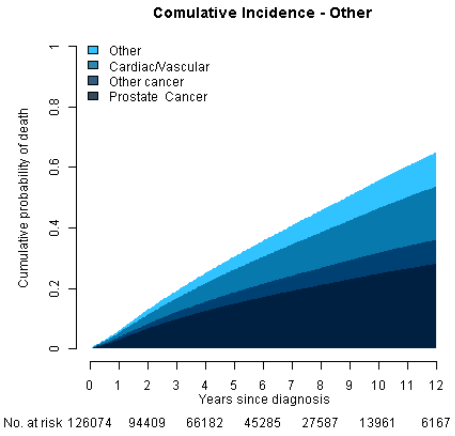


Figure 4b

Considering men missing data of risk stage categorisation, cardiac/vascular and other cancer are similar in proportion while other causes of death are less likely and so is death by prostate cancer as seen in figure 4a-b. On the contrary, men who do not miss data of risk stage categorisation have a larger proportion of death by prostate cancer and a smaller proportion of death by other cancer. This is reflected by Gray's test which produces p-values < 0.002 for all competing risks.

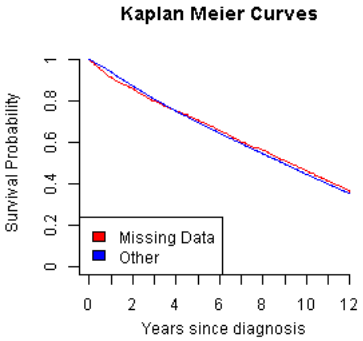


Figure 4c

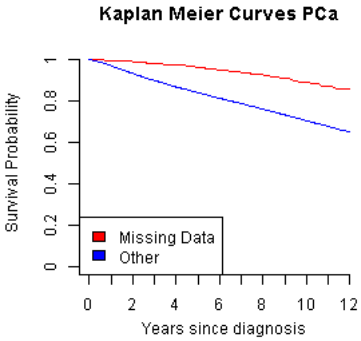


Figure 4d

In total, the mortality rate is similar, at 64% (61-67%) after 12 years from diagnosis for men missing data of risk stage categorisation and 65% (64-65%) for other (fig. 4c). The log-rank test for overall survival indicates no difference, with a p-value of 0.44. For prostate cancer (fig. 4d), on the other hand, the mortality rate is 15% (12-17%) for men missing data of risk stage categorisation and 35% (35-36%) for other men, and the log-rank test produced a p-value of < 0.001, indicating a difference between the curves, where men missing data of risk stage categorisation generally have a higher survival probability.

Looking at reasons for missing data of risk stage categorisation we see even more apparent differences between the competing risks, and also between total survival, for the different subsets. Most remarkable is the low mortality rate for men missing Gleason, and for the other reasons, the large proportion of death by other cancer.

5.3.2 Reasons for Missing Data of Risk Stage Categorisation

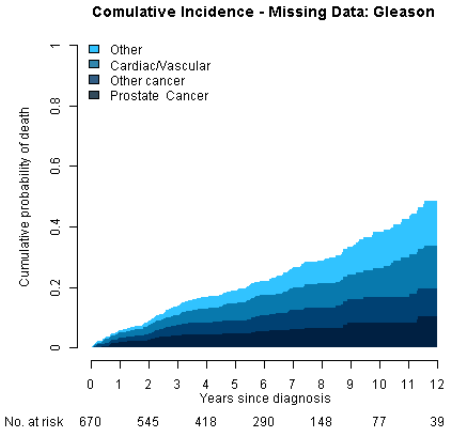


Figure 5a

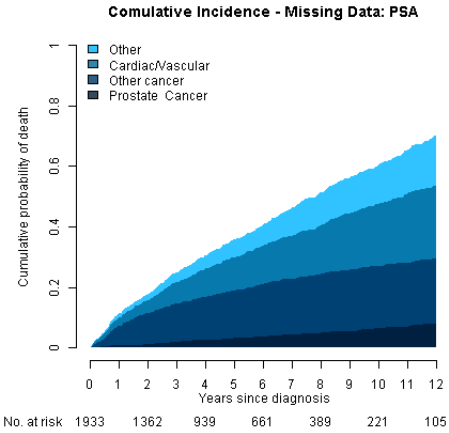


Figure 5b

Men missing data of risk stage categorisation due to missing Gleason have similar proportions of causes of death (fig. 5a-b). Men missing data of risk stage categorisation due to missing PSA have a low proportion of death by prostate cancer, and larger proportions of death due to cardiac/vascular.

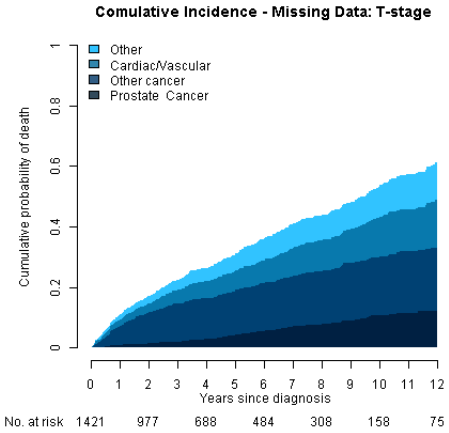


Figure 5c

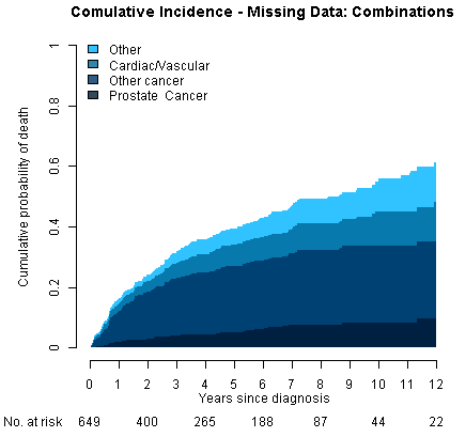


Figure 5d

Men missing data of risk stage categorisation due to T-stage have a larger proportion of death by other cancer, while the other causes are similar in proportion (fig. 5c-d). Men missing data of risk stage categorisation due to combinations have a low proportion of death by prostate cancer and a large proportion of death by other cancer.

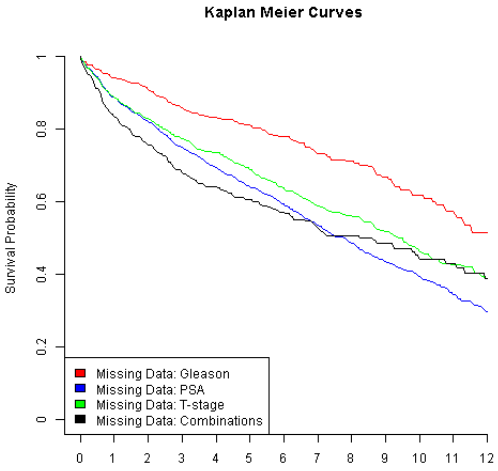


Figure 5e

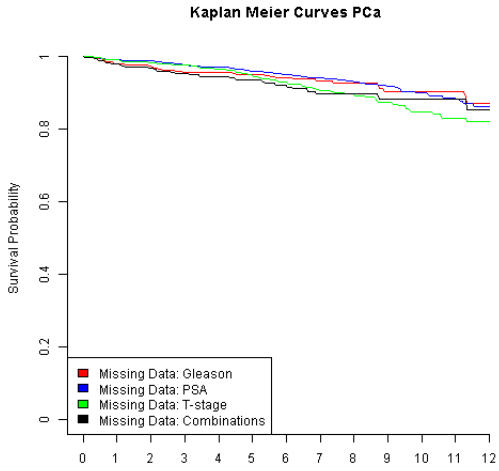


Figure 5f

The overall mortality rate (fig. 5e) differs between the reasons, where combinations has a faster decrease in the beginning of the time period and Gleason has the overall lowest mortality rate over the entire period. Men missing Gleason have an overall mortality rate of about 48% (41-56%) after 12 years, while it is higher for men missing PSA at and about 71% (67-74%). Similarly, the mortality rate for missing T-stage is around 62% (57-66%) and for combinations it is around 63% (55-71%) after 12 years of diagnosis. Since the subsets, or reasons, overlap no log-rank or Gray test was applied.

When considering death by PCa only (fig. 5f) the mortality rate is fairly similar for all reasons, where men missing Gleason have 13% (7-18%), PSA 14% (10-17%), T-stage 18% (14-22%) and combinations 15% (8-21%) after 12 years of diagnosis.

5.3.3 Other subsets: Treatment

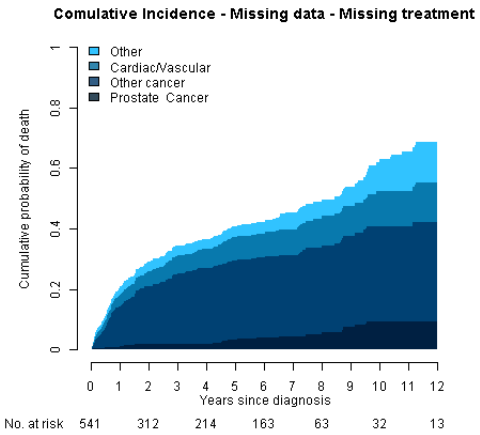


Figure 6a

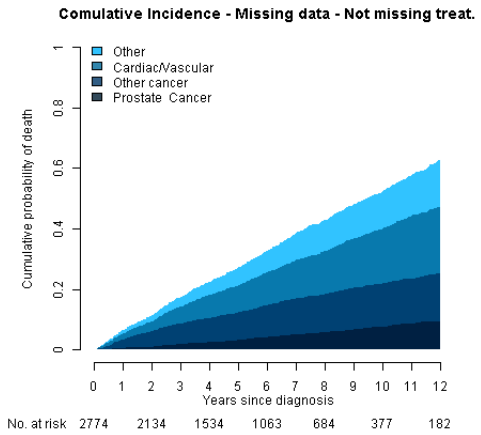


Figure 6b

Men missing data of risk stage categorisation and missing treatment data (including curative therapy other/missing and death before treatment decision/initiation) have a significantly larger proportion of probability of death by other cancer, while the other causes seem to be similar in proportion (fig. 6a-b). Men missing data of risk stage categorisation and not missing treatment data have a similar proportion of death by prostate cancer, while the other causes are slightly larger and of approximately same proportions. For prostate cancer and other causes the Gray test produces p-values 0.87 resp. 0.11 while for cardiac/vascular and other cancer the p-values are < 0.05, so we may conclude that the CIC's for cardiac/vascular differ and CIC's for other cancer differ.

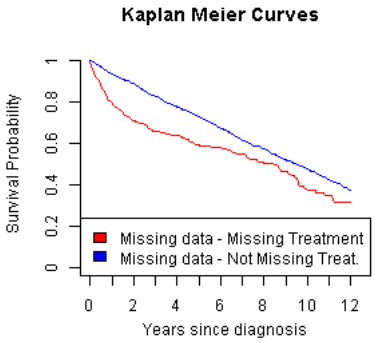


Figure 6c

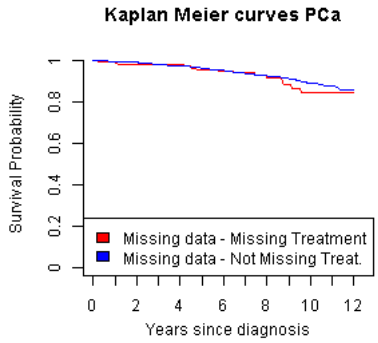


Figure 6d

The mortality rate is about 68% (60-71%) after 12 years from diagnosis for men missing treatment data, while it is is about 63% (60-66%) for men not missing treatment data (fig. 6c). The log-rank test yields the p-value < 0.001 for overall survival, indicating a difference. For PCa only (fig. 6d) the mortality rate is 16% (8-24%) for men missing treatment and 14% (12-17%) for other, and the log-rank yields $p = 0.37$ so no major difference.

5.3.4 Other subsets: Mode of detection

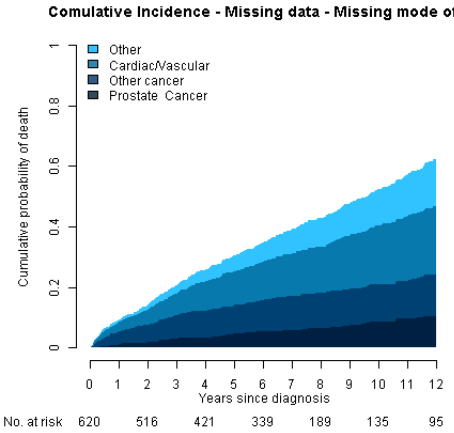


Figure 7a

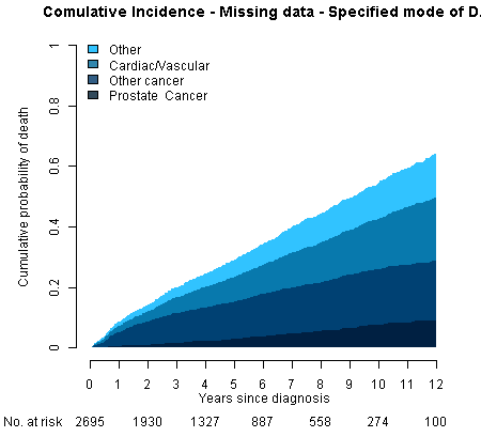


Figure 7b

Men missing data of risk stage categorisation and missing mode of detection data have a slightly larger proportion of probability of death by cardiac/vascular than the other causes (fig. 7a-b), and a slight decrease in proportion of death by cardiac/vascular. The Gray test produced a p-value < 0.01 for cardiac/vascular and p-values > 0.11 for the other risks, so we may conclude that the CIC's for cardiac/vascular risk differ.

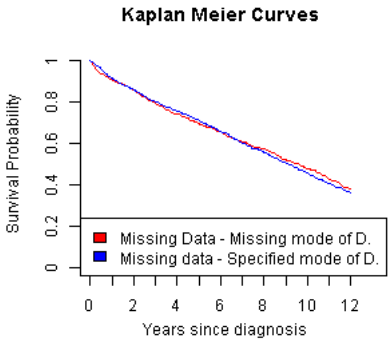


Figure 7c

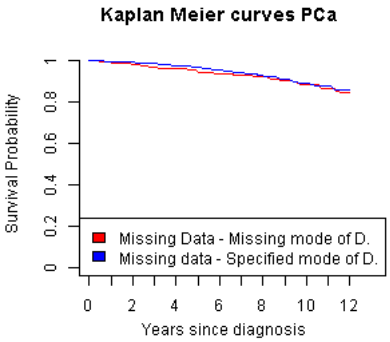


Figure 7d

In total, the mortality rate is just below 64% (61-67%) after 12 years from diagnosis for both men missing mode of detection data and those who do not (fig. 7c), and the log-rank test yields the p-value 0.82 for overall survival. For PCa only (fig. 7d) the mortality rate is 15% (12-17%) for men missing mode of detection data and others, with $p = 0.33$, indicating no difference in either case.

5.3.5 Other subsets: Physicians

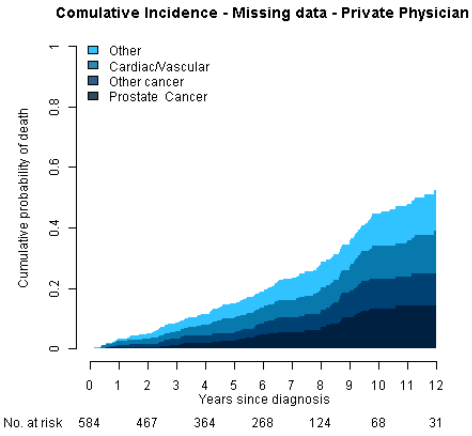


Figure 8a

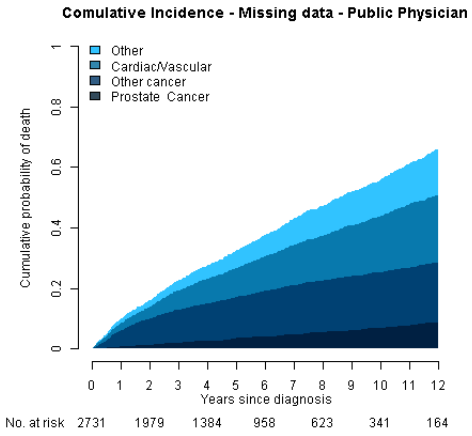


Figure 8b

Men missing data of risk stage categorisation and attending private physicians have a slightly larger proportion of death by prostate cancer (fig. 8a-b). Men missing data of risk stage categorisation and attending public physicians have a low proportion of death by prostate cancer, while death by other cancer and cardiac/vascular are larger in proportion. Gray's test produced a p-value 0.25 for other, and p-values < 0.05 for the other risks, so we may conclude that the CIC's for all risk differ, except for other.

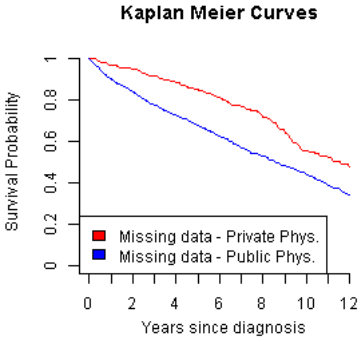


Figure 8c

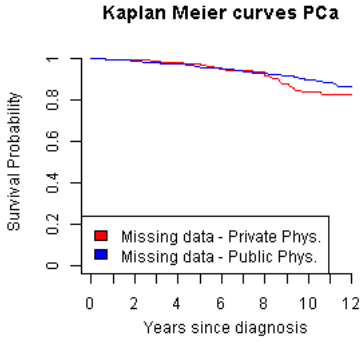


Figure 8d

The mortality rate is about 54% (46-62%) for men attending private physicians after 12 years from diagnosis, with a lower mortality in the first years of diagnosis and higher at the end (fig. 8c). In contrast to men attending private physicians the mortality rate is higher, about 66% (63-69%) after 12 years from diagnosis, and a larger proportion of men die earlier. The log-rank test yielded the p-value < 0.001 for overall survival indicating difference. For PCa only (fig. 8d) the mortality rate is 18% (11-24%) for private physicians and 14% (11-17%) for public, with $p = 0.31$, indicating no major difference.

5.3.6 Other subsets: Hospitals

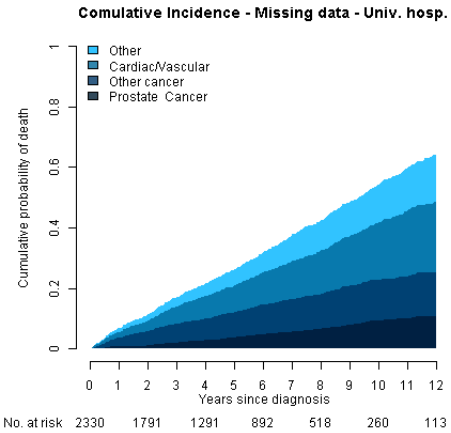


Figure 9a

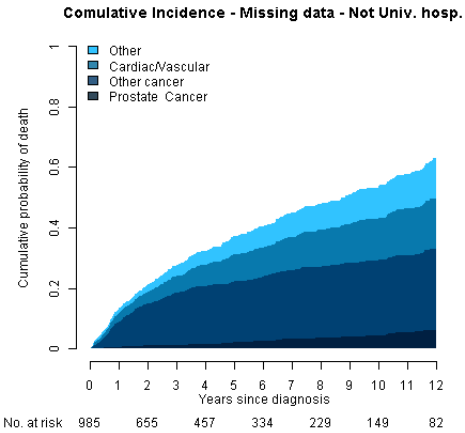


Figure 9b

Men missing data of risk stage categorisation and attending university hospitals have evenly distributed proportions of causes of death, with a slight increase in cardiac/vascular (fig. 9a-b). Men missing data of risk stage categorisation and attending other hospitals than university hospitals have a large proportion of death by other cancer, especially at the beginning of the time period, and a small proportion of death by prostate cancer. Gray's test produced a p-value 0.26 for other causes, and p-values < 0.03 for the other risks, so we may conclude that the CIC's for all competing risks differ, except for men that died by other causes.

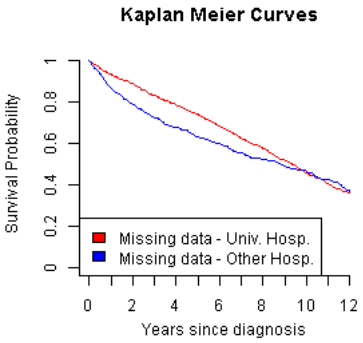


Figure 9c

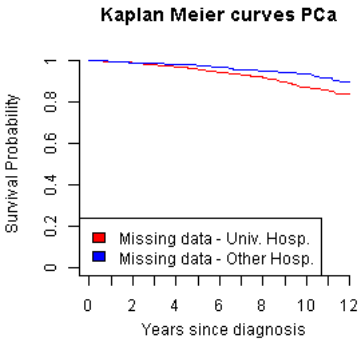


Figure 9d

For men attending university hospitals the mortality rate is about 64% (61-68%) after 12 years from diagnosis, and similarly, for men attending other hospitals it is about 63% (59-68%) (fig. 9c). The log-rank test yields the p-value 0.002 for overall survival, indicating difference. While for PCa only (fig. 9d), the mortality rate is 16% (13-20%) for university hospitals and 11% (6-15%) for other hospitals, with $p = 0.004$, indicating that the curves differ.

6 Discussion and Conclusions

6.1 Missing data of Risk Stage Categorisation

One goal of the thesis was to describe the men who are missing data of risk stage categorisation and compare them to other men in the database in order to understand why they are missing data or risk stage categorisation. In general, men missing data of risk stage categorisation had higher comorbidity levels, a lower proportion of death by PCa and a larger proportion of death by other cancer (fig. 2), indicating that their PCa risk stage was most likely low. About 98.2% of men missing data of risk stage categorisation (fig. 1a) have at least one of Gleason, PSA or T-stage specified, and they had generally lower Gleason score, PSA levels and T-stages, again indicating that they had mostly low risk prostate cancer (table 1). This was also indicated by their lower odds ratios for being treated by RT and RP, which aim at curing, and GnRH+AA, which is a retarding treatment for incurable PCa (fig. 2).

Regarding mode of detection, among men missing data of risk stage categorisation, there were no major differences of survival and competing risks, except for cardiac/vascular, between men with missing and men with specified mode of detection data (fig. 7a-d). On the other hand, men missing treatment data (fig. 6a-d) have a significantly larger proportion of death by other cancer and have a larger probability of death closer to the time of diagnosis, compared to men with specified treatment data. One reason for this might be that another cancer demanded more urgent attention and the treatment for PCa was of passive nature and not registered.

Men attending private physicians had a higher survival probability, with a larger proportion of death by PCa and lower proportions of death by other cancer and cardiac/vascular, compared to men attending public physicians (fig. 8a-d). This could hypothetically be explained by the fact that private clinics might not want to accept patients with a pessimistic forecast, that is with several ongoing diseases, or that patients with a pessimistic forecast might not prioritise an active choice of physician, assuming that they in this case probably would attend a public physician.

On the other hand, men attending university hospitals had a larger probability of death by PCa and a smaller probability of death by other cancer (fig. 9a-d). Men attending other hospitals had a larger probability of death by other cancer and it is more pronounced closer to time of diagnosis. It was more likely that men attending other hospitals were missing data of risk stage categorisation. Possibly, men attending university hospitals had less comorbidity or at least had more urgent need to handle the PCa, due to the larger proportion of death by PCa, and therefore also had lower probability of missing data of risk stage categorisation, compared to men attending other hospitals.

6.2 Reasons for Missing Data of Risk Stage Categorisation

Another goal of the thesis was to understand why men are missing data of risk categorisation by looking at the different reasons of missing data, and to describe and compare men in the subgroups defined by these reasons.

Firstly, men missing Gleason were about 20.3% of all men missing data of risk stage categorisation, and were surprisingly younger and had less comorbidity (fig. 3a). This was reflected by their high survival probability. One reason for missing Gleason could have been the choice of not performing the necessary biopsy due to risk of infection and pain. Since men missing Gleason level were a small portion of all men missing data for risk categorisation, and about a third of these men were missing more than just Gleason level. Their cancer were more likely to have been detected by PSA-testing since non-symptomatic mode of detection had an increased odds ratio. Since the Gleason level was more difficult and risky to assess it it could possibly have been avoided if PSA and/or T-stage had been assessed and indicating low risk PCa. Since the men missing Gleason were younger and had lower comorbidity it is possible that indication of low risk sometimes motivated the decision to not perform the biopsy, and hence the formal risk categorisation was less prioritised.

Secondly, about 58.3% of all men missing data of risk stage categorisation were missing PSA (fig. 3b). These men were old, had high comorbidity levels, were more likely to be treated by expectance/ surveillance, RP or missing treatment, and to attend public physicians and other hospitals than university. They had a low probability of death by PCa and a high probability of death by other cancer and cardiac/vascular, with a larger proportion of death closer to the time of diagnosis, and a high probability of death in general. This was reflected by their higher age and comorbidity. The increased odds for expectance/surveillance could be connected to the high comorbidity and age, and the lower odds ratio when mode of detection was non-symptomatic is expected since the method uses PSA-testing. Since this only required a simple blood sample the PSA value was probably found but not registered. This could be explained by the high age and high comorbidity due to some other kind of cancer that required more immediate attention.

Thirdly, men missing data of risk stage categorisation and missing T-stage (fig. 3c) were about 43% of all men missing data, and had a large proportion of death by other cancer and a slightly larger proportion of death by PCa compared to men missing PSA (fig. 5a-f). High comorbidity did not seem to have been a major factor, though the results were inconclusive when comparing the uni- and multivariate analysis. Surprisingly, men treated with RT, attending private physicians and/or other hospitals than university were more likely to be missing T-stage. Since the assessment of T-stage was simpler than that of Gleason level, but depended on an experienced urologist, having knowledge of how to perform the analysis by feeling with a finger, it might have been that the procedure was avoided in a larger extent by private physicians and/or other hospitals, especially if there was another cancer that required more attention. Another explanation could be that the procedure was performed but not registered cause of inexperience and/or uncertainty of the precise categorisation. Perhaps the PCa in average was slightly more severe , as indicated by the larger proportion of death by PCa.

In addition, men with high comorbidity levels and/or attending other hospitals than university were more likely to miss data due to missing combinations (fig. 3d). These men were about 19.6% of all men missing data, and had a significantly larger proportion of death by other cancer and a large proportion of death closer to the time of diagnosis (fig. 5a-f). As argued above, missing any combination, or possibly all three of Gleason level, PSA level and T-stage could indicate that there was another disease, probably another cancer, that required more attention thus making testing for PCa less interesting or that tests indicated that the PCa was of lower risk. This was reflected by the high comorbidity and large proportion of death by other cancer. The most probable explanation to why men were missing all three could be a combination of high comorbidity and an assessment of PSA, T-stage or Gleason that was not registered properly but that probably indicated lower risk, due to the low proportion of death by PCa.

Lastly, there were some fluctuations between the proportions of the reasons of missing data (fig. 1b) with a decrease between 1998 and 2005, a peak around 2006 and a change of proportions between the reasons. This was reflected in the logistic regression (fig. 2, 3a-d), where there also were some indications of collinearity between age when diagnosed and year of diagnosis. The peak around 2006 could possibly be explained by the change of system of registration of data to the INCA platform, which occurred at the time, leaving a general increase of missing data of risk stage categorisation behind it, perhaps due to a looser control during the transition and unfamiliarity of the new system. Some of the other variations could possibly be explained by similar changes on other levels or unknown factors. Changes in interpretation of the risk stage categorisation algorithm or in the assessment of Gleason level, PSA value and T-stage could possibly explain the some of the variation and the connection between age and year of diagnosis.

6.3 Weaknesses and Strengths of the Study

A weakness of this study was the limited timespan of 1998-2012 when there was more data available. The main reason was that before 1998 NPCR did not cover the whole nation and therefore the data from earlier years did not reflect similar circumstances.

On the other hand, a strength of the study is the large amount of data, both counted in number of men registered and the number of variables available.

Another major strength of this study was the use of a large database with detailed information which is generalizable to all men in Sweden with prostate cancer, with a capture rate of 98% of all cases of prostate cancer registered in Sweden compared with the Cancer Registry to which registration is mandated by law, since 1998, Tomic et. al. (2015).

6.4 Conclusions

Men missing data of risk stage categorisation generally had high comorbidity, and are more likely to die by other cancer, and less likely to die by PCa. They most likely had a PCa that could have been categorised as low risk. The correlation between missing data of risk stage categorisation and high comorbidity levels was not surprising, though the assessment of risk (biopsy to acquire Gleason score, blood sample for PSA level or rectal examination to obtain a T-stage) could possibly have been affected by comorbidity and age, especially in combination with another cancer and a low risk PCa, resulting in unregistered results or unperformed tests and hence rendering missing data of risk stage categorisation. Surprisingly, men attending private physicians and/or treated by RT were likely to be missing T-stage possibly due to slightly higher PCa risk and inexperience/uncertainty related to the assessment of T-stage. Younger men with low comorbidity were more likely to be missing Gleason level, and a possible explanation is that they had indications of low risk PCa via PSA-testing. Men missing PSA-value were older and had high CCI, and were probably missing PSA due to failed registration due to indications of low risk PCa and a more imminent other cancer. The variations seen when looking at the proportions of missing data by year of diagnosis could partly be explained by the change of IT-system for registration, assuming a negative dependency between successful registrations and recent changes in the system. There were also indications of a connection between year of diagnosis and age when diagnosed.

7 Recommendations and Future Work

It would be interesting to further assess the relationship between missing data of risk stage categorisation and missing treatment and mode of detection data, comorbidity and death by other cancer, as well as the results regarding private physicians, RT-treatment and missing T-stage. It would also be of interest to investigate how much men missing data of risk stage categorisation differ compared to men with low risk PCa.

The results regarding the low age of diagnosis and comorbidity for men missing Gleason are baffling and should be given more attention. There also seems to be a relation between changes in system of registration and missing data of risk stage categorisation, where changes cause more missing data of risk stage categorisation for a period of time, which is understandable. The connection between age and year of diagnosis, and other factors and connection not considered in this thesis, should be looked upon to assess and improve the quality of the register over time. In fact, it would be recommended to consider the main results presented in this paper when discussing quality maintenance of the register, perhaps including a regional perspective, and the quality of the registration and care given to patients with indicated but not registered low risk PCa, especially in combination with a high comorbidity level.

8 References

Theory

Bouliotis, George, Billingham, Lucinda, *Crossing survival curves: alternatives to the logrank test*, From Clinical Trials Methodology Conference 2011, Bristol, UK. 4-5 October 2011, http://www.trialsjournal.com/content/12/S1/A137, accessed 23-04-2015.

Dignam, James J. and Kocherginsky, Maria N., *Choice and Interpretation of Statistical Tests Used When Competing Risks Are Present*, Journal of Clinical Oncology - Statistics in Oncology, Vol. 26, Nr. 24 Aug. 20, 2008

Dobson, Annette J. *An Introduction to Generalized Linear Models*. 2nd ed. Chapman & Hall/CRC, 2002

Gray, Robert J, *A class of K-sample tests for comparing the cumulative incidence of a competing risk*, The annals of Statistics, vol 16, No. 3 1141-1154, 1988

Kleinbaum, David G., Klein, Mitchel. *Logistic Regression - A Self Learning Text*. 3rd ed., Springer, 2010

Kleinbaum, David G., Klein, Mitchel. *Survival Analysis - A Self-learning text*. 3rd ed., Springer, 2012

Moeschberger, Melvin L., Klein, John P. *Survival Analysis*. Springer, 1997

Sainani, Kristin L. *Statistically Speaking: Understanding Odds Ratios*. American Academy of Physical Medicine and Rehabilitation, Vol. 3, 263-267, March 2011

Sainani, Kristin L. *Statistically Speaking: Logistic Regression*. American Academy of Physical Medicine and Rehabilitation, Vol. 6, 1157-1162, December 2014

Rashid, Mamunur, Naima Shifa. *Consistency of the Maximum Likelihood Estimator in Logistic Regression Model: A Different Approach*, Journal of Statistics Volume 16, 2009, pp. 1-11

Zwanzig, Silvelyn, Liero, Hannelore. *Introduction to the Theory of Statistical Inference*. CRC press, 2012

Research and Publications

Charlson, Mary E., et. al., *A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation*, Journal of Chronic Diseases 40 (5): 373-383, 1987

Van Hemelrijck, Mieke, et. al., *Cohort Profile: The National Prostate Cancer Register of Sweden and Prostate Cancer data Base Sweden 2.0*, International Journal of Epidemiology 42:956967, 2013

Mohler, James, et. al., *Prostate Cancer Clinical Practice Guidelines in Oncology, Journal of the National Comprehensive Cancer Network*, Vol. 8 Num. 2, Feb. 2010

Sandin, Fredrik, Wigertz, Annette. *Prostatacancer, Årsrapport från Nationella prostatacancerregistret 2013*, Regionalt cancercentrum, Uppsala Örebro, 2014

Statistics Sweden, *Causes of Death 2012*, Official Statistics of Sweden, Statistics Health and Medical Care, 2013, http://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/19175/2013-8-6.pdf, accessed 11-05-2015

Stattin, Pär, Bratt, Ola, et al. *Manual Nationella prostatacancerregistret (NPCR) Diagnostik och primärbehandling*, Regionalt cancercentrum Uppsala/ Örebro, 2013

Tomic, Katarina et. al.,*Capture rate and representativity of The National Prostate Cancer Register of Sweden*, Acta Oncologica, 54: 158163, 2015

9 Appedix

9.1 Table 1

	Missing		Other		Total	
	n	(%)	n	(%)	n	(%)
Total	3315	(100)	126074	(100)	129389	(100)
Age when diagnosed						
mean (sd)	71.1	(9)	70.4	(9.2)	70.4	(9.2)
0-59	106	(3.2)	4605	(3.7)	4711	(3.6)
60-64	715	(21.6)	30578	(24.3)	31293	(24.2)
65-69	1214	(36.6)	47975	(38.1)	49189	(38)
70-79	1011	(30.5)	32486	(25.8)	33497	(25.9)
79+	269	(8.1)	10430	(8.3)	10699	(8.3)
Education						
Low	1316	(39.7)	52398	(41.6)	53714	(41.5)
Middle	1244	(37.5)	45354	(36)	46598	(36)
High	689	(20.8)	26392	(20.9)	27081	(20.9)
Missing	66	(2)	1930	(1.5)	1996	(1.5)
Year of diagnosis						
mean (sd)	2005.4	(4.3)	2005.5	(4.2)	2005.5	(4.2)
1998-2000	592	(17.9)	19926	(15.8)	20518	(15.9)
2001-2003	565	(17)	23346	(18.5)	23911	(18.5)
2004-2006	709	(21.4)	27935	(22.2)	28644	(22.1)
2007-2009	722	(21.8)	27533	(21.8)	28255	(21.8)
2010-2012	727	(21.9)	27334	(21.7)	28061	(21.7)
Treatment						
AA	64	(1.9)	6754	(5.4)	6818	(5.3)
Death before treatment d/i*	23	(0.7)	459	(0.4)	482	(0.4)
Exspectance/surveillance	1753	(52.9)	32470	(25.8)	34223	(26.4)
GnRH	83	(2.5)	10955	(8.7)	11038	(8.5)
GnRH + AA	65	(2)	19899	(15.8)	19964	(15.4)
Non-curative therapy o/m**	64	(1.9)	1168	(0.9)	1232	(1)
Surgical castration	31	(0.9)	6088	(4.8)	6119	(4.7)
Surgical castration + AA	1	(0)	293	(0.2)	294	(0.2)
Curative therapy o/m**	315	(9.5)	789	(0.6)	1104	(0.9)
RP	590	(17.8)	27802	(22.1)	28392	(21.9)
RT brachy	20	(0.6)	2312	(1.8)	2332	(1.8)
RT external	80	(2.4)	10344	(8.2)	10424	(8.1)
RT external + brachy	21	(0.6)	3718	(2.9)	3739	(2.9)
RT unknown	2	(0.1)	135	(0.1)	137	(0.1)
Missing	203	(6.1)	2888	(2.3)	3091	(2.4)
Hospital						
Univerity	2330	(70.3)	98848	(78.4)	101178	(78.2)
Other	985	(29.7)	27226	(21.6)	28211	(21.8)
Physician						
Public	2731	(82.4)	103311	(81.9)	106042	(82)
Private	584	(17.6)	22763	(18.1)	23347	(18)
Mode of detection						
Not symptomatic	435	(13.1)	36954	(29.3)	37389	(28.9)
Symptomatic	2260	(68.2)	75154	(59.6)	77414	(59.8)
Missing data	620	(18.7)	13966	(11.1)	14586	(11.3)
CCI						
0	1819	(54.9)	93060	(73.8)	94879	(73.3)
1	411	(12.4)	17313	(13.7)	17724	(13.7)
2	702	(21.2)	9754	(7.7)	10456	(8.1)
3+	383	(11.6)	5947	(4.7)	6330	(4.9)
T-stage						
Missing	239	(7.2)	161	(0.1)	400	(0.3)
T0	198	(6)	328	(0.3)	526	(0.4)
T1(abc missing)	20	(0.6)	105	(0.1)	125	(0.1)
T1a	756	(22.8)	3682	(2.9)	4438	(3.4)
T1b	285	(8.6)	2759	(2.2)	3044	(2.4)
T1c	492	(14.8)	46769	(37.1)	47261	(36.5)
T2	341	(10.3)	39301	(31.2)	39642	(30.6)
T3	0	(0)	27002	(21.4)	27002	(20.9)
T4	0	(0)	4942	(3.9)	4942	(3.8)
TX	984	(29.7)	1025	(0.8)	2009	(1.6)
N-stage						
N0	501	(15.1)	13963	(11.1)	14464	(11.2)
N1	0	(0)	2519	(2)	2519	(1.9)
NX	2602	(78.5)	108622	(86.2)	111224	(86)
Missing	212	(6.4)	970	(0.8)	1182	(0.9)

	Missing		Other		Total	
	n	(%)	n	(%)	n	(%)
M-stage						
M0	864	(26.1)	45154	(35.8)	46018	(35.6)
M1	0	(0)	11625	(9.2)	11625	(9)
MX	2263	(68.3)	68468	(54.3)	70731	(54.7)
Missing	188	(5.7)	827	(0.7)	1015	(0.8)
Gleason						
Gleason 2-6	2048	(61.8)	50906	(40.4)	52954	(40.9)
Gleason 3+4	255	(7.7)	20091	(15.9)	20346	(15.7)
Gleason 4+3	85	(2.6)	12493	(9.9)	12578	(9.7)
Gleason 7***	30	(0.9)	5333	(4.2)	5363	(4.1)
Gleason 8	0	(0)	12851	(10.2)	12851	(9.9)
Gleason 9-10	0	(0)	10704	(8.5)	10704	(8.3)
WHO 1	179	(5.4)	3280	(2.6)	3459	(2.7)
WHO 2	48	(1.4)	5444	(4.3)	5492	(4.2)
WHO 3	0	(0)	2856	(2.3)	2856	(2.2)
Missing Gleason/WHO	670	(20.2)	2116	(1.7)	2786	(2.2)
PSA						
median (1:st Qu., 3:rd Qu.)	7.20	(4.70, 10.58)	12.00	(6.00, 32.00)	12.00	(6.00, 32.00)
0-4	248	(7.5)	9546	(7.6)	9794	(7.6)
5-10	755	(22.8)	44998	(35.7)	45753	(35.4)
11-20	379	(11.4)	26170	(20.8)	26549	(20.5)
21-50	0	(0)	20705	(16.4)	20705	(16)
51-100	0	(0)	9213	(7.3)	9213	(7.1)
100+	0	(0)	14161	(11.2)	14161	(10.9)
missing	1933	(58.3)	1281	(1)	3214	(2.5)

*decision/initiation **other/missing ***not 3+4/4+3 or unspecified